A Distance Covariance-based Estimator*

Emmanuel Selorm Tsyawo† Abdul-Nasah Soale‡

September 27, 2024

Abstract

This paper introduces an estimator that significantly weakens the relevance condition of conventional instrumental variable (IV) methods, allowing endogenous covariates to be weakly correlated, uncorrelated, or even mean-independent, though not independent of instruments. As a result, the set of relevant instruments is maximised in any given empirical setting. Identification without excludability is feasible, and the disturbance term does not need to possess any moments. Under a weak conditional median independence condition for pairwise differences in disturbances, along with mild regularity assumptions, identification is achieved. Furthermore, the estimator is shown to be consistent and asymptotically normal. To enhance its practical utility, this paper also demonstrates the testability of the relevance condition.

Keywords: distance covariance, dependence, weak instrument, endogeneity, U-statistics

JEL classification: C13, C14, C26

*This paper benefitted from valuable feedback from Brantly Callaway, Feiyu Jiang, Weige Huang, Oleg Rytchkov, and Dai Zusai.

†Corresponding author, emmanuel.tsyawo@um6p.ma, AIRESS & FGSES, Université Mohammed VI Polytechnique.

‡abdul-nasah.soale@case.edu, Department of Mathematics, Applied Mathematics, and Statistics, Case Western Reserve University.

1 Introduction

Empirical work in economics often relies on instrumental variable (IV) methods. However, when instruments are weakly correlated with endogenous covariates, conventional IV methods such as two-stage least squares (2SLS), the control function (CF) method, and the generalised method of moments (GMM) become unreliable, leading to biased estimates and hypothesis tests with significant size distortions. Furthermore, IV methods are infeasible when excluded instruments are unavailable or uncorrelated with the endogenous variables. These conventional methods are also highly sensitive to outliers and in scenarios where the disturbance term U lacks finite first moments. While much of the econometric literature on the weak instrument problem is focused on detection and weak-instrument-robust inference, theoretical progress on estimation is scant (Andrews, Stock, and Sun, 2019). This paper introduces a new estimator that minimises the distributional dependence between a parameterised disturbance $U(\theta)$ and a set of instruments Z using the distance covariance measure (dCov) proposed by Székely, Rizzo, and Bakirov (2007). The proposed Minimum Dependence estimator (MDep) substantially relaxes the instrument relevance requirement, allows for instruments that are not independent of covariates X, and remains robust even when the disturbance term U lacks finite moments.

The MDep has remarkable features that render it fundamentally different from existing IV methods. (1) The non-independence identifying variation means the MDep can exploit the maximum number of instruments available in any given empirical setting. (2) In the absence of excluded instruments, MDep inference remains reliable as long as covariates X are not independent of instruments Z. (3) Without estimating quantile models, the MDep shares the "robustness" feature of quantile estimators, e.g., Powell (1991) and Oberhofer and Haupt (2016), in the sense that its asymptotic properties do not depend on the existence of moments of U. By replacing Z with a bounded one-to-one mapping such that Z and the mapping generate the same Euclidean Borel field, one obviates moment existence conditions

on Z as well.[1] This third feature is important as economic theory can go as far as justifying the exogeneity of instruments but typically *cannot* go far enough to justify the existence of moments of U. This paper is the first, to the best of the authors' knowledge, to introduce an IV estimator that exploits arbitrary "first stage" distributional dependence of unknown form for identification in a general class of models.

Unlike conventional IV methods, the form of identifying variation needs to be neither known nor specified, effectively eliminating the sensitivity of estimates to first-stage model specification.[2] Although this property is also shared by integrated conditional moment estimators (ICM hereafter), e.g., Domínguez and Lobato (2004), Escanciano (2006), Antoine and Lavergne (2014), Escanciano (2018), and Tsyawo (2023), it is worth emphasising that the MDep relevance condition is more general. For example, $\mathbb{E}[X|Z] \neq \mathbb{E}[X]$ *almost surely* (*a.s.*) does not imply nor is implied by $\mathbb{E}[Z|X] \neq \mathbb{E}[Z]$ *a.s.* The MDep exploits both forms of dependence, while ICM estimators can only exploit the former. The MDep can achieve identification without excludability; this is more general than similar identification highlighted in Tsyawo (2023) and Gao and Wang (2023). Within the class of models under consideration in this paper, the MDep simply requires that no non-trivial linear combination X be independent of Z, unlike the ICM class. Not requiring the existence of moments of the disturbance makes the MDep robust to outliers and forms of contamination of the disturbance that do not possess any moments.

Related Literature

The literature on ICM estimators is perhaps the mostly closely related to the MDep estimator. Thanks to objective functions based on ICM measures of mean dependence, this class of estimators minimise the mean dependence of a parameterised disturbance term and a set of instruments. Examples of ICM estimators are proposed by Domínguez and Lobato

[1] An example of such a mapping is $\text{atan}(Z)$.

[2] Dieterle and Snell (2016), for example, uncovers substantial sensitivity of conclusions to specification (linear versus quadratic) of the first stage.

(2004), Escanciano (2006), Antoine and Lavergne (2014), Escanciano (2018), Wang (2018), Antoine and Sun (2022), and Tsyawo (2023). A related category of estimators converts mean independence restrictions into several or a continuum of unconditional moment conditions indexed by a nuisance parameter on an index set, and estimation is typically conducted via IV methods such 2SLS or GMM – see, e.g., Carrasco and Florens (2000), Donald, Imbens, and Newey (2003), Hsu and Kuan (2011), and Carrasco and Tchuente (2015). In spite of the advantages of both classes of estimators, two key differences set the MDep apart. First, the MDep exploits identifying variation in instruments of which endogenous covariates can be mean-independent but distributionally dependent, e.g., at some quantile(s) that need not be known or determined. Second, unlike the above ICM and continuum-moment estimators, which require the existence of at least the first two moments of the disturbance for consistency and asymptotic inference, the MDep obviates the existence of any moment of the disturbance.

Some existing works considered IV estimation without excludability by exploiting and modelling non-linear forms of dependence between endogenous and exogenous covariates, e.g., Lewbel (1997), Cragg (1997), Dagenais and Dagenais (1997), Lewbel (1997), Erickson and Whited (2002), Rigobon (2003), Klein and Vella (2010), and Gao and Wang (2023). Unlike the foregoing, the MDep does not require the practitioner to construct moments or model first-stage relationships. It suffices that there be dependence between covariates and instruments that does not need to be known or modelled.

The econometric literature on weak instruments largely focuses on detection and weak-instrument-robust inference (e.g., Staiger and Stock (1997), Stock and Yogo (2005), Andrews, Moreira, and Stock (2006), Kleibergen and Paap (2006), Olea and Pflueger (2013), Andrews and Mikusheva (2016), Sanderson and Windmeijer (2016), and Andrews and Armstrong (2017)) – see Andrews, Stock, and Sun (2019) for a review. Normal distributions of conventional IV estimates can be poor and hypothesis tests based on them can be unreliable when instruments are weak (Nelson and Startz, 1990a; Nelson and Startz, 1990b; Bound, Jaeger, and Baker, 1995). Although the econometric literature on weak instruments does

not focus much on estimation, Hirano and Porter (2015) and Andrews and Armstrong (2017) do, however, constitute exceptions. The MDep gives a new perspective to handling weak IVs in empirical practice; IV- or ICM-irrelevant instruments can be MDep-strong.

By extracting non-linear identifying variation in instruments in order to boost instrument strength, a number of works employ flexible methods such as the non-parametric IV, e.g., Donald and Newey (2001), Newey and Powell (2003), Donald, Imbens, and Newey (2003), Kitamura, Tripathi, and Ahn (2004), and Das (2005), machine learning techniques, e.g., Chen, Chen, and Lewis (2020), and regularisation or moment selection schemes, e.g., Ng and Bai (2009), Darolles, Fan, Florens, and Renault (2011), Belloni, Chen, Chernozhukov, and Hansen (2012), Hansen and Kozbur (2014), and Carrasco and Tchuente (2015). While it is conceivable to take transformations of instruments to extract more identifying variation, this approach may be limited, for example, when available instruments are not monotone in endogenous covariates.[3] Further, the aforementioned approach usually results in high dimensionality, unlike the MDep, which is parsimonious in the elementary instruments Z.

The dCov measure is used primarily to test independence between two random variables of arbitrary dimensions. It is consistent against all types of dependent alternatives, including linear, non-linear, monotone, and non-monotone forms of dependence. Several applications of the dCov measure have emerged since the seminal work Székely, Rizzo, and Bakirov (2007) – see, e.g., Sheng and Yin (2013), Székely, Rizzo, et al. (2014), Shao and Zhang (2014), Park, Shao, Yao, et al. (2015), Su and Zheng (2017), Davis, Matsui, Mikosch, Wan, et al. (2018), and Xu and Chen (2020). Unlike the current paper, this literature is largely concerned with tests of distributional and mean independence. The closest work to the current paper is perhaps Sheng and Yin (2013); it proposes an estimator of single-index models as a tool for sufficient dimension reduction using the dCov as a criterion. The reader is referred to Edelmann, Fokianos, and Pitsillou (2019) for a review.

[3]Consider an outcome $Y = X + U$, endogenous covariate $X = X^* + U$, and instrument $Z = |X^*|$ with $\text{cov}[U, Z] = 0$. For a symmetric mean-zero random variable X^*, $\text{cov}[X, Z] = 0$. It is not feasible, without further information, to transform Z in order to induce correlation with X.

The rest of the paper is organised as follows. Section 2 describes the dCov measure and presents the MDep estimator. Section 3 derives important theoretical results viz. identification, consistency, asymptotic normality, and consistency of the covariance matrix estimator. Section 4 provides an empirical application, and Section 5 concludes. All proofs are relegated to the Appendix. Additional theoretical results are available in the Supplementary Appendix.

Notation: Define $\mathbb{E}_n[\xi_i] \equiv \frac{1}{n}\sum_{i=1}^n \xi_i$ and $\mathbb{E}_n[\xi_{ij}] \equiv \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \xi_{ij}$. For a random variable ξ, let $\xi^\dagger$ denote its independent and identical copy. $\widetilde{\xi} \equiv \xi - \xi^\dagger$ is its symmetrised version, $\widetilde{\xi}_{ij} \equiv \xi_i - \xi_j$, and p_ξ is the dimension of ξ.

2 The MDep Estimator

This section presents illustrative examples highlighting the MDep's key features, the underlying dCov measure, the class of models being considered, and the corresponding objective function.

2.1 Illustrative examples

The MDep estimator has unique strengths relative to existing estimators. To explore these, consider the linear model: $Y = \theta_1 X_1 + \theta_2 X_2 + U$ in the following examples where $\theta_1 = 0.5$ is the parameter of interest and X_1 is endogenous.

Example 2.1 (Non-monotone relationships)**.** *Consider the relationship* $X_1 = Z^* + V$ *and the instrument set* $Z = [\mathrm{I}(|Z^*| < -\Phi(0.25)^{-1}), X_2]$ *where* $Z^* \sigma \mathcal{N}(0,1)$*. There is no possible transformation, without extra information, of the non-monotone relationship* $Z_1 = \mathrm{I}(|Z^*| < -\Phi(0.25)^{-1})$ *that induces mean-dependence or correlation of* Z_1 *with* X_1*.* Z_1 *is thus ICM- and IV-irrelevant, while it is MDep-relevant.*

Generally, it is well-known that $\mathbb{E}[Z|X] \neq \mathbb{E}[Z]$ *a.s.* does not imply (nor is implied by)

$\mathbb{E}[X|Z] \neq \mathbb{E}[X]$ *a.s.* The MDep framework exploits both forms of dependence for identification, while the ICM, conventional IV, and non-parametric IV methods can only exploit the latter.

Example 2.2 (Identification without excludability I - non-monotonicity)**.** *Like in the ICM framework, identification without excludability in the MDep framework is possible as long as no linear combination of covariates is independent of instruments. Consider the setting in Example 2.1 with a single included instrument* $X_2 = Z = 0.2Z^* + Z^{*2}$.

Example 2.3 (Identification without excludability II - skedastic function)**.** *The identifying variation, without excludability in this example comes from the skedastic function of the endogenous covariate on the exogenous covariate:* $X_1 = 0.2Z + V\sqrt{1+Z^2}$, $Z = X_2$, *and* $\mathbb{E}[V|Z] = 0$ *a.s. Since a non-trivial linear combination of* $X \equiv [X_1, X_2]$ *is mean-independent of* $Z = X_2$, *while none is independent of* Z, *the ICM (and conventional IV estimators by construction) cannot exploit this type of exogenous variation for identification, unlike the MDep.*

Example 2.4 (Non-existent first moment of U)**.** *The disturbance,* U, *follows a mixture of the standard normal and Cauchy distributions, namely* $U \sim 0.7\mathcal{N}(0,1) + 0.3\mathcal{C}(0,1)$. *It is well known that conventional IV, non-parametric IV, and ICM estimators are inconsistent when the first moment of* U *does not exist. The MDep, on the other hand, is consistent as it does not require the existence of any moment of* U.

Table 2.1 presents summaries from a limited simulation exercise to highlight the features in the preceding illustrative examples. One observes a robust performance of the MDep across all four examples, whereas the conventional IV, e.g., the 2SLS and the ICM estimator of Tsyawo (2023), namely the MMD does not perform well. Existing methods use identifying variations of mean-dependence of X on Z, which for conventional IV and non-parametric IV methods can induced by taking transformations of Z in order to induce correlation with X.

Table 2.1: Illustrative Examples

	Med-t	MAD	RMSE	Rej.		Med-t	MAD	RMSE	Rej.
	Example 2.1					Example 2.2			
MDep	-0.060	0.059	0.105	0.029		-0.018	0.037	0.066	0.040
MMD	-0.255	0.181	0.637	0.018		-0.080	0.366	18.125	0.003
ESC6	-0.247	0.164	0.292	0.023		-0.124	0.574	7.903	0.000
2SLS	-0.050	0.676	25.938	0.001					
LIML	-0.050	0.676	25.938	0.001					
	Example 2.3					Example 2.4			
MDep	-0.080	0.060	0.099	0.064		-0.050	0.112	0.165	0.030
MMD	-0.240	0.266	8.44	0.010		-0.052	2.467	896.069	0.000
ESC6	-0.306	0.321	30.563	0.015		-0.031	2.881	835.218	0.000

Notes: 1000 random samples, sample size $n = 500$, $[\theta_1, \theta_2] = [0.5, -0.5]$, $U \sim (\chi^2_1 - 1)/\sqrt{2}$, $V = -0.2U + (1 - 0.2^2)^{1/2}\widetilde{V}$, $\widetilde{V} \sim \mathcal{U}[-\sqrt{3}, \sqrt{3}]$. Med-$t$, MAD, RMSE, and Rej. denote the median t-statistic, median absolute deviation, root mean-squared error, and 5% empirical rejection rates of the test $\mathbb{H}_o: \ \theta_1 = 0.5$.

Example 2.1 shows that this is not always feasible when Z is a non-monotone transformation of X.

2.2 The dCov measure

Section 2.1 showcases certain remarkable features of the MDep via illustrative examples. It remains to explore the mechanics of the MDep in order to explain why the features hold. An appropriate starting point is a brief presentation of the distance covariance (dCov) measure of Székely, Rizzo, and Bakirov (2007) on which the MDep is based. The dCov measures the dependence between two random variables of arbitrary dimension. Let $\varphi_{\Upsilon,Z}(t,s)$ denote the joint characteristic function of $[\Upsilon, Z]$, and $[\varphi_\Upsilon(t), \varphi_Z(s)]$ denote their marginal characteristic functions.

Definition 2.1. *The square of the distance covariance between random variables* Υ *and* Z

with finite first moments is defined by

$$
\begin{aligned}
\mathcal{V}^2(\Upsilon, Z) &\equiv |\varphi_{\Upsilon,Z}(t,s) - \varphi_{\Upsilon}(t)\varphi_Z(s)|_w^2 \\
&= \int |\varphi_{\Upsilon,Z}(t,s) - \varphi_{\Upsilon}(t)\varphi_Z(s)|^2 w(t,s)dtds \\
&= \int \left|\mathbb{E}\big[\exp(\iota(t'\Upsilon + s'Z))\big] - \mathbb{E}\big[\exp(\iota t'\Upsilon)\big]\mathbb{E}\big[\exp(\iota s'Z)\big]\right|^2 w(t,s)dtds
\end{aligned} \tag{2.1}
$$

where $\iota = \sqrt{-1}$ *and the weight* $w(t,s)$ *is an arbitrary positive function for which the integration exists.* $|\zeta|^2 = \zeta\bar{\zeta}$*, where* $\bar{\zeta}$ *is the complex conjugate of* ζ*.*

Using the weight function $w(t,s) = (c_{p_\Upsilon} c_{p_Z} ||t||^{1+p_\Upsilon} ||s||^{1+p_Z})^{-1}$ where $c_p = \frac{\pi^{(1+p)/2}}{\Gamma((1+p)/2)}$, $p \geq 1$, $||\cdot||$ is the Euclidean norm, and $\Gamma(\cdot)$ is the complete gamma function, Székely, Rizzo, and Bakirov (2007) obtains a distance covariance measure, which can be shown to be equivalent to $\mathcal{V}^2(\Upsilon, Z) = \mathbb{E}[\mathcal{Z}||\Upsilon - \Upsilon^\dagger||]$ where $\mathcal{Z} \equiv h(Z, Z^\dagger)$, $Z^\dagger$ denotes an independent and identical copy of Z, and $h(z_a, z_b) \equiv ||z_a - z_b|| - \mathbb{E}[||z_a - Z|| + ||Z - z_b||] + \mathbb{E}[||Z - Z^\dagger||]$. From (2.1), one observes that $|\varphi_{U,Z}(t,s) - \varphi_U(t)\varphi_Z(s)|_w^2 \geq 0$.

This paper follows Székely, Rizzo, et al. (2014) in using the following algebraically equivalent form of the unbiased estimator of the distance covariance measure

$$
\mathcal{V}_n^2(\Upsilon, Z) \equiv \frac{1}{n(n-3)} \sum_{i \neq j} \mathcal{Z}_{ij,n} ||\widetilde{\Upsilon}_{ij}|| \tag{2.2}
$$

where $\mathcal{Z}_{ij,n} \equiv h_n(Z_i, Z_j)$ such that

$$
h_n(Z_i, Z_j) = \begin{cases} ||\widetilde{Z}_{ij}|| - \frac{1}{n-2}\sum_{l=1}^n (||\widetilde{Z}_{il}|| + ||\widetilde{Z}_{lj}||) + \frac{1}{(n-1)(n-2)} \sum_{k,l} ||\widetilde{Z}_{kl}||, & i \neq j \\ 0, & i = j. \end{cases}
$$

Székely, Rizzo, and Bakirov's (2007) weight function $w(t,s) = (c_{p_\Upsilon} c_{p_Z} ||t||^{1+p_\Upsilon} ||s||^{1+p_Z})^{-1}$, besides yielding a reliable measure of dependence, results in a computationally tractable measure, which does not require numerical integration, obviates the choice of smoothing parameters (e.g., bandwidth or number of approximating terms in non-parametric approaches), and

admits multiple instruments Z. While the empirical analogue of $|\varphi_{\Upsilon,Z}(t,s)-\varphi_{\Upsilon}(t)\varphi_{Z}(s)|^2_{w,n}$ of the dCov measure is natural and intuitive as a non-negative measure of distributional dependence, the form $\mathcal{V}^2_n(\Upsilon,Z)$ is computationally tractable. The MDep estimator proposed in this paper is based on the latter.

For the estimator presented in this paper, the simplified formulation (2.2) has two main advantages: the permutation symmetry of $\mathcal{Z}_{ij,n}$, i.e., $\mathcal{Z}_{ij,n}=\mathcal{Z}_{ji,n}$, simplifies the application of U-statistic theory in the proof of asymptotic normality, and lessens computational time in evaluating (2.2). The Székely, Rizzo, and Bakirov (2007) integrating measure is not the only possibility – see Davis, Matsui, Mikosch, Wan, et al. (2018) for other possible integrating measures. From an estimation and inference point of view, this paper's choice is detailed later in Remark 2.1.

For ease of reference, the properties of the dCov measure in Székely, Rizzo, and Bakirov (2007) and Székely and Rizzo (2009) are stated below.

Properties of dCov. *The following properties hold for the distance covariance measure under the condition* $\mathbb{E}[||\Upsilon||+||Z||]<\infty$*:*

(a) $\mathcal{V}^2(\Upsilon,Z)\geq 0$*;*

(b) $\mathcal{V}^2(\Upsilon,Z)=0$ *if and only if* Υ *is independent of* Z*;*

(c) $\mathcal{V}^2(\Upsilon,Z)=\mathbb{E}[\mathcal{Z}||\Upsilon-\Upsilon^{\dagger}||]$*; and*

(d) $\mathbb{E}[\mathcal{V}^2_n(\Upsilon,Z)]=\mathcal{V}^2(\Upsilon,Z)$ *for* $n>3$ *and i.i.d. samples.*

The properties are provided and proved in the following papers: Property (a) in Székely and Rizzo (2009, Theorem 4 (i)), Property (b) in Székely, Rizzo, and Bakirov (2007), and Property (d) in Székely, Rizzo, et al. (2014, Proposition 1).

2.3 Model specification

For considerations of space and scope, the current paper is limited to regression models, which assume the outcome Y_i is generated as $Y_i = G\big(\theta_{o,c} + g(X_i\theta_o) + U_i\big)$, where $G(\cdot)$ is a known invertible function, $g(\cdot)$ is a known differentiable function with unknown parameter vector $\theta_o \in \mathbb{R}^{p_\theta}$, and $\theta_{o,c}$ is the intercept. Observed data $\{[Y_i, X_i, Z_i]\}_{i=1}^n$ include the outcome Y_i, a vector of covariates X_i, and a set of instruments Z_i – see Jurecková and Procházka (1994) for a similar class of models.

By the invertibility of $G(\cdot)$, the parameterised disturbance function can be written as $U_i(\theta) = G^{-1}(Y_i) - g(X_i\theta)$. Like Honoré and Powell (1994), the intercept is ignored as it is differenced out in the objective function (2.3). The dependence of $U_i(\theta)$ on covariates X_i is suppressed for notational convenience. While the class of models under consideration includes interesting examples such as the linear model $U_i(\theta) = Y_i - X_i\theta$, non-linear parametric models, e.g., $U_i(\theta) = Y_i - \exp(X_i\theta)$, fractional response models, e.g., $U_i(\theta) = \log(Y_i/(1-Y_i)) - X_i\theta$, and special cases of Box-Cox models e.g., $U_i(\theta) = \log(Y_i) - X_i\theta$, it excludes equally interesting ones such as binary response models and quantile regression.

2.4 The objective function

The objective function is based on the squared distance covariance of the parameterised disturbance $U(\theta)$ and Z. The MDep estimator is defined as

$$\widehat{\theta}_n = \arg\min_{\theta\in\Theta} \mathcal{V}_n^2(U(\theta), Z).$$

Define $\{W_i : 1 \le i \le n\}$, $W_i \equiv [Y_i, X_i, Z_i]$ on a probability space $(\mathcal{W}, \mathscr{W}, \mathbb{P})$. Following Huber (1967), the normalised minimand is

$$Q_n(\theta) \equiv \frac{1}{n(n-3)} \sum_{i\neq j} q_n(W_i, W_j; \theta), \tag{2.3}$$

where $q_n(W_i, W_i; \theta) \equiv \mathcal{Z}_{ij,n}(|\widetilde{U}_{ij}(\theta)| - |\widetilde{U}_{ij}|)$. Normalising the minimand avoids unnecessary moment conditions on U – see, e.g., Powell (1991) and Oberhofer and Haupt (2016). This holds even though the dCov measure itself requires the existence of the first moment of U. An additional advantage to using the normalised minimand $Q_n(\theta)$ is that by Property (d), it is unbiased for $Q(\theta) \equiv \mathbb{E}[q(W, W^\dagger; \theta)]$ where $q(W, W^\dagger; \theta) \equiv \mathcal{Z}(|\widetilde{U}(\theta)| - |\widetilde{U}|)$.

Remark 2.1. *The preceding discussion also highlights a unique advantage to using the Székely, Rizzo, and Bakirov (2007) integrating measure – moments of U need not exist thus rendering the MDep robust to contamination of the disturbance. Arbitrary choices of valid integrating measures do not deliver such a robustness feature by construction.*

(2.3) is a sum of specially weighted absolute pairwise differences in parameterised disturbances. As the objective function is based on a U-statistic, the MDep is thus related to the estimators of Honoré and Powell (1994), Honoré and Powell (2005), and Jochmans (2013). Observe from the objective function (2.3) that it is based on symmetrised disturbance function $\widetilde{U}_{ij}(\theta)$. The idea of symmetrising the disturbance term is not new – see, e.g., Honoré and Powell (1994). This paper, however, appears to be the first to propose an estimator in the presence of endogeneity with remarkable features as outlined in Section 2.1. (2.3) suggests that the asymptotic behaviour of the MDep estimator is akin to that of quantile estimators. A key difference, however, is that (2.3) is non-convex as $\mathcal{Z}_{ij,n}$ is not always non-negative.

It is instructive to note how the MDep relates to moment estimators, which include conventional IV methods such as ordinary least squares (OLS), IV, 2SLS, method of moments (MM), and GMM and ICM estimators, e.g., Domínguez and Lobato (2004), Antoine and Lavergne (2014), Escanciano (2018), and Tsyawo (2023). A moment estimator typically minimises the (weighted) linear dependence between $U(\theta)$ and Z. ICM estimators minimise the mean dependence of $U(\theta)$ on Z. The MDep minimises the (distributional) dependence between $U(\theta)$ and Z. As distributional independence is the strongest form of independence between variables, the MDep thus minimises a more general form of dependence between $U(\theta)$ and Z.

3 Asymptotic Theory

One notices from (2.3) that the asymptotic theory for the MDep falls within the category of estimators of models with U-statistics-based objective functions, e.g., Honoré and Powell (1994), non-smooth objective functions such as quantile regression (QR) methods, e.g., Koenker and Bassett Jr (1978), Powell (1991), and Oberhofer and Haupt (2016), and instrumental variable QR methods, e.g., Chernozhukov and Hansen (2006) and Chernozhukov and Hansen (2008), and the CF approach to QR of Lee (2007). Define the Jacobian $X_i^g(\theta)$ where $X_i^g(\theta) \equiv \frac{\partial U_i(\theta)}{\partial \theta'} = -g'(X_i\theta)X_i$, the symmetrised Jacobian $\widetilde{X}_{ij}^g(\theta) \equiv X_i^g(\theta) - X_j^g(\theta)$, and the parameter vector θ_o. Also, let $\widetilde{X^g}(\theta) \equiv X^g(\theta) - X^{g\dagger}(\theta)$ and $X^g(\theta)$ be any random variables defined on the supports of $\widetilde{X}_{ij}^g(\theta)$ and $X_i^g(\theta)$, respectively.

3.1 Regularity Conditions

Two sets of regularity conditions imposed in the paper guarantee the consistency of the MDep estimator $\widehat{\theta}_n$. The first set comprises sampling, smoothing, and dominance conditions which ensure that the difference between the normalised minimand and its expectation converges to zero uniformly in $\theta \in \Theta$.

Assumption 1 (Regularity)**.**

(a) $\{W_i : 1 \leq i \leq n\}$ *are independently and identically (iid) distributed random vectors, and the outcome is generated as* $Y_i = G\big(\theta_{o,c} + g(X_i\theta_o) + U_i\big)$ *where* $G(\cdot)$ *is a known invertible function and* $g(\cdot)$ *is a known continuously differentiable function.*

(b) $U(\theta)$ *is measurable in* $[U, X]$ *for all* θ *and is twice differentiable in* θ *for all* $[U, X]$ *in the support of* $[U_i, X_i]$*. For all* θ*,* $X^g(\theta) \equiv -g'(X\theta)X$ *is measurable in* X *for all* θ *and* $g'(X\theta) \neq 0$ *a.s.*

(c) $\mathbb{E}\big[\sup_{\theta\in\Theta} ||\max\{|\mathcal{Z}|, 1\}\widetilde{X^g}(\theta)||^4\big] \leq C$.

(d) The parameter space Θ *is compact.*

Assumptions1(a) to (d) above suffice for the application of the uniform law of large numbers in Theorem 1 of Honoré and Powell (1994). The condition on the data generating process in Assumption 1(a) and the differentiability requirement in Assumption 1(b) characterise the class of models considered in this paper, e.g., the linear model. They exclude models such as Koenker and Bassett Jr's (1978) QR and binary response models.[4] $\widetilde{U}(\theta) = \widetilde{U} + \widetilde{X^g}(\bar{\theta})(\theta - \theta_o)$ for some $\bar{\theta}$ between θ and θ_o is a useful expression for subsequent analyses thanks to Assumption 1(b) and the mean-value theorem (MVT). Observe that the technical requirement $g'(X\theta) \neq 0$ $a.s.$ is important for identification as the expression $\widetilde{U}(\theta) = \widetilde{U} + \widetilde{X^g}(\bar{\theta})(\theta - \theta_o)$ shows that $\widetilde{U}(\theta)$ can equal $\widetilde{U}$ $a.s.$ for some $\theta \neq \theta_o$ if it is violated.

Assumption 1(c) is an MDep analogue of a standard condition that uniformly bounds the fourth moment of both the weighted and unweighted Jacobian, i.e., $\mathbb{E}\big[\sup_{\theta\in\Theta} ||\max\{|\mathcal{Z}|, 1\}\widetilde{X^g}(\theta)||^4\big] \leq C$ implies both $\mathbb{E}\big[\sup_{\theta\in\Theta} ||\mathcal{Z}\widetilde{X^g}(\theta)||^4\big] \leq C$ and $\mathbb{E}\big[\sup_{\theta\in\Theta} ||\widetilde{X^g}(\theta)||^4\big] \leq C$. Assumption 1(c) can be further weakened by replacing Z with bounded one-to-one mappings, e.g., $\text{atan}(Z)$ such that Z and the mapping generate the same Euclidean Borel field – see Bierens (1982, p. 108) and Székely, Rizzo, and Bakirov (2007, Remark 1), thereby allowing Z (in addition to U) to have no integer moments. In that case, Assumption 1(c) can simply be replaced with $\mathbb{E}\big[\sup_{\theta\in\Theta} ||\widetilde{X^g}(\theta)||^4\big] \leq C$. It can be shown (see Lemma A.1) that under Assumption 1(c), $\mathcal{B}(X, X^\dagger, Z, Z^\dagger) \equiv \sup_{\theta\in\Theta} ||\mathcal{Z}\widetilde{X^g}(\theta)||$, which does not depend on $[U, U^\dagger]$, is integrable, and $|q(W, W^\dagger; \theta)| \leq \mathcal{B}(X, X^\dagger, Z, Z^\dagger)||\theta - \theta_o||$ a.s. for all $\theta \in \Theta$ where $\mathbb{E}[q(W, W^\dagger; \theta)] = \mathbb{E}[q_n(W_i, W_j; \theta)]$ $i \neq j$ under Assumption 1(a). Assumption 1(d) is required as the objective function is non-convex. From Assumption 1(d) and the foregoing, the expected value of $q_n(W_i, W_j; \theta)$ exists even if the expected value of U does not exist.

Remark 3.1. *Although regression quantiles do not fall in the category of estimators considered, the MDep shares the "robustness" feature of quantile estimators over the class of models in Assumption 1(a) where asymptotic properties do not depend on the existence of moments of the disturbance U. Further, by taking bounded one-to-one mappings of Z, only*

[4] The differentiability requirement is for ease of treatment as the case of non-smooth objective functions is beyond the scope of the current paper.

moments of $\widetilde{X}^g(\theta)$ $\theta \in \Theta$ *are required to exist.*

The remaining set of regularity conditions for consistency (Assumptions 2(a) and 2(b)) are identification conditions which ensure that the expected value of the minimand is uniquely minimised.

Assumption 2 (Identification)**.**

(a) $\text{med}[U - U^\dagger | X, X^\dagger, Z, Z^\dagger] = 0$.

(b) $\tau \neq 0$ *implies* $X\tau \not\perp Z$.

These assumptions are perhaps the most important because one easily sees the advantage the MDep estimator has vis-à-vis conventional IV methods. The relevance condition (Assumption 2(b)), relative to IV and ICM methods, is much weaker.

Assumption 2(a) requires that the median of $U - U^\dagger$ conditional on $[X, X^\dagger, Z, Z^\dagger]$ be zero. Compared to exogeneity conditions imposed on levels, e.g., Escanciano (2018, eqn. 1.2), Assumption 2(a) is imposed on pairwise differences. One can glean a similar exogeneity condition based on symmetrised disturbances from the first-order condition in Honoré and Powell (1994, p. 245) – see also Honoré and Powell (2005) and Jochmans (2013). An advantage to symmetrising U is that it renders the location (which coincides with the median of $\widetilde{U}$) zero by construction and a natural point on which to impose an exogeneity condition. Under the stronger assumption that $\mathbb{E}[U]$ exists, $\text{med}[\widetilde{U}]$ coincides with $\mathbb{E}[\widetilde{U}]$. Further, $\text{med}[\widetilde{U}] = \mathbb{E}[\widetilde{U}] = 0$ coincides with the mode if $\widetilde{U}$ is unimodal. This point is crucial, especially in structural economic models, where the exogeneity of Z given X can follow from economic theory but the function of the conditional distribution of U where the exogeneity condition is imposed and the existence thereof typically does not emanate from economic theory. Assumption 2(a) can therefore be perceived as meaningfully addressing this concern.

Unlike Assumption 2(a) which is imposed on pairwise differences in disturbances, similar exclusion restrictions on conditional quantiles are imposed on the levels of disturbances for quantile estimators under (possible) endogeneity, see e.g., Chernozhukov and

Hansen (2006, Assumption A.2), Lee (2007, Assumption 3.6), and Powell (1991, Assumption B2). Specific to the median of U, Lee (2007), for example, requires that the median of U, conditional on $[X, Z]$ be constant. Noting that Assumption 2(a) can be expressed as $\mathbb{E}[\mathrm{I}(\widetilde{U} \leq 0) - 0.5|X, X^{\dagger}, Z, Z^{\dagger}] = 0$ *a.s.* is testable using, e.g., integrated conditional moment (ICM) specification tests – see Bierens (1982), Domínguez and Lobato (2015), Su and Zheng (2017), Xu and Chen (2020), and Jiang and Tsyawo (2022).[5]

Assumption 2(b) is the condition of non-independence between non-trivial linear combinations of X, namely $X\tau, \tau \neq 0$ and Z; it is the MDep analogue of the relevance condition in the IV/CF setting (see, e.g., Wooldridge (2010, Assumption 2SLS.2(b))) and an MDep analogue of the linear completeness condition in ICM estimators, e.g., Escanciano (2018) and Tsyawo (2023). In the IV setting, the relevance condition requires that no non-zero linear combination of X be uncorrelated with Z. The ICM linear completeness condition requires that no non-zero linear combination of X be mean-independent of Z. Assumption 2(b) requires that no non-zero linear combination X be independent of Z. As independence implies mean independence which in turn implies uncorrelatedness, it is obvious that the MDep relevance condition Assumption 2(b) is the weakest possible. In a simple case with a univariate X, Assumption 2(b) allows X to be uncorrelated, or mean-independent of Z as long as X is not independent of Z. All IV-strong or ICM-strong Z are MDep-strong by construction. The converse is, however, not true. Like in the case of ICM estimators, Assumption 2(b) can hold even if there are fewer instruments than covariates – see, e.g., Tsyawo (2023). This feature of the MDep can be explored to attain identification without excludability – see Examples 2.2 and 2.3.

Remark 3.2. *MDep admits the largest set of instruments possible. In addition to IV- and ICM-relevant instruments that the MDep admits by construction, MDep admits instruments that can be IV- or ICM-irrelevant but non-independent of covariates such that Assumption 2(b) holds.*

[5]This task, however, is left for future work due to considerations of scope and space.

3.2 Parameters of interest and interpretation

When the conditional mean function $\mathbb{E}[Y|X]$ exists, the MDep estimate has a standard interpretation as the partial effect can be obtained with respect to $\mathbb{E}[Y|X]$. For ease of exposition, assume the simple linear model $Y = X\theta + U$ without endogeneity, i.e., $Z = X$ and $p_X = 1$. Under Assumption 2(a), $\mathbb{E}[\widetilde{Y}|\widetilde{X}] = \widetilde{X}\theta_o \Leftrightarrow \mathbb{E}[Y|\widetilde{X}] = X\theta_o - (\mathbb{E}[Y^\dagger|\widetilde{X}] - X^\dagger\theta_o)$ is identified, thus the partial effect of X with respect to Y is simply θ_o.[6] If the first moment of U does not exist, the above interpretation no longer holds. In that case, what is identified is $\text{med}[Y - Y^\dagger|X, X^\dagger] = (X - X^\dagger)\theta_o$, i.e., the median difference in outcomes conditional on $X, X^\dagger$. This conditional median of differences is attributable to differences in observed characteristics. For two observationally equivalent economic agents, $\text{med}[Y - Y^\dagger|X, X^\dagger] = 0$. Using treatment effect terminology, θ_o is the median improvement in the outcome from treatment relative to that of an observationally equivalent untreated agent.

Unlike the simple linear example above, the partial effect of X is not constant when $g(X\theta)$ is non-linear in X. As Assumption 1(c) implies the existence of the first moment of the partial effect of X on $\text{med}[Y - Y^\dagger|X, X^\dagger]$, interesting summaries of the heterogeneous partial effects such as averages or medians can be reported. The MDep is a robust estimator of the conditional mean function under the assumption of the existence of the conditional mean. In the presence of contamination from outliers or random noise with non-existent moments, the MDep $\widehat{\theta}_n$ remains consistent for θ_o.

3.3 Identification and Consistency

The MDep objective function (2.3) is non-convex because $\mathcal{Z}_{ij,n}$ is not non-negative. This makes typical QR identification proof techniques based on the convexity of the objective function – e.g., Koenker and Bassett Jr (1978), Powell (1991), Honoré and Powell (1994), and Oberhofer and Haupt (2016) – inapplicable to the MDep. The proof of identification is thus non-trivial. This paper leverages the non-negativity and "omnibus" properties of the

[6] That of $X^\dagger$ is zero as $X^\dagger$ is independent of Y (Assumption 1(a)).

dCov measure (Property (a) and (b), respectively) to establish identification and consistency.

Theorem 1 (Identification)**.** *Suppose Assumptions 1 and 2 hold, then (a) for any* $\varepsilon > 0$*, there exists a* $\delta_\varepsilon > 0$ *such that* $\inf_{\{\theta\in\Theta:||\theta-\theta_o||\geq\varepsilon\}} Q(\theta) > \delta_\varepsilon$ *and (b)* $\widehat{\theta}_n \xrightarrow{p} \theta_o$*.*

Theorem 1(a) shows that under the given assumptions, the minimand $Q(\theta)$ has a unique minimum. Theorem 1(b) provides the consistency result for the MDep.

3.4 Asymptotic normality

Define the score function $\mathcal{S}_n(\theta) \equiv \mathbb{E}_n[\psi(W_i, W_j; \theta)]$ where $\psi(W_i, W_j; \theta) \equiv \mathcal{Z}_{ij}\big(1-2\mathrm{I}(\widetilde{U}_{ij}(\theta) \leq 0)\big)\widetilde{X}^g_{ij}(\theta)'$, with $\psi(W_i, W_j) \equiv \psi(W_i, W_j; \theta_o)$. $\mathcal{Z}_{ij} = \mathcal{Z}_{ji}$ and $\big(1 - 2\mathrm{I}(\widetilde{U}_{ij} \leq 0)\big)\widetilde{X}^g_{ij} = \big(1 - 2\mathrm{I}(\widetilde{U}_{ji} \leq 0)\big)\widetilde{X}^g_{ji}$ with $\widetilde{X}^g_{ij} \equiv \widetilde{X}^g_{ij}(\theta_o)$ hence $\psi(\cdot,\cdot)$ is permutation symmetric. Denote the cumulative distribution function and the probability density functions of $\widetilde{U}$ conditional on $[X, X^\dagger, Z, Z^\dagger]$ by $F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\cdot)$ and $f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\cdot)$, respectively. Further, define $v_n(\theta) \equiv \sqrt{n}\big(\mathcal{S}_n(\theta) - \mathcal{S}(\theta)\big)$, $\psi^{(1)}(W_i) \equiv \mathbb{E}[\psi(W_i, W_j)|W_i]$, and $\mathcal{H} := 2\mathbb{E}[f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(0)\mathcal{Z}\widetilde{X^g}'\widetilde{X^g}]$. Finally, let $\partial^-|\hat{q}|$ and $\partial^+|\hat{q}|$, respectively, denote the left- and right-derivatives of $|q|$ with respect to q at $q = \hat{q}$.

Assumption 3 (Asymptotic Linearity of $\widehat{\theta}_n$)**.**

(a) θ_o *is an interior point of* Θ*;*

(b) In an open neighbourhood Θ_o *of* θ_o*,* $\sup_{\theta\in\Theta_o} \sum_{i\neq j}^{n} \mathrm{I}\big(\partial^-|\widetilde{U}_{ij}(\theta)| \neq \partial^+|\widetilde{U}_{ij}(\theta)|\big) = \mathcal{O}_p(1)$*;*

(c) $F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\cdot)$ *is differentiable with density* $f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\cdot)$*,* $f_o^{-1} < f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\epsilon) \leq \sup_{e\in\mathbb{R}} f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(e) \leq f_o^{1/4}$ *a.s., and* $|f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\epsilon_1) - f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\epsilon_2)| \leq f_o^{1/4}|\epsilon_1 - \epsilon_2|$ *a.s. for all* $\epsilon, \epsilon_1, \epsilon_2$ *in a neighbourhood of zero and a positive constant* $f_o < \infty$*;*

(d) $\mathbb{E}\big[||\widetilde{Z}||^4\big] \leq C$*;*

(e) In some open neighbourhood Θ_o *of* θ_o*,* $\sup_{\theta\in\Theta_o} \dfrac{||v_n(\theta) - v_n(\theta_o)||}{1 + \sqrt{n}||\mathcal{S}(\theta)||} = o_p(1)$*; and*

(f) $\mathcal{H}$ *is non-singular.*

Assumption 3(a) is a standard assumption. Assumption 3(b) is useful in ensuring that the normalised sub-gradient $\sqrt{n}\mathcal{S}_n(\theta)$ tends to zero when evaluated at $\widehat{\theta}_n$ – cf. Honoré and Powell (1994, Assumption F3). Assumption 3(c) is a standard assumption for regression quantile estimators – see, e.g., Lee (2007, Assumption 3.6), Chernozhukov and Hansen (2006, Assumption 2 R.4), Chernozhukov and Hansen (2008, Assumption R.4), Powell (1991, Assumption C4. (i) and (ii)), Oberhofer and Haupt (2016, Assumption A.14)), and Xu and He (2021, Condition D.1). It ensures the Hessian is well-defined. Assumption 3(d) is useful in establishing the almost sure convergence of $|\mathcal{Z}_{ij,n} - Z_{ij}|$ to zero for any $i, j = 1, \ldots, n$ – see Lemma E.1. The motivation for Lemma E.1 is primarily computational and theoretical – one is able to base theoretical arguments of the MDep on U-statistics of order 2 instead of order 4 as given in Székely, Rizzo, et al. (2014). Assumption 3(e) is a stochastic equi-continuity condition, which is crucial in establishing the asymptotic linearity and asymptotic normality of estimators based on non-smooth moments – see, e.g., Honoré and Powell (1994), Pollard (1985), Pakes and Pollard (1989), and Cheng and Liao (2015). It is verified in Lemma E.4. As $\mathcal{Z}$ is not a non-negative random variable, the Hessian $\mathcal{H} = 2\mathbb{E}[f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(0)\mathcal{Z}\widetilde{X^g}'\widetilde{X^g}]$ cannot be positive definite by construction; non-singularity (Assumption 3(f)) is thus imposed – cf. Honoré and Powell (1994, Assumption N2). Define $\Omega \equiv 4\mathbb{E}[\psi^{(1)}(W)\psi^{(1)}(W)']$. The following theorem states the asymptotic linearity and asymptotic normality of the MDep estimator.

Theorem 2. *The MDep* $\widehat{\theta}_n$

(a) satisfies the asymptotic linearity condition

$$\sqrt{n}(\widehat{\theta}_n - \theta_o) = -\mathcal{H}^{-1}\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\psi^{(1)}(W_i) + o_p(1), \text{ and}$$

(b) is asymptotically normal

$$\sqrt{n}(\widehat{\theta}_n - \theta_o) \xrightarrow{d} \mathcal{N}(0, \mathcal{H}^{-1}\Omega\mathcal{H}^{-1}).$$

3.5 Consistent covariance matrix estimation

The preceding subsection presents the asymptotic normality of the MDep estimator. This subsection provides the covariance matrix estimator and proves its consistency. Consistency of the covariance estimator is essential for statistical inference procedures viz. Wald tests, and confidence intervals. Define $\widehat{\psi}^{(1)}(W_i) \equiv \dfrac{1}{n-1}\sum_{j=1}^{n}\widehat{\psi}(W_i, W_j)$ where $\widehat{\psi}(W_i, W_j) \equiv \mathcal{Z}_{ij,n}\big(1 - 2\mathrm{I}(\widetilde{U}_{ij}(\widehat{\theta}_n) < 0)\big)\widetilde{X}^g_{ij}(\widehat{\theta}_n)'$. The estimators of Ω and $\mathcal{H}$ are given by

$$\widehat{\Omega}_n = 4\mathbb{E}_n[\widehat{\psi}^{(1)}(W_i)\widehat{\psi}^{(1)}(W_i)']$$

and

$$\widehat{\mathcal{H}}_n = \frac{1}{n^2\hat{c}_n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\mathrm{I}(|\widetilde{U}_{ij}(\widehat{\theta}_n)| \leq \hat{c}_n)\mathcal{Z}_{ij,n}\widetilde{X}^g_{ij}(\widehat{\theta}_n)'\widetilde{X}^g_{ij}(\widehat{\theta}_n)\right\}$$

respectively where $\hat{c}_n$ is a (possibly random) bandwidth sequence and the uniform kernel, as proposed by Powell (1991), replaces the conditional density in $\mathcal{H}$. The estimator of the covariance matrix is $\widehat{\mathcal{H}}_n^{-1}\widehat{\Omega}_n\widehat{\mathcal{H}}_n^{-1}$. An additional condition is imposed on the bandwidth sequence $\hat{c}_n$ to ensure the consistency of $\widehat{\mathcal{H}}_n$.

Assumption 4. *For some non-stochastic sequence c_n with $c_n \to 0$ and $\sqrt{n}c_n \to \infty$,* $\operatorname{plim}_{n\to\infty}(\hat{c}_n/c_n) = 1$.

Assumption 4 is Powell (1991, Assumption D1) for proving the consistency of $\widehat{\mathcal{H}}_n$; it means $\hat{c}_n$ satisfies the rate conditions $\hat{c}_n = o_p(1)$ and $\hat{c}_n^{-1} = o_p(\sqrt{n})$. The following theorem states the consistency of the covariance matrix estimator.

Theorem 3. *Under Assumptions 1 to 4,* $\operatorname{plim}_{n\to\infty}\widehat{\mathcal{H}}_n^{-1}\widehat{\Omega}_n\widehat{\mathcal{H}}_n^{-1} = \mathcal{H}^{-1}\Omega\mathcal{H}^{-1}$.

Estimating the asymptotic covariance matrix involves specifying the bandwidth $\hat{c}_n$. This paper follows Koenker (2005, Sect. 3.4.2) in using the bandwidth sequence $\hat{c}_n = \widehat{k}(\Phi^{-1}(0.5+\eta_n) - \Phi^{-1}(0.5-\eta_n))$ where $\eta_n = n^{-1/3}\Phi^{-1}(0.975)^{2/3}\left(\frac{3}{4\pi}\right)^{1/3}$ is the Hall and Sheather (1988) bandwidth sequence, $\widehat{k}_n = \min\{\widehat{\sigma}_{\widetilde{U}}, \text{IQR}(\widetilde{\widehat{U}})\}/1.34$ is a robust estimate of scale, $\widehat{\sigma}_{\widetilde{U}}$ and $\text{IQR}(\widetilde{\widehat{U}})$ are, respectively, the sample standard deviation and inter-quartile range of pairwise differenced residuals $\{\widetilde{\widehat{U}}_{i,j} \equiv \widehat{U}_i - \widehat{U}_j, 1 \leq i, j \leq n\}$.

3.6 Testing the MDep relevance condition

The weak relevance condition (Assumption 2(b)) makes the MDep a very powerful tool in a practitioner's toolkit especially in dealing with unavailable or weak instruments. The practical usefulness of the MDep thus crucially lies in its testability. This subsection demonstrates the testability of the MDep relevance condition (Assumption 2(b)) in the single-endogenous-covariate setting.[7] Partition X into $X = [D, \dot{X}]$ where D is the univariate endogenous covariate. Define $\mathcal{E}(\gamma) \equiv D - \dot{X}\gamma$. The following theorem shows the testability of the MDep relevance condition.

Theorem 4. *Suppose Assumption 1(c) holds, then a test of Assumption 2(b) in the presence of a single endogenous covariate D can be formulated via the following hypotheses:*

$$\mathbb{H}_o : \mathcal{E}(\gamma^*) \perp\!\!\!\perp Z \text{ for some } \gamma^* \text{ in a compact subset of } \mathbb{R}^{p_x - 1};$$
$$\mathbb{H}_a : \mathcal{E}(\gamma) \not\perp\!\!\!\perp Z \text{ for all } \gamma \text{ in a compact subset of } \mathbb{R}^{p_x - 1}.$$

Theorem 4 suggests that Assumption 2(b) is testable using a test of independence between MDep regression residuals and a set of instruments, e.g., Sen and Sen (2014), Davis, Matsui, Mikosch, Wan, et al. (2018), and Xu and He (2021).

[7] The extension to multiple endogenous covariates is out of the scope of the current paper.

4 Empirical Example - Demand for fish

This section uses an empirical example to illustrate the usefulness of the MDep estimator in real-world data settings. The application considers a simple demand estimation problem using data from Graddy (1995). The instrument is IV-strong and hence MDep-strong by construction, in a small sample setting of $n = 111$ observations. Thus, the example illustrates the agreement between the conventional OLS/IV and the MDep while illustrating the ability of the MDep to provide more precise estimates thanks to its more robust features discussed in the paper.

The parameter of interest is the elasticity of demand for fish in a (log)-linear demand model with additive disturbance. The disturbance expressed in terms of parameters is specified as

$$U(\theta) = \ln(Q_p) - \theta_c - \theta_1 \ln(P) - \theta_{-1} W$$

where Q_p is the total quantity of fish sold per day, P is the daily average price, θ_1 is the price elasticity of demand, and W is a vector of day dummies – see Graddy (1995) for more details. In a typical demand estimation framework, price P is endogenous. Graddy (1995) proposes an instrument *Stormy* which measures weather conditions at sea. Specifically, *Stormy* is a dummy variable which indicates whether wave height is greater than 4.5 ft and wind speed is greater than 18 knots.

The upper panel of Table 4.1 presents coefficients and standard errors while specification and relevance tests are presented in the lower panel. The tests comprise the MDep relevance test following Theorem 4, and the specification tests of Xu and He (2021) for the MDep and Su and Zheng (2017) for the OLS/IV. The null hypothesis of the Xu and He (2021) test is $U \perp\!\!\!\perp Z$ while that of the Su and Zheng (2017) is $\mathbb{E}[U|Z] = 0$ *a.s.* The MDep relevance p-values suggest a strong and statistically significant distributional dependence between price and the instrument *Stormy*, whereas the p-values of the specification tests suggest failure to reject a null hypothesis of correct model specification. The rk Wald F-statistic of Kleibergen

Table 4.1: Estimates - Fish Demand Function - Graddy (1995)

	MDep (1)	OLS (1)	MDep (2)	IV (2)	MDep (3)	OLS (3)	MDep (4)	IV (4)
$\ln(P)$	-0.558	-0.541	-1.105	-1.082	-0.453	-0.563	-1.231	-1.119
	(0.103)	(0.168)	(0.265)	(0.484)	(0.098)	(0.158)	(0.405)	(0.460)
Day Dummies					✓	✓	✓	✓
Excluded Instr.			✓	✓			✓	✓
Relevance								
MDep.Relv			0.001				0.000	
KB F-Stat.				20.663				22.929
Specification								
T_n p-value	0.751	0.406	0.720	0.892	0.316	0.952	0.471	0.793

Notes: The number of observations in each specification is 111. For each specification, standard errors (in parentheses) are heteroskedasticity-robust. 999 bootstrap samples are used for the relevance and specification tests. The test of Xu and He (2021) is used for the MDep specification and relevance tests (the latter is based on Theorem 4). The Su and Zheng (2017) specification test is applied to the OLS/IV. Excluded Instr. - whether the excluded instrument is used. KB F-Stat. denotes the rk Wald F-statistic of Kleibergen and Paap (2006) for the test of instrument relevance.

and Paap (2006) is provided as a measure of IV-relevance.

MDep parameter estimates of the price elasticity of demand generally agree with OLS/IV estimates. However, one observes that the MDep is more precise than the OLS/IV across all four specifications. Taking specification (4) which also controls for day dummies as the preferred specification, one cannot reject the null hypothesis that the demand for fish is unit-elastic, i.e., a one percent increase in price induces a one percent decrease in the quantity demanded. More interestingly, one fails to reject the null of unit-elastic demand in favour of the alternative of elastic demand. One notices from the respective specification tests that no specification is rejected by the data. The MDep relevance test, like Kleibergen and Paap's (2006) rank test, confirms the strength of the instrument *Stormy*.[8]

[8]The rk Wald F-statistics of Kleibergen and Paap (2006) are above Staiger and Stock's (1997) rule-of-thumb cut-off of 10.

5 Conclusion

This paper introduces the MDep estimator, which weakens the relevance condition of IV, ICM, and non-parametric IV methods to non-independence and hence exploits the maximum number of relevant instruments possible in any given empirical setting. The estimator is based on the distance covariance measure of Székely, Rizzo, and Bakirov (2007), which measures the distributional dependence between random variables of arbitrary dimensions.

The MDep provides a fundamentally different and practically useful approach to dealing with (1) the unavailability of excluded instruments, (2) the weak-IV problem, and (3) the non-existence or contamination of the disturbance term by outliers or random noise with possibly non-existent moments. Consistent estimation and reliable inference is feasible without excludability provided endogenous covariates are non-linearly dependent (in the distributional sense) on exogenous covariates. The MDep is able to handle the weak IV problem by admitting instruments of which the endogenous covariates may be uncorrelated or even mean-independent but not independent. Additionally, no moments for the disturbance term are required, and bounded one-to-one mappings of Z eliminate the need for moment bounds on the instruments.

Consistency and asymptotic normality of the MDep estimator hold under mild conditions. Illustrative examples backed by simulations showcase the remarkable properties of the MDep estimator vis-à-vis existing methods. This paper is a first step in developing a distance covariance-based estimator for econometric models under possible endogeneity. Future work could extend the estimator in several directions, such as incorporating clustering and accommodating temporal dependence in the observations.

References

[1] Andrews, Donald WK, Marcelo J Moreira, and James H Stock. "Optimal two-sided invariant similar tests for instrumental variables regression". *Econometrica* 74.3 (2006), pp. 715–752.

[2] Andrews, Isaiah and Timothy B Armstrong. "Unbiased instrumental variables estimation under known first-stage sign". *Quantitative Economics* 8.2 (2017), pp. 479–503.

[3] Andrews, Isaiah and Anna Mikusheva. "Conditional inference with a functional nuisance parameter". *Econometrica* 84.4 (2016), pp. 1571–1612.

[4] Andrews, Isaiah, James H Stock, and Liyang Sun. "Weak instruments in instrumental variables regression: Theory and practice". *Annual Review of Economics* 11 (2019), pp. 727–753.

[5] Antoine, Bertille and Pascal Lavergne. "Conditional moment models under semi-strong identification". *Journal of Econometrics* 182.1 (2014), pp. 59–69.

[6] Antoine, Bertille and Xiaolin Sun. "Partially linear models with endogeneity: a conditional moment-based approach". *The Econometrics Journal* 25.1 (2022), pp. 256–275.

[7] Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen. "Sparse models and methods for optimal instruments with an application to eminent domain". *Econometrica* 80.6 (2012), pp. 2369–2429.

[8] Bernstein, Dennis S. *Matrix Mathematics: Theory, Facts, and Formulas*. 2009.

[9] Bierens, Herman J. "Consistent model specification tests". *Journal of Econometrics* 20.1 (1982), pp. 105–134.

[10] Bound, John, David A Jaeger, and Regina M Baker. "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak". *Journal of the American Statistical Association* 90.430 (1995), pp. 443–450.

[11] Carrasco, Marine and Jean-Pierre Florens. "Generalization of GMM to a continuum of moment conditions". *Econometric Theory* 16.6 (2000), pp. 797–834.

[12] Carrasco, Marine and Guy Tchuente. "Regularized LIML for many instruments". *Journal of Econometrics* 186.2 (2015), pp. 427–442.

[13] Chen, Jiafeng, Daniel L Chen, and Greg Lewis. "Mostly Harmless Machine Learning: Learning Optimal Instruments in Linear IV Models". *arXiv preprint arXiv:2011.06158* (2020).

[14] Cheng, Xu and Zhipeng Liao. "Select the valid and relevant moments: An information-based lasso for gmm with many moments". *Journal of Econometrics* 186.2 (2015), pp. 443–464.

[15] Chernozhukov, Victor and Christian Hansen. "Instrumental quantile regression inference for structural and treatment effect models". *Journal of Econometrics* 132.2 (2006), pp. 491–525.

[16] Chernozhukov, Victor and Christian Hansen. "Instrumental variable quantile regression: A robust inference approach". *Journal of Econometrics* 142.1 (2008), pp. 379–398.

[17] Cragg, John G. "Using higher moments to estimate the simple errors-in-variables model". *Rand Journal of Economics* (1997), S71–S91.

[18] Dagenais, Marcel G and Denyse L Dagenais. "Higher moment estimators for linear regression models with errors in the variables". *Journal of Econometrics* 76.1-2 (1997), pp. 193–221.

[19] Darolles, Serge, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. "Nonparametric instrumental regression". *Econometrica* 79.5 (2011), pp. 1541–1565.

[20] Das, Mitali. "Instrumental variables estimators of nonparametric models with discrete endogenous regressors". *Journal of Econometrics* 124.2 (2005), pp. 335–361.

[21] Davis, Richard A, Muneya Matsui, Thomas Mikosch, Phyllis Wan, et al. "Applications of distance correlation to time series". *Bernoulli* 24.4A (2018), pp. 3087–3116.

[22] Dieterle, Steven G and Andy Snell. "A simple diagnostic to investigate instrument validity and heterogeneous effects when using a single instrument". *Labour Economics* 42 (2016), pp. 76–86.

[23] Domínguez, Manuel A and Ignacio N Lobato. "Consistent estimation of models defined by conditional moment restrictions". *Econometrica* 72.5 (2004), pp. 1601–1615.

[24] Domínguez, Manuel A and Ignacio N Lobato. "A simple omnibus overidentification specification test for time series econometric models". *Econometric Theory* 31.4 (2015), pp. 891–910.

[25] Donald, Stephen G, Guido W Imbens, and Whitney K Newey. "Empirical likelihood estimation and consistent tests with conditional moment restrictions". *Journal of Econometrics* 117.1 (2003), pp. 55–93.

[26] Donald, Stephen G and Whitney K Newey. "Choosing the number of instruments". *Econometrica* 69.5 (2001), pp. 1161–1191.

[27] Edelmann, Dominic, Konstantinos Fokianos, and Maria Pitsillou. "An updated literature review of distance correlation and its applications to time series". *International Statistical Review* 87.2 (2019), pp. 237–262.

[28] Erickson, Timothy and Toni M Whited. "Two-step GMM estimation of the errors-in-variables model using high-order moments". *Econometric Theory* 18.3 (2002), pp. 776–799.

[29] Escanciano, J Carlos. "A consistent diagnostic test for regression models using projections". *Econometric Theory* 22.6 (2006), pp. 1030–1051.

[30] Escanciano, Juan Carlos. "A simple and robust estimator for linear regression models with strictly exogenous instruments". *The Econometrics Journal* 21.1 (2018), pp. 36–54.

[31] Gao, Wayne Yuan and Rui Wang. "IV Regressions without Exclusion Restrictions". *arXiv preprint arXiv:2304.00626* (2023).

[32] Graddy, Kathryn. "Testing for imperfect competition at the Fulton fish market". *The RAND Journal of Economics* (1995), pp. 75–92.

[33] Hall, Peter and Simon J Sheather. "On the distribution of a studentized quantile". *Journal of the Royal Statistical Society: Series B (Methodological)* 50.3 (1988), pp. 381–391.

[34] Hansen, Christian and Damian Kozbur. "Instrumental variables estimation with many weak instruments using regularized JIVE". *Journal of Econometrics* 182.2 (2014), pp. 290–308.

[35] Hirano, Keisuke and Jack R Porter. "Location properties of point estimators in linear instrumental variables and related models". *Econometric Reviews* 34.6-10 (2015), pp. 720–733.

[36] Hoeffding, Wassily. *The strong law of large numbers for U-statistics.* Tech. rep. North Carolina State University. Dept. of Statistics, 1961.

[37] Honoré, Bo E and James Powell. *Pairwise difference estimators for nonlinear models.* Cambridge University Press, 2005.

[38] Honoré, Bo E and James L Powell. "Pairwise difference estimators of censored and truncated regression models". *Journal of Econometrics* 64.1-2 (1994), pp. 241–278.

[39] Hsu, Shih-Hsun and Chung-Ming Kuan. "Estimation of conditional moment restrictions without assuming parameter identifiability in the implied unconditional moments". *Journal of Econometrics* 165.1 (2011), pp. 87–99.

[40] Huber, Peter J. "The behavior of maximum likelihood estimates under nonstandard conditions". *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 1. University of California Press. 1967, pp. 221–233.

[41] Jiang, Feiyu and Emmanuel Selorm Tsyawo. "A Consistent ICM-based χ^2 Specification Test". *arXiv preprint arXiv:2208.13370* (2022).

[42] Jochmans, Koen. "Pairwise-comparison estimation with non-parametric controls". *The Econometrics Journal* 16.3 (2013), pp. 340–372.

[43] Jurecková, J and Bohumír Procházka. "Regression quantiles and trimmed least squares estimator in nonlinear regression model". *Journal of Nonparametric Statistics* 3.3 (1994), pp. 201–222.

[44] Kitamura, Yuichi, Gautam Tripathi, and Hyungtaik Ahn. "Empirical likelihood-based inference in conditional moment restriction models". *Econometrica* 72.6 (2004), pp. 1667–1714.

[45] Kleibergen, Frank and Richard Paap. "Generalized reduced rank tests using the singular value decomposition". *Journal of Econometrics* 133.1 (2006), pp. 97–126.

[46] Klein, Roger and Francis Vella. "Estimating a class of triangular simultaneous equations models without exclusion restrictions". *Journal of Econometrics* 154.2 (2010), pp. 154–164.

[47] Knight, Keith. "Limiting distributions for L1 regression estimators under general conditions". *Annals of statistics* (1998), pp. 755–770.

[48] Koenker, Roger. *Quantile regression*. Vol. 38. Cambridge university press, 2005.

[49] Koenker, Roger and Gilbert Bassett Jr. "Regression quantiles". *Econometrica: Journal of the Econometric Society* (1978), pp. 33–50.

[50] Lee, A J. *U-statistics: Theory and Practice*. Routledge, 1990.

[51] Lee, Sokbae. "Endogeneity in quantile regression models: A control function approach". *Journal of Econometrics* 141.2 (2007), pp. 1131–1158.

[52] Lewbel, Arthur. "Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and R&D". *Econometrica: journal of the econometric society* (1997), pp. 1201–1213.

[53] Nelson, Charles R and Richard Startz. "The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one". *Journal of business* (1990), S125–S140.

[54] Nelson, Charles R. and Richard Startz. "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator". *Econometrica* 58.4 (1990), pp. 967–976.

[55] Newey, Whitney K and James L Powell. "Instrumental variable estimation of nonparametric models". *Econometrica* 71.5 (2003), pp. 1565–1578.

[56] Ng, Serena and Jushan Bai. "Selecting instrumental variables in a data rich environment". *Journal of Time Series Econometrics* 1.1 (2009).

[57] Oberhofer, Walter and Harry Haupt. "Asymptotic theory for nonlinear quantile regression under weak dependence". *Econometric Theory* 32.3 (2016), p. 686.

[58] Olea, José Luis Montiel and Carolin Pflueger. "A robust test for weak instruments". *Journal of Business & Economic Statistics* 31.3 (2013), pp. 358–369.

[59] Pakes, Ariel and David Pollard. "Simulation and the asymptotics of optimization estimators". *Econometrica: Journal of the Econometric Society* (1989), pp. 1027–1057.

[60] Park, Trevor, Xiaofeng Shao, Shun Yao, et al. "Partial martingale difference correlation". *Electronic Journal of Statistics* 9.1 (2015), pp. 1492–1517.

[61] Pollard, David. "New ways to prove central limit theorems". *Econometric Theory* (1985), pp. 295–313.

[62] Powell, James L. "Estimation of monotonic regression models under quantile restrictions". *Nonparametric and semiparametric methods in Econometrics* (1991), pp. 357–384.

[63] Rigobon, Roberto. "Identification through heteroskedasticity". *Review of Economics and Statistics* 85.4 (2003), pp. 777–792.

[64] Sanderson, Eleanor and Frank Windmeijer. "A weak instrument F-test in linear IV models with multiple endogenous variables". *Journal of Econometrics* 190.2 (2016), pp. 212–221.

[65] Sen, Arnab and Bodhisattva Sen. "Testing independence and goodness-of-fit in linear models". *Biometrika* 101.4 (2014), pp. 927–942.

[66] Shao, Xiaofeng and Jingsi Zhang. "Martingale difference correlation and its use in high-dimensional variable screening". *Journal of the American Statistical Association* 109.507 (2014), pp. 1302–1318.

[67] Sheng, Wenhui and Xiangrong Yin. "Direction estimation in single-index models via distance covariance". *Journal of Multivariate Analysis* 122 (2013), pp. 148–161.

[68] Staiger, Douglas and James H. Stock. "Instrumental Variables Regression with Weak Instruments". *Econometrica* 65.3 (1997), pp. 557–586.

[69] Stock, James and Motohiro Yogo. "Testing for Weak Instruments in Linear IV Regression". *Identification and Inference for Econometric Models*. Ed. by Andrews, Donald W.K. New York: Cambridge University Press, 2005, pp. 80–108.

[70] Su, Liangjun and Xin Zheng. "A martingale-difference-divergence-based test for specification". *Economics Letters* 156 (2017), pp. 162–167.

[71] Székely, Gábor J and Maria L Rizzo. "Brownian distance covariance". *The Annals of Applied Statistics* (2009), pp. 1236–1265.

[72] Székely, Gábor J, Maria L Rizzo, et al. "Partial distance correlation with methods for dissimilarities". *The Annals of Statistics* 42.6 (2014), pp. 2382–2412.

[73] Székely, Gábor J, Maria L Rizzo, and Nail K Bakirov. "Measuring and testing dependence by correlation of distances". *The Annals of Statistics* 35.6 (2007), pp. 2769–2794.

[74] Tsyawo, Emmanuel Selorm. "Feasible IV regression without excluded instruments". *The Econometrics Journal* 26.2 (2023), pp. 235–256.

[75] Wang, Xuexin. *Consistent Estimation Of Models Defined By Conditional Moment Restrictions Under Minimal Identifying Conditions*. Working Papers 2018-10-29. Wang Yanan Institute for Studies in Economics (WISE), Xiamen University, Oct. 2018. URL: https://ideas.repec.org/p/wyi/wpaper/002382.html.

[76] Wooldridge, Jeffrey M. *Econometric analysis of cross section and panel data*. MIT Press, 2010.

[77] Xu, Kai and Fangxue Chen. "Martingale-difference-divergence-based tests for goodness-of-fit in quantile models". *Journal of Statistical Planning and Inference* 207 (2020), pp. 138–154.

[78] Xu, Kai and Daojiang He. "Omnibus Model Checks of Linear Assumptions through Distance Covariance". *Statistica Sinica* 31 (2021), pp. 1055–1079.

[79] Yao, Shun, Xianyang Zhang, and Xiaofeng Shao. "Testing mutual independence in high dimension via distance covariance". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.3 (2018), pp. 455–480.

Appendix

The proofs of the results in the main text are organised in building blocks of lemmata.

A Proof of Theorem 1

The proof of Theorem 1 begins with a uniform law of large numbers result. The following lemma essentially verifies the conditions of Honoré and Powell (1994, Theorem 1).

Lemma A.1. *Suppose Assumption 1 holds, then*

(a) there exists a function $\mathcal{B} : \mathcal{W} \times \mathcal{W} \to \mathbb{R}_+$ *with* $\mathbb{E}[\mathcal{B}(X, X^\dagger, Z, Z^\dagger)] \leq C^{1/4}$ *such that for any* $\theta_1, \theta_2 \in \Theta$, $|q(W, W^\dagger; \theta_1) - q(W, W^\dagger; \theta_2)| \leq \mathcal{B}(X, X^\dagger, Z, Z^\dagger)||\theta_1 - \theta_2||$;

(b) $Q(\theta)$ *is continuous in* θ *uniformly, and* $\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{a.s.} 0$.

Proof of Lemma A.1.

Part (a): First, let $\mathcal{B}(X, X^\dagger, Z, Z^\dagger) \equiv \sup_{\theta \in \Theta} ||\mathcal{Z}\widetilde{X^g}(\theta)||$. By Lyapunov's inequality and Assumption 1(c),

$$\mathbb{E}[\mathcal{B}(X, X^\dagger, Z, Z^\dagger)] \equiv \mathbb{E}[\sup_{\theta \in \Theta} ||\mathcal{Z}\widetilde{X^g}(\theta)||] \leq (\mathbb{E}[\sup_{\theta \in \Theta} ||\mathcal{Z}\widetilde{X^g}(\theta)||^4])^{1/4} \leq C^{1/4}.$$

Second, for any $W, W^\dagger$ defined on the support of W_i and $\theta_1, \theta_2, \bar{\theta}_{1,2} \in \Theta$ where $\bar{\theta}_{1,2}$, by Assumption 1(b) and the Mean-Value Theorem (MVT), satisfies $\widetilde{U}(\theta_1) - \widetilde{U}(\theta_2) = \widetilde{X^g}(\bar{\theta}_{1,2})(\theta_1 - \theta_2)$,

$$\begin{aligned}
\mathcal{B}(X, X^\dagger, Z, Z^\dagger) \cdot ||\theta_1 - \theta_2|| &\equiv \sup_{\theta \in \Theta} ||\mathcal{Z}\widetilde{X^g}(\theta)|| \cdot ||\theta_1 - \theta_2|| \\
&\geq |\mathcal{Z}| \cdot |\widetilde{X^g}(\bar{\theta}_{1,2})(\theta_2 - \theta_1)| \\
&= |\mathcal{Z}| \cdot |\widetilde{U}(\theta_1) - \widetilde{U}(\theta_2)| \\
&\geq \left| \mathcal{Z} \Big((|\widetilde{U}(\theta_1)| - |\widetilde{U}(\theta_o)|) - (|\widetilde{U}(\theta_2)| - |\widetilde{U}(\theta_o)|) \Big) \right| \\
&= |q(W, W^\dagger; \theta_1) - q(W, W^\dagger; \theta_2)|.
\end{aligned}$$

The first and second inequalities follow from the Schwartz and triangle inequalities, respectively.

Part (b): From Assumption 1(d), there exists a constant $C_\theta < \infty$ such that $||\theta_1 - \theta_2|| < C_\theta$ for all $\theta_1, \theta_2 \in \Theta$. It thus follows from part (a) above that $|q(W, W^\dagger; \theta)| = |q(W, W^\dagger; \theta) - q(W, W^\dagger; \theta_0)| < C_\theta \mathcal{B}(X, X^\dagger, Z, Z^\dagger)$, and this verifies Honoré and Powell (1994, Assumption C3).

Assumptions 1(a) and 1(b) imply the measurability of $q(W, W^\dagger; \theta) \equiv \mathcal{Z}(|\widetilde{U}(\theta)| - |\widetilde{U}|)$ in $[W, W^\dagger]$ for all $\theta \in \Theta$. Assumption 1(b) and the continuity of the absolute value function imply $q(W, W^\dagger; \theta)$ is continuous in $\theta \in \Theta$ on the support of $[W, W^\dagger]$. $\mathbb{E}[Q_n(\theta)] = Q(\theta) \equiv \mathbb{E}[q(W, W^\dagger; \theta)]$ by Property (d). As the expectation operator is continuous, $Q(\theta)$ is continuous. This verifies Honoré and Powell (1994, Assumption C2). In addition to Assumption 1(d), the conclusion follows from Theorem 1 of Honoré and Powell (1994). □

As a next step for the proof of Theorem 1, Lemma A.2 is useful in proving the identification of θ_o. Define the random variable $\tilde{\tau}_\theta \equiv F_{\widetilde{U}|\tilde{\sigma}(X,Z)}\big(\lambda \widetilde{X^g}(\bar{\theta})(\theta - \theta_o)\big)$ and the random set-valued map indexed by $\theta \in \Theta$, $\mathscr{A}_\theta : [0,1] \twoheadrightarrow \mathcal{W}$ such that

$$\mathscr{A}_\theta(\tau) = \big\{X, X^\dagger, Z, Z^\dagger \in \mathcal{W} : F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\lambda \widetilde{X^g}(\bar{\theta})(\theta - \theta_o)) = \tau\big\},$$

$\lambda \in (0,1)$, $F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\cdot)$ is the conditional distribution function, $\bar{\theta}$ satisfies $\widetilde{U}(\theta) = \widetilde{U} + \widetilde{X^g}(\bar{\theta})(\theta - \theta_o)$, and $\lambda \in (0,1)$ under Assumption 1 satisfies $\int_0^x F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\eta)d\eta = F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\lambda x)x$ by the (integral) MVT. The following provides an important result for identification.

Lemma A.2. *Suppose Assumptions 1(b) and 2(a) hold, then for any $W, W^\dagger$ defined on the support of W_i,*

$$\mathbb{E}[q(W, W^\dagger; \theta)] = \int_0^1 |2\tau - 1| \Big(\int_{\mathscr{A}_\theta(\tau)} \big[\mathcal{Z}\big|\widetilde{X^g}(\bar{\theta})(\theta - \theta_o)\big|\big] d\mathbb{P}_{\widetilde{X^g}, \mathcal{Z}} \Big) d\tau.$$

The inner integrand $\mathcal{Z}\big|\widetilde{X^g}(\bar{\theta})(\theta - \theta_o)\big|$ has the same form as the integrand of the distance

covariance measure $\mathcal{V}^2(U,Z) = \mathbb{E}[\mathcal{Z}|\widetilde{U}|]$. This is important because the expectation of the normalised minimand can be expressed in terms of the distance covariance between $X^g(\bar{\theta})(\theta-\theta_o)$ and Z for all $\theta \neq \theta_o$ and $\bar{\theta}$, which satisfies $||\bar{\theta}-\theta_o|| \leq ||\theta-\theta_o||$.

Proof of Lemma A.2. By Assumption 1(b), the equality $\widetilde{U}(\theta) = \widetilde{U} - \widetilde{X^g}(\bar{\theta})(\theta-\theta_o)$ holds by the MVT for any pair of random vectors $W, W^\dagger$ where $\bar{\theta}$ satisfies $||\theta-\bar{\theta}|| \leq ||\theta-\theta_o||$. Knight's identity (Knight, 1998) is given by

$$|\xi-b|-|\xi| = -b\big(\mathrm{I}(\xi>0)-\mathrm{I}(\xi<0)\big) + 2\int_0^b \big(\mathrm{I}(\xi\leq\eta)-\mathrm{I}(\xi\leq 0)\big)d\eta.$$

Applying expectations to a continuously distributed ξ, one has by the MVT

$$\begin{aligned}\mathbb{E}[|\xi-b|-|\xi|] &= (2F_\xi(0)-1)b + 2\int_0^b (F_\xi(\eta)-F_\xi(0))d\eta \\ &= (2F_\xi(0)-1)b + 2(F_\xi(\lambda b)-F_\xi(0))b \\ &= (2F_\xi(\lambda b)-1)b\end{aligned}$$

for some $\lambda \in (0,1)$. It follows from the foregoing and the LIE that

$$\begin{aligned}\mathbb{E}[q(W,W^\dagger;\theta)] &= \mathbb{E}[\mathcal{Z}(|\widetilde{U}(\theta)|-|\widetilde{U}|)] \\ &= \mathbb{E}\big[\mathcal{Z}\big(|\widetilde{U}-\widetilde{X^g}(\bar{\theta})(\theta-\theta_o)|-|\widetilde{U}|\big)\big] \\ &= \mathbb{E}\Big[\mathcal{Z}\Big(2F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\lambda\widetilde{X^g}(\bar{\theta})(\theta-\theta_o))-1\Big)\widetilde{X^g}(\bar{\theta})(\theta-\theta_o)\Big].\end{aligned} \tag{A.1}$$

for some $\lambda \in (0,1)$ thanks to the MVT and the law of iterated expectations (LIE).

$\Big(2F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\lambda b)-1\Big)b \geq 0$ for all $b \in \mathbb{R}$ and $\lambda \in (0,1)$. To see this, observe that if $b>0$, $\lambda b>0$, $F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\lambda b) \geq 1/2$ by Assumption 2(a) and the monotonicity property of (conditional) cumulative distribution functions, thus $\Big(2F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\lambda b)-1\Big)b \geq 0$ if $b \geq 0$. The same sequence of arguments as the foregoing shows that $\Big(2F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\lambda b)-1\Big)b \geq 0$ if $b \leq 0$. Hence, $\Big(2F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\lambda\widetilde{X^g}(\bar{\theta})(\theta-\theta_o))-1\Big)\widetilde{X^g}(\bar{\theta})(\theta-\theta_o) = \Big|\Big(2F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\lambda\widetilde{X^g}(\bar{\theta})(\theta-$

$\theta_o)) - 1\Big)\widetilde{X^g}(\bar{\theta})(\theta - \theta_o)\Big|$ under Assumption 2(a). It therefore follows from (A.1) that

$$\begin{aligned}
\mathbb{E}[q(W, W^\dagger; \theta)] &= \mathbb{E}\Big[\mathcal{Z}\Big|\Big(2F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\lambda\widetilde{X^g}(\bar{\theta})(\theta - \theta_o)) - 1\Big)\widetilde{X^g}(\bar{\theta})(\theta - \theta_o)\Big|\Big] \\
&= \mathbb{E}\big[\mathcal{Z}\big|2F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\lambda\widetilde{X^g}(\bar{\theta})(\theta - \theta_o)) - 1\big| \times \big|\widetilde{X^g}(\bar{\theta})(\theta - \theta_o)\big|\big] \\
&= \mathbb{E}\big[|2\tilde{\tau}_\theta - 1| \times \mathcal{Z}|\widetilde{X^g}(\bar{\theta})(\theta - \theta_o)|\big] \\
&= \int_0^1 |2\tau - 1|\mathbb{E}\big[(\mathcal{Z}|\widetilde{X^g}(\bar{\theta})(\theta - \theta_o)|)|\tilde{\tau}_\theta = \tau\big]d\tau \\
&= \int_0^1 |2\tau - 1|\Big(\int_{\mathscr{A}_\theta(\tau)} \big[\mathcal{Z}\big|\widetilde{X^g}(\bar{\theta})(\theta - \theta_o)\big|\big]d\mathbb{P}_{\widetilde{X^g},\mathcal{Z}}\Big)d\tau.
\end{aligned}$$

The fourth equality follows from the LIE. □

With Lemmata A.1 and A.2 in hand, the identification result is proved next.

Proof of Theorem 1(a). Under the assumptions of Lemma A.2,

$$\begin{aligned}
Q(\theta) &= \mathbb{E}[q(W, W^\dagger; \theta)] \\
&= \int_0^1 |2\tau - 1|\Big(\int_{\mathscr{A}_\theta(\tau)} \big[\mathcal{Z}\big|\widetilde{X^g}(\bar{\theta})(\theta - \theta_o)\big|\big]d\mathbb{P}_{\widetilde{X^g},\mathcal{Z}}\Big)d\tau \\
&= \int_0^1 |2\tau - 1|\mathcal{V}^2_{\cdot,\tau}(X^g(\bar{\theta})(\theta - \theta_o), Z)d\tau
\end{aligned}$$

where

$$\mathcal{V}^2_{\cdot,\tau}(X^g(\bar{\theta})(\theta - \theta_o), Z) \equiv \int_{\mathscr{A}_\theta(\tau)} \mathcal{Z}\big|\widetilde{X^g}(\bar{\theta})(\theta - \theta_o)\big|d\mathbb{P}_{\widetilde{X^g},\mathcal{Z}}$$

is the distance covariance between $X^g(\bar{\theta})(\theta - \theta_o)$ and Z over the measurable set $\mathscr{A}_\theta(\tau)$.

By the LIE,

$$\begin{aligned}
\int_0^1 \mathcal{V}^2_{\cdot,\tau}(X^g(\bar{\theta})(\theta - \theta_o), Z)d\tau &= \int_0^1 \mathbb{E}\big[(\mathcal{Z}|\widetilde{X^g}(\bar{\theta})(\theta - \theta_o)|)|\tilde{\tau}_\theta = \tau\big]d\tau \\
&= \mathbb{E}\big[\mathcal{Z}\big|\widetilde{X^g}(\bar{\theta})(\theta - \theta_o)\big|\big] \\
&\equiv \mathcal{V}^2\big(X^g(\bar{\theta})(\theta - \theta_o), Z\big).
\end{aligned} \tag{A.2}$$

For the class of models considered, $X^g(\theta) \equiv -g'(X\theta)X$ is measurable in X for all θ (Assumption 1(b)). Moreover, for all non-zero linear combinations of $X^g(\theta)$, $X^g(\theta)\tau = -g'(X\theta)X\tau \not\perp Z$ is automatically satisfied as long as at least one element of X is not independent of Z and $g'(X\theta)$ is not constant *a.s.* This form of "automatic" identification is not uniform in θ, however, as it fails when $g'(0) \neq 0$ is constant. Assumption 2(b) is thus sufficient for $\mathcal{V}^2(X^g(\bar{\theta})(\theta-\theta_o), Z) > 0$ to hold whenever $\theta \neq \theta_o$. Formally, Assumption 2(b) and Properties (a) and (b) imply that there exists a $\tilde{\delta}_\varepsilon > 0$ such that $\inf_{\{\theta\in\Theta:||\theta-\theta_o||\geq\varepsilon\}} \mathcal{V}^2(X^g(\bar{\theta})(\theta-\theta_o), Z) > \tilde{\delta}_\varepsilon$. By Property (a),

$$\inf_{\{\theta\in\Theta:||\theta-\theta_o||\geq\varepsilon\}} \mathcal{V}^2_{\cdot,\tau}(X^g(\bar{\theta})(\theta-\theta_o), Z) \geq 0$$

for any $\tau \in [0,1]$.

Define functions $f_\theta(\tau) = \mathcal{V}^2_{\cdot,\tau}(X^g(\bar{\theta})(\theta-\theta_o), Z)/\mathcal{V}^2(X^g(\bar{\theta})(\theta-\theta_o), Z)$ and $F_\theta(\tau) = \int_0^\tau f_\theta(t)dt$. For all $\theta \neq \theta_o$, $f_\theta(\tau) \geq 0$, $\int_0^1 f_\theta(t)dt = 1$ by (A.2), and

$$\int_0^1 |2\tau-1| f_\theta(\tau)d\tau = 1 - 2\Big(\int_{0.5}^1 F_\theta(\tau)d\tau - \int_0^{0.5} F_\theta(\tau)d\tau\Big)$$

using integration by parts. By the MVT, there exist $\bar{\tau}^*_\theta$ and $\underline{\tau}^*_\theta$, $0 < \underline{\tau}^*_\theta < 0.5 < \bar{\tau}^*_\theta < 1$ such that $\int_0^{0.5} F_\theta(\tau)d\tau = 0.5F_\theta(\underline{\tau}^*_\theta)$, and $\int_{0.5}^1 F_\theta(\tau)d\tau = 0.5F_\theta(\bar{\tau}^*_\theta)$. Further, $0 < F_\theta(\underline{\tau}^*_\theta) \leq F_\theta(\bar{\tau}^*_\theta) < 1$, and this implies

$$\begin{aligned} Q(\theta) &= \mathcal{V}^2(X^g(\bar{\theta})(\theta-\theta_o), Z)\int_0^1 |2\tau-1| f_\theta(\tau)d\tau \\ &= \big(1 - (F_\theta(\bar{\tau}^*_\theta) - F_\theta(\underline{\tau}^*_\theta))\big)\mathcal{V}^2(X^g(\bar{\theta})(\theta-\theta_o), Z) \\ &\equiv d(\theta)\mathcal{V}^2(X^g(\bar{\theta})(\theta-\theta_o), Z). \end{aligned}$$

$0 < \inf_{\{\theta\in\Theta:||\theta-\theta_o||\geq\varepsilon\}} d(\theta) \leq \sup_{\{\theta\in\Theta:||\theta-\theta_o||\geq\varepsilon\}} d(\theta) < 1$ for all $\varepsilon > 0$. Set $\delta_\varepsilon = \tilde{\delta}_\varepsilon \inf_{\{\theta\in\Theta:||\theta-\theta_o||\geq\varepsilon\}} d(\theta)$. As Assumption 2(b) is sufficient for $X^g(\bar{\theta})(\theta-\theta_o) \not\perp Z$ for all $\theta \neq \theta_o$, i.e., $\delta_\varepsilon > 0$, the proof is complete. □

Proof of Theorem 1(b). Under the assumptions of Lemmata A.1, A.2, and Theorem 1, the conclusion follows from Corollary 1 of Honoré and Powell (1994). □

B Proof of Theorem 2

Define the following score functions:

(B.1)

$$\widehat{\mathcal{S}}_n(\theta) \equiv \mathbb{E}_n\big[(1 - 2\mathrm{I}(\widetilde{U}_{ij}(\theta) \leq 0))\mathcal{Z}_{ij,n}\widetilde{X}^g_{ij}(\theta)\big] \text{ and } \mathcal{S}_n(\theta) \equiv \mathbb{E}_n\big[(1 - 2\mathrm{I}(\widetilde{U}_{ij}(\theta) \leq 0))\mathcal{Z}_{ij}\widetilde{X}^g_{ij}(\theta)\big].$$

Observe that $\widehat{\mathcal{S}}_n(\theta) \equiv \frac{\partial Q_n(\theta)}{\theta}$ and $\mathcal{S}_n(\theta)$ uses $\mathcal{Z}_{ij} = h(Z_i, Z_i)$ instead of $\mathcal{Z}_{ij,n} = h_n(Z_i, Z_j)$. The latter is a second-order U-statistic, while the former can be shown to be a fourth-order U-statistic – see, e.g., Yao, Zhang, and Shao (2018, Lemma 1).

The following result provides convergence rates on the score functions evaluated at the estimator $\widehat{\theta}_n$.

Lemma B.1. *Under Assumption 1((b) and (c)) and Assumption 3((a) and (b)),* $\sqrt{n}||\mathcal{S}_n(\widehat{\theta}_n)|| = o_p(n^{-1})$ *and* $\sqrt{n}||\widehat{\mathcal{S}}_n(\widehat{\theta}_n)|| = o_p(n^{-1/2})$.

Proof. Applying the chain rule,

$$\frac{\partial^-(q(W_i, W_j; \theta))}{\partial\theta} = \mathcal{Z}_{ij}\partial^-|\widetilde{U}_{ij}(\theta)| \times \frac{\partial^-\widetilde{U}_{ij}(\theta)}{\partial\theta} = \partial^-|\widetilde{U}_{ij}(\theta)|\mathcal{Z}_{ij}\widetilde{X}^g_{ij}(\theta)$$

where the last equality follows by the continuous differentiability of $\widetilde{U}(\theta)$ (Assumption 1(b)). By the consistency of the MDep (Theorem 1(b)), the left- and right-differentiability of the absolute value function, and Assumption 3(b), the left and right derivatives of $Q_n(\theta)$ at $\widehat{\theta}_n$ are of opposite signs. In addition to the consistency of $\widehat{\theta}_n$ (Theorem 1(a)) given an open

neighbourhood Θ_o of θ_o, it follows that

$$\begin{aligned}
||\sqrt{n}\mathcal{S}_n(\widehat{\theta}_n)|| &\leq \frac{1}{n^{3/2}}\sum_{i\neq j}^{n}\left|\partial^-|\widetilde{U}_{ij}(\widehat{\theta}_n)| - \partial^+|\widetilde{U}_{ij}(\widehat{\theta}_n)|\right| \cdot ||\mathcal{Z}_{ij}\widetilde{X}_{ij}^g(\widehat{\theta}_n)|| \\
&\leq \frac{1}{n^{1/2}}\sqrt{\sup_{\theta\in\Theta_o}\sum_{i\neq j}^{n}\left|\partial^-|\widetilde{U}_{ij}(\theta)| - \partial^+|\widetilde{U}_{ij}(\theta)|\right|^2} \cdot \sqrt{\frac{1}{n^2}\sup_{\theta\in\Theta}\sum_{i\neq j}^{n}||\mathcal{Z}_{ij}\widetilde{X}_{ij}^g(\theta)||^2} \\
&\leq \frac{\sqrt{2}}{n^{1/2}}\sqrt{\sup_{\theta\in\Theta_o}\sum_{i\neq j}^{n}\left|\partial^-|\widetilde{U}_{ij}(\theta)| - \partial^+|\widetilde{U}_{ij}(\theta)|\right|} \cdot \sqrt{\frac{1}{n^2}\sup_{\theta\in\Theta}\sum_{i\neq j}^{n}||\mathcal{Z}_{ij}\widetilde{X}_{ij}^g(\theta)||^2} \\
&= \mathcal{O}_p(n^{-1/2})
\end{aligned}$$

by the triangle inequality, Cauchy-Schwartz inequality, that $2 \geq \left|\partial^-|\widetilde{U}_{ij}(\theta)| - \partial^+|\widetilde{U}_{ij}(\theta)|\right|$, Assumption 3(b), and Assumption 1(c). From the foregoing,

$$\begin{aligned}
\sqrt{n}||\widehat{\mathcal{S}}_n(\widehat{\theta}_n) - \mathcal{S}_n(\widehat{\theta}_n)|| &\leq \frac{1}{n^{3/2}}\sum_{i\neq j}^{n}\left|\partial^-|\widetilde{U}_{ij}(\widehat{\theta}_n)| - \partial^+|\widetilde{U}_{ij}(\widehat{\theta}_n)|\right| \cdot |\mathcal{Z}_{ij,n} - \mathcal{Z}_{ij}| \cdot ||\widetilde{X}_{ij}^g(\widehat{\theta}_n)|| \\
&\leq \mathcal{O}_p(n^{-1/2}) \times \frac{1}{n^{3/2}}\sum_{i\neq j}^{n}\left|\partial^-|\widetilde{U}_{ij}(\widehat{\theta}_n)| - \partial^+|\widetilde{U}_{ij}(\widehat{\theta}_n)|\right| \cdot ||\mathcal{Z}_{ij}\widetilde{X}_{ij}^g(\widehat{\theta}_n)|| \\
&= \mathcal{O}_p(n^{-1})
\end{aligned}$$

where $|\mathcal{Z}_{ij,n} - \mathcal{Z}_{ij}| = \mathcal{O}_p(n^{-1/2})$ holds under the conditions of Lemma E.1.

Thus, by the triangle inequality,

$$||\sqrt{n}\widehat{\mathcal{S}}_n(\widehat{\theta}_n)|| \leq \sqrt{n}||\widehat{\mathcal{S}}_n(\widehat{\theta}_n) - \mathcal{S}_n(\widehat{\theta}_n)|| + ||\sqrt{n}\mathcal{S}_n(\widehat{\theta}_n)|| = \mathcal{O}_p(n^{-1/2}).$$

□

The next result obtains an asymptotically linear expression for the MDep $\widehat{\theta}_n$.

Lemma B.2. *Under the conditions of Lemma B.1, Lemma B.2, Assumption 2(a), and*

Assumption 3(e), the MDep $\widehat{\theta}_n$ satisfies the asymptotic linearity relation

$$\sqrt{n}(\widehat{\theta}_n - \theta_o) = -\mathcal{H}^{-1}\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\psi^{(1)}(W_i) + o_p(1).$$

Proof. Under Assumption 2(a) and the LIE,

$$\mathcal{S}(\theta_o) = \mathbb{E}\big[(1 - 2\mathrm{I}(\widetilde{U} \leq 0))\mathcal{Z}\widetilde{X^g}\big] = \mathbb{E}\big[(1 - 2F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(0))\mathcal{Z}\widetilde{X^g}\big] = 0.$$

Under the conditions of Lemma E.2, $\mathcal{S}(\theta)$ is differentiable. Expanding around $\widehat{\theta}_n$, one has $0 = \mathcal{S}(\theta_o) = \mathcal{S}(\widehat{\theta}_n) - \mathcal{H}(\bar{\theta}_n)(\widehat{\theta}_n - \theta_o)$ where $\bar{\theta}_n$ satisfies $||\bar{\theta}_n - \theta_o|| \leq ||\widehat{\theta}_n - \theta_o||$. Thus,

$$\sqrt{n}(\widehat{\theta}_n - \theta_o) = \mathcal{H}(\bar{\theta}_n)^{-1}\sqrt{n}\mathcal{S}(\widehat{\theta}_n). \tag{B.2}$$

Denote $v_n(\theta) := \sqrt{n}\big(\mathcal{S}_n(\theta) - \mathcal{S}(\theta)\big)$. The following studies the term $\sqrt{n}\mathcal{S}(\widehat{\theta}_n)$ in the above expression. Consider the expansion

$$\begin{aligned}
\sqrt{n}\mathcal{S}(\widehat{\theta}_n) &= -\sqrt{n}\big(\mathcal{S}_n(\widehat{\theta}_n) - \mathcal{S}(\widehat{\theta}_n)\big) + \sqrt{n}\mathcal{S}_n(\widehat{\theta}_n) \\
&\equiv -v_n(\widehat{\theta}_n) + \sqrt{n}\mathcal{S}_n(\widehat{\theta}_n) \\
&= -v_n(\theta_o) + r_{1n} \\
&= -\sqrt{n}\binom{n}{2}^{-1}\sum_{i<j}^{n}\psi(W_i, W_j) + r_{1n} \\
&= -\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\psi^{(1)}(W_i) + r_{1n} + r_{2n}
\end{aligned} \tag{B.3}$$

where the last equality follows by the Hoeffding decomposition,

$$\begin{aligned}
r_{1n} &\equiv -(v_n(\widehat{\theta}_n) - v_n(\theta_o)) + \sqrt{n}\mathcal{S}_n(\widehat{\theta}_n) \text{ and} \\
r_{2n} &\equiv \frac{2}{n^{1/2}(n-1)}\sum_{i\neq j}^{n}\big[\psi(W_i, W_j) - \psi^{(1)}(W_i) - \psi^{(1)}(W_j)\big].
\end{aligned}$$

From Lemma B.1, $\sqrt{n}||\mathcal{S}(\widehat{\theta}_n)|| = \sqrt{n}||\mathbb{E}[\mathcal{S}_n(\widehat{\theta}_n)]|| = o(n^{-1})$. The consistency of the MDep $\widehat{\theta}_n$ (Theorem 1(b)) and the stochastic equi-continuity assumption (Assumption 3(e)) imply

$$||v_n(\widehat{\theta}_n) - v_n(\theta_o)|| = o_p(1) \times (1 + \sqrt{n}||\mathcal{S}(\widehat{\theta}_n)||) = o_p(1) + o_p(1) \times o(n^{-1}) = o_p(1)$$

whence $r_{1n} = o_p(1)$. In addition, $r_{2n} = o_p(1)$ by Lee (1990, Theorem 3, Sect. 1.3).

Combining Equations (B.2) and (B.3) and the foregoing,

$$\sqrt{n}(\widehat{\theta}_n - \theta_o) = -\mathcal{H}(\bar{\theta}_n)^{-1}\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\psi^{(1)}(W_i) + o_p(1).$$

As $||\bar{\theta}_n - \theta_o|| \leq ||\widehat{\theta}_n - \theta_o||$, the conclusion follows from the continuity of the inverse at a non-singular matrix Assumption 3((f)), the continuous mapping theorem, and Theorem 1(b). □

The conclusion follows from the asymptotic linearity in Lemma B.2, Assumption 1(a), the Lindberg-Lévy Central Limit Theorem, and the continuous mapping theorem.

C Proof of Theorem 3

The following expression is useful in subsequent analyses. For any positive ϵ_1, ϵ_2 in a neighbourhood of zero,

(C.1)

$$\begin{aligned}\mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}[\mathrm{I}(|\widetilde{U}| \leq \epsilon_1)]/(2\epsilon_2) &= \frac{F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\epsilon_1) - F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(-\epsilon_1)}{2\epsilon_2} \\ &= \frac{F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\epsilon_1) - \left(F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\epsilon_1) - f_{\widetilde{U}|\tilde{\sigma}(X,Z)}((1-2\lambda)\epsilon_1)(2\epsilon_1)\right)}{2\epsilon_2} \\ &= (\epsilon_1/\epsilon_2) f_{\widetilde{U}|\tilde{\sigma}(X,Z)}((1-2\lambda)\epsilon_1)\end{aligned}$$

for some $\lambda \in (0,1)$ by Assumption 3(c) and the MVT (taken about $-\epsilon_1$). Recall $\widehat{\mathcal{H}}_n = \frac{1}{n^2\hat{c}_n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\mathrm{I}(|\widetilde{U}_{ij}(\widehat{\theta}_n)| \leq \hat{c}_n)\mathcal{Z}_{ij,n}\widetilde{X}_{ij}^{g}(\widehat{\theta}_n)'\widetilde{X}_{ij}^{g}(\widehat{\theta}_n)\right\}$ and

$\mathcal{H} = \mathbb{E}[f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(0)\mathcal{Z}\widetilde{X^g}'\widetilde{X^g}]$. Define $\mathcal{H}_n \equiv \frac{1}{n^2 c_n}\sum_{i=1}^n\sum_{j=1}^n \Big\{\mathrm{I}(|\widetilde{U}_{ij}| \le c_n)\mathcal{Z}_{ij}\widetilde{X}_{ij}^{g\prime}\widetilde{X}_{ij}^g\Big\}$.

A first step in the proof is to show that $||\widehat{\mathcal{H}}_n - \mathcal{H}_n|| = o_p(1)$. By the triangle inequality, $||\widehat{\mathcal{H}}_n - \mathcal{H}_n|| \le \frac{c_n}{\hat{c}_n}(A_{n,0} + A_{n,1} + A_{n,2} + A_{n,3})$ where

$$
\begin{aligned}
A_{n,0} &\equiv \frac{1}{n^2 c_n}\sum_{i=1}^n\sum_{j=1}^n \Big\{|\mathcal{Z}_{ij,n} - \mathcal{Z}_{ij}| \times \mathrm{I}(|\widetilde{U}_{ij}(\widehat{\theta}_n)| \le \hat{c}_n) \times ||\widetilde{X}_{ij}^g(\widehat{\theta}_n)'\widetilde{X}_{ij}^g(\widehat{\theta}_n)||\Big\},\\
A_{n,1} &\equiv \frac{1}{n^2 c_n}\sum_{i=1}^n\sum_{j=1}^n \Big\{|\mathrm{I}(|\widetilde{U}_{ij}(\widehat{\theta}_n)| \le \hat{c}_n) - \mathrm{I}(|\widetilde{U}_{ij}| \le \hat{c}_n)| \times ||\mathcal{Z}_{ij}\widetilde{X}_{ij}^g(\widehat{\theta}_n)'\widetilde{X}_{ij}^g(\widehat{\theta}_n)||\Big\},\\
A_{n,2} &\equiv \frac{1}{n^2 c_n}\sum_{i=1}^n\sum_{j=1}^n \Big\{\mathrm{I}(|\widetilde{U}_{ij}| \le \hat{c}_n) \times ||\mathcal{Z}_{ij}[\widetilde{X}_{ij}^g(\widehat{\theta}_n)'\widetilde{X}_{ij}^g(\widehat{\theta}_n) - \widetilde{X}_{ij}^{g\prime}\widetilde{X}_{ij}^g]||\Big\}, \text{ and}\\
A_{n,3} &\equiv \frac{1}{n^2 c_n}\sum_{i=1}^n\sum_{j=1}^n \Big\{|\mathrm{I}(|\widetilde{U}_{ij}| \le \hat{c}_n) - \mathrm{I}(|\widetilde{U}_{ij}| \le c_n)| \times ||\mathcal{Z}_{ij}\widetilde{X}_{ij}^{g\prime}\widetilde{X}_{ij}^g||\Big\}.
\end{aligned}
$$

The following lemma studies the terms $A_{n,0}$, $A_{n,1}$, $A_{n,2}$, and $A_{n,3}$.

Lemma C.1. *Under Assumptions 1 to 4, $A_{n,0} = o_p(1)$, $A_{n,1} = o_p(1)$, $A_{n,2} = o_p(1)$, and $A_{n,3} = o_p(1)$.*

Proof. The verification of terms $A_{n,0}, A_{n,1}, A_{n,2}$, and $A_{n,3}$ proceeds in the following parts.

$A_{n,0}$:

By the Schwartz inequality, the Cauchy-Schwartz (CS) inequality, and the identical distribution of the data Assumption 1(a),

$$
\begin{aligned}
\mathbb{E}[A_{n,0}] &= \frac{1}{n^2 c_n}\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}\big[|\mathcal{Z}_{ij,n} - \mathcal{Z}_{ij}| \times \mathrm{I}(|\widetilde{U}_{ij}(\widehat{\theta}_n)| \le \hat{c}_n) \times ||\widetilde{X}_{ij}^g(\widehat{\theta}_n)'\widetilde{X}_{ij}^g(\widehat{\theta}_n)||\big]\\
&\le \frac{(\mathbb{E}[\sup_{\theta\in\Theta}||\widetilde{X^g}(\theta)||^4])^{1/2}}{n^2 c_n}\sum_{i=1}^n\sum_{j=1}^n (\mathbb{E}[(\mathcal{Z}_{ij,n} - \mathcal{Z}_{ij})^2])^{1/2}.
\end{aligned}
$$

Under the assumptions of Lemma E.1, Assumption 1(c), and Assumption 4, it follows that $A_{n,0} = \mathcal{O}_p((\sqrt{n}c_n)^{-1}) = o_p(1)$.

$A_{n,1}$:

For the term $A_{n,1}$, note that for $\bar{\theta}_n$ that satisfies $||\bar{\theta}_n - \theta_o|| \leq ||\widehat{\theta}_n - \theta_o||$. Thus,

$$\begin{aligned}
&\mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}\big[|\mathrm{I}(|\widetilde{U}_{ij}(\widehat{\theta}_n)| \leq \hat{c}_n) - \mathrm{I}(|\widetilde{U}_{ij}| \leq \hat{c}_n)|\big] \\
&= \mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}\big[|\mathrm{I}(|\widetilde{U}_{ij} + \widetilde{X}^g_{ij}(\bar{\theta}_n)(\widehat{\theta}_n - \theta_o)| \leq \hat{c}_n) - \mathrm{I}(|\widetilde{U}_{ij}| \leq \hat{c}_n)|\big] \\
&= \mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}\big[\mathrm{I}(\hat{c}_n < \widetilde{U}_{ij} \leq \hat{c}_n - \widetilde{X}^g_{ij}(\bar{\theta}_n)(\widehat{\theta}_n - \theta_o))\big] \\
&+ \mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}\big[\mathrm{I}(-\hat{c}_n - \widetilde{X}^g_{ij}(\bar{\theta}_n)(\widehat{\theta}_n - \theta_o) \leq \widetilde{U}_{ij} < -\hat{c}_n)\big] \\
&\equiv \mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}[\tilde{I}^1_{ij}] + \mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}[\tilde{I}^2_{ij}]
\end{aligned}$$

by Assumption 1(b), and the MVT. For ease of notation, let $\widetilde{J}_{x,\theta} \equiv \widetilde{X}^g_{ij}(\bar{\theta}_n)(\widehat{\theta}_n - \theta_o)$. By Assumption 3(c) and the MVT,

$$\begin{aligned}
\mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}[\tilde{I}^1_{ij}] &= \max\{0, F_{\widetilde{U}|\tilde{\sigma}(X,Z)}\big(\hat{c}_n - \widetilde{J}_{x,\theta}\big) - F_{\widetilde{U}|\tilde{\sigma}(X,Z)}\big(\hat{c}_n\big)\} \\
&= \max\{0, -f_{\widetilde{U}|\tilde{\sigma}(X,Z)}\big(\hat{c}_n - \lambda_1\widetilde{J}_{x,\theta}\big)\widetilde{J}_{x,\theta}\} \text{ and} \\
\mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}[\tilde{I}^1_{ij}] &= \max\{0, F_{\widetilde{U}|\tilde{\sigma}(X,Z)}\big(-\hat{c}_n\big) - F_{\widetilde{U}|\tilde{\sigma}(X,Z)}\big(-\hat{c}_n - \widetilde{J}_{x,\theta}\big)\} \\
&= \max\{0, f_{\widetilde{U}|\tilde{\sigma}(X,Z)}\big(-\hat{c}_n - \lambda_2\widetilde{J}_{x,\theta}\big)\widetilde{J}_{x,\theta}\}.
\end{aligned}$$

Since $f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\cdot) \leq f_o^{1/4}$ a.s. by Assumption 3(c), $|\widetilde{J}_{x,\theta}| = \mathcal{O}_p(n^{-1/2})$ by Assumption 1(c) cum Theorem 2, and $\frac{\widetilde{J}_{x,\theta}}{c_n} = o_p(1)$ by Assumption 4, $\frac{\mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}[\tilde{I}^1_{ij}] + \mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}[\tilde{I}^2_{ij}]}{2c_n} = o_p(1)$.

It follows from the foregoing, the LIE, the CS inequality, Lyapunov's inequality, Assumption 1(c), and the identical sampling of the data (Assumption 1(a)) that

$$\begin{aligned}
\mathbb{E}[A_{n,1}] &= \frac{1}{n^2 c_n}\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}\Big[\mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}[|\mathrm{I}(|\widetilde{U}_{ij}(\widehat{\theta}_n)| \leq \hat{c}_n) - \mathrm{I}(|\widetilde{U}_{ij}| \leq \hat{c}_n)|] \times ||\mathcal{Z}_{ij}\widetilde{X}^g_{ij}(\widehat{\theta}_n)'\widetilde{X}^g_{ij}(\widehat{\theta}_n)||\Big] \\
&\leq \frac{2}{n^2}\sum_{i=1}^n\sum_{j=1}^n \Big\{\mathbb{E}\Big[\Big(\frac{\mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}[\tilde{I}^1_{ij}] + \mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}[\tilde{I}^2_{ij}]}{2c_n}\Big)^2\Big] \times \mathbb{E}\big[||\max\{|\mathcal{Z}_{ij}|, 1\}\widetilde{X}^g_{ij}(\widehat{\theta}_n)||^4\big]\Big\}^{1/2} \\
&\leq 2\sup_{\theta\in\Theta}(\mathbb{E}\big[||\max\{|\mathcal{Z}|, 1\}\widetilde{X}^g(\theta)||^4\big])^{1/2} \times \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \Big\{\mathbb{E}\Big[\Big(\frac{\mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}[\tilde{I}^1_{ij}] + \mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}[\tilde{I}^2_{ij}]}{2c_n}\Big)^2\Big]\Big\}^{1/2}.
\end{aligned}$$

From the foregoing, $A_{n,1} = o_p(1)$ thanks to the Markov inequality.

$A_{n,2}$:

First, by Assumption 3(c), eq. (C.1), and the MVT,

$$\mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}[\mathrm{I}(|\widetilde{U}_{ij}| \leq \hat{c}_n)]/(2c_n) = f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\lambda \hat{c}_n)(\hat{c}_n/c_n)$$

for some $\lambda \in (0,1)$. It follows from the LIE, Assumption 3(c), Assumption 4, the CS inequality, the continuity of the Jacobian (Assumption 1(b)), the continuous mapping theorem (CMT), and the consistency of the MDep (Theorem 1) that

$$\begin{aligned}
\mathbb{E}[A_{n,2}] &= \frac{1}{n^2 c_n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\Big[\mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}[\mathrm{I}(|\widetilde{U}_{ij}| \leq \hat{c}_n)] \times ||\mathcal{Z}_{ij}[\widetilde{X}_{ij}^g(\widehat{\theta}_n)'\widetilde{X}_{ij}^g(\widehat{\theta}_n) - \widetilde{X}_{ij}^{g\prime}\widetilde{X}_{ij}^g]||\Big] \\
&\leq \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big\{\Big(\mathbb{E}\big[\big((\hat{c}_n/c_n)f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\lambda\hat{c}_n)\big)^2\big]\Big)^{1/2}\Big(\mathbb{E}\big[\mathcal{Z}_{ij}^2||\widetilde{X}_{ij}^g(\widehat{\theta}_n)'\widetilde{X}_{ij}^g(\widehat{\theta}_n) - \widetilde{X}_{ij}^{g\prime}\widetilde{X}_{ij}^g||^2\big]\Big)^{1/2}\Big\} \\
&\leq 4C^{3/4} \times o(1) = o(1)
\end{aligned}$$

noting in particular that $\rho(\theta) := \mathbb{E}\big[\mathcal{Z}^2||\widetilde{X}^g(\theta)'\widetilde{X}^g(\theta) - \widetilde{X}^g(\theta_o)'\widetilde{X}^g(\theta_o)||^2\big]$ under Assumption 1(b) is continuous in θ. From the foregoing, $A_{n,2} = o_p(1)$ thanks to the Markov inequality.

$A_{n,3}$:

$$\begin{aligned}
&\mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}[|\mathrm{I}(|\widetilde{U}_{ij}| \leq \hat{c}_n) - \mathrm{I}(|\widetilde{U}_{ij}| \leq c_n)|]/(2c_n) = \frac{\mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}[\mathrm{I}(c_n < \widetilde{U}_{ij} \leq \hat{c}_n) + \mathrm{I}(-\hat{c}_n \leq \widetilde{U}_{ij} < -c_n)]}{2c_n} \\
&= \max\left\{0, \frac{F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\hat{c}_n) - F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(c_n)}{2c_n}\right\} + \max\left\{0, \frac{F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(-c_n) - F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(-\hat{c}_n)}{2c_n}\right\} \\
&= \max\{0, 0.5 f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\lambda_1\hat{c}_n + (1-\lambda_1)\hat{c}_n)(\hat{c}_n/c_n - 1)\} \\
&+ \max\{0, -0.5 f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(-\lambda_2 c_n - (1-\lambda_2)\hat{c}_n)(\hat{c}_n/c_n - 1)\}
\end{aligned}$$

for some $\lambda_1, \lambda_2 \in (0,1)$ by Assumption 3(c), Assumption 4, eq. (C.1), and the MVT. By arguments analogous to the case of $A_{n,1}$, $A_{n,3} = o_p(1)$. □

To complete the proof of consistency of $\widehat{\mathcal{H}}_n$, it suffices to show that $\mathcal{H}_n - \mathcal{H}$ converges to zero in quadratic mean. This is shown in the following lemma.

Lemma C.2. *Under Assumptions 1((a) and (c)), 3(c), and 4, $\mathcal{H}_n - \mathcal{H}$ converges to zero in quadratic mean.*

Proof. $|\mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}[\mathrm{I}(|\widetilde{U}_{ij}| \leq c_n)]/(2c_n) - f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(0)| = |f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\lambda c_n) - f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(0)| \leq \lambda f_o^{1/4} c_n$ a.s. for some $\lambda \in (0,1)$ by Assumption 3(c) and the MVT. In addition to Assumption 1(c), Assumption 4, and the CS inequality, this implies

$$
\begin{aligned}
&||\mathbb{E}[\mathcal{H}_n] - \mathcal{H}|| \leq \\
&\big(\mathbb{E}[\big(\mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}[\mathrm{I}(|\widetilde{U}_{ij}| \leq c_n)]/(2c_n) - f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(0)\big)^2]\big)^{1/2} \big(\mathbb{E}[||\max\{|\mathcal{Z}_{ij}|,1\}\widetilde{X}_{ij}^g||^4]\big)^{1/2} \\
&\leq \lambda f_o^{1/4} C^{1/2} c_n = \mathcal{O}(c_n) = o(1).
\end{aligned}
$$

Let τ_1 and τ_2 be two $p_x \times 1$ vectors with $||\tau_1|| = ||\tau_2|| = 1$, then

$$
\begin{aligned}
&\mathrm{var}(\tau_1' \mathcal{H}_n \tau_2) \\
&= \frac{1}{n^4 c_n^2} \sum_{i=1}^n \sum_{j=1}^n \sum_{i'=1}^n \sum_{j'=1}^n \mathrm{cov}\Big(\{\mathrm{I}(|\widetilde{U}_{ij}| \leq c_n)\mathcal{Z}_{ij}\tau_1'\widetilde{X}_{ij}^{g\prime}\widetilde{X}_{ij}^g\tau_2\}, \{\mathrm{I}(|\widetilde{U}_{i'j'}| \leq c_n)\mathcal{Z}_{i'j'}\tau_1'\widetilde{X}_{i'j'}^{g\prime}\widetilde{X}_{i'j'}^g\tau_2\}\Big) \\
&= \frac{1}{n^4 c_n^2} \sum_{i=1}^n \sum_{j=1}^n \mathrm{var}\Big(\mathrm{I}(|\widetilde{U}_{ij}| \leq c_n)\mathcal{Z}_{ij}\tau_1'\widetilde{X}_{ij}^{g\prime}\widetilde{X}_{ij}^g\tau_2\Big) \\
&+ \frac{2}{n^4 c_n^2} \sum_{i=1}^n \sum_{j=1}^n \sum_{i'=1}^n \mathrm{cov}\Big(\{\mathrm{I}(|\widetilde{U}_{ij}| \leq c_n)\mathcal{Z}_{ij}\tau_1'\widetilde{X}_{ij}^{g\prime}\widetilde{X}_{ij}^g\tau_2\}, \{\mathrm{I}(|\widetilde{U}_{i'j}| \leq c_n)\mathcal{Z}_{i'j}\tau_1'\widetilde{X}_{i'j}^{g\prime}\widetilde{X}_{i'j}^g\tau_2\}\Big) \\
&\leq \frac{1}{n^4 c_n^2} \sum_{i=1}^n \sum_{j=1}^n \mathrm{var}\Big(\mathrm{I}(|\widetilde{U}_{ij}| \leq c_n)\mathcal{Z}_{ij}\tau_1'\widetilde{X}_{ij}^{g\prime}\widetilde{X}_{ij}^g\tau_2\Big) \\
&+ \frac{2}{n^4 c_n^2} \sum_{i=1}^n \sum_{j=1}^n \sum_{i'=1}^n \Big(\mathrm{var}\big(\mathrm{I}(|\widetilde{U}_{ij}| \leq c_n)\tau_1'\mathcal{Z}_{ij}\widetilde{X}_{ij}^{g\prime}\widetilde{X}_{ij}^g\tau_2\big) \cdot \mathrm{var}\big(\mathrm{I}(|\widetilde{U}_{i'j}| \leq c_n)\tau_1'\mathcal{Z}_{i'j}\widetilde{X}_{i'j}^{g\prime}\widetilde{X}_{i'j}^g\tau_2\big)\Big)^{1/2} \\
&\leq \frac{1}{n^4 c_n^2} \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}[||\max\{|\mathcal{Z}_{ij}|,1\}\widetilde{X}_{ij}^g||^4] \\
&+ \frac{2}{n^4 c_n^2} \sum_{i=1}^n \sum_{j=1}^n \sum_{i'=1}^n \Big(\mathbb{E}[||\max\{|\mathcal{Z}_{ij}|,1\}\widetilde{X}_{ij}^g||^4] \cdot \mathbb{E}[||\max\{|\mathcal{Z}_{i'j}|,1\}\widetilde{X}_{i'j}^g||^4]\Big)^{1/2} \\
&\leq \frac{C}{n^2 c_n^2} + \frac{2C}{n c_n^2}.
\end{aligned}
$$

The second equality follows from Assumption 1(a), the first inequality follows from the CS

inequality, and the last inequality follows from Jensen's inequality. The second inequality holds because

$$\begin{aligned}\text{var}(\tau_1' M \tau_2) &\leq \mathbb{E}[(\tau_1' M \tau_2)^2] = \mathbb{E}[(\text{vec}(\tau_1' M \tau_2))^2] = \mathbb{E}[\text{vec}(M)'(\tau_2' \otimes \tau_1')'(\tau_2' \otimes \tau_1')\text{vec}(M)] \\ &\leq \mathbb{E}[||\text{vec}(M)||^2 \cdot ||\tau_1' \otimes \tau_2'||^2] = \mathbb{E}[||\text{vec}(M)||^2 \cdot ||\tau_1||^2 \cdot ||\tau_2||^2] = \mathbb{E}[||M||^2]\end{aligned}$$

for a matrix-valued random variable M, and $||\tau_1' \otimes \tau_2'|| = ||\tau_1|| \cdot ||\tau_2||$ by Bernstein (2009, Fact 9.7.27). Thanks to Assumptions 1(c) and 4, $\text{var}(\tau_1' \mathcal{H}_n \tau_2) \leq 3C/(nc_n^2) = o(1)$, and the assertion is proved as claimed. □

Combining Lemmata C.1 and C.2 shows that $\operatorname{plim}_{n\to\infty} \widehat{\mathcal{H}}_n = \mathcal{H}$. The next part of the proof of Theorem 3 concerns the consistency of $\widehat{\Omega}_n$. The result is stated in the following lemma.

Lemma C.3. *Under Assumptions 1 to 3,* $\operatorname{plim}_{n\to\infty} \widehat{\Omega}_n = \Omega$.

Proof. Recall $\widehat{\Omega}_n = 4\mathbb{E}_n[\widehat{\psi}^{(1)}(W_i)\widehat{\psi}^{(1)}(W_i)']$ where

$$\widehat{\psi}^{(1)}(W_i) \equiv \frac{1}{n-1}\sum_{j=1}^{n} \mathcal{Z}_{ij,n}\big(1 - 2\text{I}(\widetilde{U}_{ij}(\widehat{\theta}_n) < 0)\big)\widetilde{X}_{ij}^{g}(\widehat{\theta}_n)'.$$

Define $\widetilde{\Omega}_n = 4\mathbb{E}_n[\widetilde{\psi}^{(1)}(W_i)\widetilde{\psi}^{(1)}(W_i)']$ where

$$\widetilde{\psi}^{(1)}(W_i) \equiv \frac{1}{n-1}\sum_{j=1}^{n} \mathcal{Z}_{ij}\big(1 - 2\text{I}(\widetilde{U}_{ij} < 0)\big)\widetilde{X}_{ij}^{g\prime}.$$

Since $||\widehat{\Omega}_n - \Omega|| \leq ||\widehat{\Omega}_n - \widetilde{\Omega}_n|| + ||\widetilde{\Omega}_n - \Omega||$ by the triangle inequality, the proof proceeds by

showing that $||\widehat{\Omega}_n - \widetilde{\Omega}_n|| = o_p(1)$ and $||\widetilde{\Omega}_n - \Omega|| = o_p(1)$. First,

$$
\begin{aligned}
||\widehat{\Omega}_n - \widetilde{\Omega}_n|| &= 4||\mathbb{E}_n[\widehat{\psi}^{(1)}(W_i)\widehat{\psi}^{(1)}(W_i)' - \widetilde{\psi}^{(1)}(W_i)\widetilde{\psi}^{(1)}(W_i)']|| \\
&= 4||\mathbb{E}_n[\widehat{\psi}^{(1)}(W_i)(\widehat{\psi}^{(1)}(W_i) - \widetilde{\psi}^{(1)}(W_i))' + (\widehat{\psi}^{(1)}(W_i) - \widetilde{\psi}^{(1)}(W_i))\widetilde{\psi}^{(1)}(W_i)']|| \\
&\leq 4\mathbb{E}_n\big[\big(||\widehat{\psi}^{(1)}(W_i)|| + ||\widetilde{\psi}^{(1)}(W_i)||\big)||\widehat{\psi}^{(1)}(W_i) - \widetilde{\psi}^{(1)}(W_i)||\big] \\
&\leq 4\big(\mathbb{E}_n\big[(||\widehat{\psi}^{(1)}(W_i)|| + ||\widetilde{\psi}^{(1)}(W_i)||)^2\big]\big)^{1/2} \times \big(\mathbb{E}_n\big[||\widehat{\psi}^{(1)}(W_i) - \widetilde{\psi}^{(1)}(W_i)||^2\big]\big)^{1/2}
\end{aligned}
$$

by Jensen's and the CS inequalities.

Second, obtain the following upper bound:

$$
\begin{aligned}
||\widehat{\psi}^{(1)}(W_i) - \widetilde{\psi}^{(1)}(W_i)|| &\leq \frac{1}{n-1}\sum_{j=1}^{n} ||\mathcal{Z}_{ij,n}\big(1 - 2\mathrm{I}(\widetilde{U}_{ij}(\widehat{\theta}_n) < 0)\big)\widetilde{X}^g_{ij}(\widehat{\theta}_n)' - \mathcal{Z}_{ij}\big(1 - 2\mathrm{I}(\widetilde{U}_{ij} < 0)\big)\widetilde{X}^{g\prime}_{ij}|| \\
&\leq \frac{1}{n-1}\sum_{j=1}^{n} ||(\mathcal{Z}_{ij,n} - \mathcal{Z}_{ij})\big(1 - 2\mathrm{I}(\widetilde{U}_{ij}(\widehat{\theta}_n) < 0)\big)\widetilde{X}^g_{ij}(\widehat{\theta}_n)|| \\
&+ \frac{2}{n-1}\sum_{j=1}^{n} ||\big(\mathrm{I}(\widetilde{U}_{ij}(\widehat{\theta}_n) < 0) - \mathrm{I}(\widetilde{U}_{ij} < 0)\big)\mathcal{Z}_{ij}\widetilde{X}^g_{ij}(\widehat{\theta}_n)|| \\
&+ \frac{1}{n-1}\sum_{j=1}^{n} ||\big(1 - 2\mathrm{I}(\widetilde{U}_{ij} < 0)\big)\mathcal{Z}_{ij}(\widetilde{X}^g_{ij}(\widehat{\theta}_n) - \widetilde{X}^g_{ij})|| \\
&\leq \frac{1}{n-1}\sum_{j=1}^{n} |\mathcal{Z}_{ij,n} - \mathcal{Z}_{ij}| \cdot \sup_{\theta\in\Theta} ||\widetilde{X}^g_{ij}(\theta)|| \\
&+ \frac{2}{n-1}\sum_{j=1}^{n} |\mathrm{I}(\widetilde{U}_{ij}(\widehat{\theta}_n) < 0) - \mathrm{I}(\widetilde{U}_{ij} < 0)| \cdot \sup_{\theta\in\Theta} ||\mathcal{Z}_{ij}\widetilde{X}^g_{ij}(\theta)|| \\
&+ \frac{1}{n-1}\sum_{j=1}^{n} ||\mathcal{Z}_{ij}(\widetilde{X}^g_{ij}(\widehat{\theta}_n) - \widetilde{X}^g_{ij})|| \\
&\equiv B_{1i,n} + B_{2i,n} + B_{3i,n}.
\end{aligned}
$$

By the c_r-inequality, $\mathbb{E}_n[||\widehat{\psi}^{(1)}(W_i) - \widetilde{\psi}^{(1)}(W_i)||^2] \leq 3\mathbb{E}_n[B^2_{1i,n}] + 3\mathbb{E}_n[B^2_{2i,n}] + 3\mathbb{E}_n[B^2_{3i,n}]$. It can be observed that $\mathbb{E}_n[B^2_{1i,n}] = \mathcal{O}_p(n^{-1/2})$ under the assumptions of Lemma E.1 and Assumption 1(c), while $\mathbb{E}_n[B^2_{3i,n}] = o_p(1)$ by the continuous mapping theorem (CMT) and Theorem 1(b). It remains to show that $\mathbb{E}_n[B^2_{2i,n}] = o_p(1)$.

$$
\begin{aligned}
|\mathrm{I}(\widetilde{U}(\theta)<0)-\mathrm{I}(\widetilde{U}<0)| &= |\mathrm{I}(\widetilde{U}+\widetilde{X^g}(\bar{\theta})(\theta-\theta_o)<0)-\mathrm{I}(\widetilde{U}<0)| \\
&= \mathrm{I}(0\le\widetilde{U}<-\widetilde{X^g}(\bar{\theta})(\theta-\theta_o))+\mathrm{I}(-\widetilde{X^g}(\bar{\theta})(\theta-\theta_o)\le\widetilde{U}<0) \\
&\equiv \tilde{I}^a(\theta)+\tilde{I}^b(\theta). \qquad \text{(C.2)}
\end{aligned}
$$

Further,

$$
\begin{aligned}
\mathbb{E}[\tilde{I}^a(\theta)+\tilde{I}^b(\theta)|\widetilde{X^g},\mathcal{Z}] &= \max\left\{0, F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(-\widetilde{X^g}(\bar{\theta})(\theta-\theta_o))-F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(0)\right\} \\
&\quad+\max\left\{0, F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(0)-F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(-\widetilde{X^g}(\bar{\theta})(\theta-\theta_o))\right\} \\
&= 2|F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(-\widetilde{X^g}(\bar{\theta})(\theta-\theta_o))-F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(0)| \\
&= 2f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(-\lambda\widetilde{X^g}(\bar{\theta})(\theta-\theta_o))\times|\widetilde{X^g}(\bar{\theta})(\theta-\theta_o)| \qquad \text{(C.3)}
\end{aligned}
$$

by the MVT and the Schwartz inequality for some $\lambda\in(0,1)$. Thanks to the foregoing, the identical distribution of data, and recalling $\mathcal{Z}_{ii}=0$ for $i=1,\ldots,n$, one has

$$
\begin{aligned}
\mathbb{E}[B_{2i,n}] &= \mathbb{E}\big[|\mathrm{I}(\widetilde{U}(\widehat{\theta}_n)<0)-\mathrm{I}(\widetilde{U}<0)|\cdot\sup_{\theta\in\Theta}||\mathcal{Z}\widetilde{X}^g(\theta)||\big] \\
&\le \big(\mathbb{E}\big[|\mathrm{I}(\widetilde{U}(\widehat{\theta}_n)<0)-\mathrm{I}(\widetilde{U}<0)|^2\big]\big)^{1/2}\cdot\big(\mathbb{E}\big[\sup_{\theta\in\Theta}||\mathcal{Z}\widetilde{X}^g(\theta)||^2\big]\big)^{1/2} \\
&\le \big(\mathbb{E}\big[|\mathrm{I}(\widetilde{U}(\widehat{\theta}_n)<0)-\mathrm{I}(\widetilde{U}<0)|\big]\big)^{1/2}\cdot\big(\mathbb{E}\big[\sup_{\theta\in\Theta}||\mathcal{Z}\widetilde{X}^g(\theta)||^4\big]\big)^{1/4} \\
&\le 2C^{1/4}\big(\mathbb{E}\big[f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(-\lambda\widetilde{X^g}(\bar{\theta}_n)(\widehat{\theta}_n-\theta_o))\times|\widetilde{X^g}(\bar{\theta}_n)(\widehat{\theta}_n-\theta_o)|\big]\big)^{1/2} \\
&= o(1).
\end{aligned}
$$

The first inequality follows by the CS, the second by the Lyapunov inequality and that $|\mathrm{I}(\widetilde{U}(\widehat{\theta}_n)<0)-\mathrm{I}(\widetilde{U}<0)|^2=|\mathrm{I}(\widetilde{U}(\widehat{\theta}_n)<0)-\mathrm{I}(\widetilde{U}<0)|$. The third inequality uses the LIE, Equation (C.2), Equation (C.3), the continuous mapping theorem, the consistency of $\widehat{\theta}_n$ (Theorem 1), and Assumption 1(c). Thus, $\mathbb{E}_n[B_{2i,n}]=o_p(1)$ by the Markov inequality.

From the foregoing, $\operatorname{plim}_{n\to\infty} ||\widehat{\Omega}_n - \widetilde{\Omega}_n|| = 0$.

$$
\begin{aligned}
\mathbb{E}[\widetilde{\Omega}_n] &= \frac{1}{n(n-1)^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{j'=1}^{n}\mathbb{E}[\psi(W_i,W_j)\psi(W_i,W_{j'})'] = \mathbb{E}[\psi(W,W^{\dagger})\psi(W,W^{\dagger\dagger})'] \\
&= \mathbb{E}\big[\mathbb{E}[\psi(W,W^{\dagger})\psi(W,W^{\dagger\dagger})'|W]\big] = \mathbb{E}\big[\mathbb{E}[\psi(W,W^{\dagger})|W]\cdot\mathbb{E}[\psi(W,W^{\dagger\dagger})|W]'\big] \\
&= \mathbb{E}[\psi^{(1)}(W)\psi^{(1)}(W)'] \equiv \Omega
\end{aligned} \tag{C.4}
$$

and

(C.5)

$$
\begin{aligned}
\mathbb{E}[||\psi(W_i,W_j)\psi(W_i,W_{j'})'||] &\leq \mathbb{E}[||\psi(W_i,W_j)||\cdot||\psi(W_i,W_{j'})||] \\
&\leq (\mathbb{E}[||\psi(W_i,W_j)||^2]\cdot\mathbb{E}[||\psi(W_i,W_{j'})'||^2])^{1/2} \\
&= \mathbb{E}[||\psi(W,W^{\dagger})||^2] = \mathbb{E}[||\mathcal{Z}\widetilde{X^g}||^2] \leq (\mathbb{E}[||\mathcal{Z}\widetilde{X^g}||^4])^{1/2} \leq C^{1/2}.
\end{aligned}
$$

$\widetilde{\Omega}_n$ is a U-statistic of order 3. Combining (C.4) and (C.5), $||\widetilde{\Omega}_n - \Omega|| = o_p(1)$ by the strong law of large numbers for U-statistics – see Hoeffding (1961). □

D Proof of Theorem 4

Let $\tau^* = \arg\inf_{\{\tau\in\mathbb{R}^{p_x}:||\tau||=1\}} \mathcal{V}^2(X\tau, Z)$. By Properties (a) and (b), a test of Assumption 2(b) can be formulated as

$$
\begin{aligned}
\mathbb{H}_o' &: \mathcal{V}^2(X\tau^*, Z) = 0 \\
\mathbb{H}_a' &: \mathcal{V}^2(X\tau^*, Z) > 0
\end{aligned}
$$

Partition $\tau^* = [\tau_1^*, \tau_{-1}^*]$ conformably, then $X\tau^* = \tau_1^* D + \dot{X}\tau_{-1}^*$. The rest of the proof relies on the following lemma.

Lemma D.1. *$\mathbb{H}_o'$ implies $\tau_1^* \neq 0$, while the converse does not hold.*

Proof. The first part of the proof proceeds by contradiction. Suppose $\tau_1^* = 0$. Then $\mathcal{V}^2(X\tau^*, Z) = \mathcal{V}^2(\tau_1^* D + \dot{X}\tau_{-1}^*, Z) = \mathcal{V}^2(\dot{X}\tau_{-1}^*, Z) > 0$ since Z contains $\dot{X}$, i.e., $\tau_1^* = 0$ implies $\mathbb{H}_a'$, thus $\mathbb{H}_o'$ implies $\tau_1^* \neq 0$.

In examining the converse, consider two cases of $\tau_1^* \neq 0$. First, $|\tau_1^*| \in (0,1)$ means $\tau_{-1}^* \neq 0$ hence $\mathcal{V}^2(X\tau^*, Z) = \mathcal{V}^2(\tau_1^* D + \dot{X}\tau_{-1}^*, Z) > 0$ since Z contains $\dot{X}$, i.e., $|\tau_1^*| \in (0,1)$ implies $\mathbb{H}_a'$. Second, $|\tau_1^*| = 1$ means $\tau_{-1}^* = 0$, thus $\mathcal{V}^2(X\tau^*, Z) = \mathcal{V}^2(\tau_1^* D + \dot{X}\tau_{-1}^*, Z) = \mathcal{V}^2(D, Z) \geq 0$ by Property (a), i.e., $|\tau_1^*| = 1$ implies either $\mathbb{H}_o'$ or $\mathbb{H}_a'$ depending on whether D is dependent on Z or not. □

With Lemma D.1 in hand, observe that for $\gamma^* = -\tau_{-1}^*/\tau_1^*$,

$$\begin{aligned}
\mathcal{V}^2(X\tau^*, Z) = \mathbb{E}[\mathcal{Z}|\widetilde{X}\tau^*|] &= \mathbb{E}[\mathcal{Z}|(D - D^\dagger)\tau_1^* + (\dot{X} - \dot{X}^\dagger)\tau_{-1}^*|] \\
&= |\tau_1^*|\mathbb{E}[\mathcal{Z}|(D - D^\dagger) + (\dot{X} - \dot{X}^\dagger)\tau_{-1}^*/\tau_1^*|] \\
&= |\tau_1^*|\mathcal{V}^2(D - \dot{X}\gamma^*, Z) \\
&= |\tau_1^*|\mathcal{V}^2(\mathcal{E}(\gamma^*), Z).
\end{aligned}$$

As $|\tau_1^*| \neq 0$, $\mathcal{V}^2(X\tau^*, Z) = 0$ is equivalent to $\mathcal{V}^2(\mathcal{E}(\gamma^*), Z) = 0$, and $\mathcal{V}^2(X\tau^*, Z) > 0$ is equivalent to $\mathcal{V}^2(\mathcal{E}(\gamma^*), Z) > 0$. The conclusion follows from Properties (a) and (b). □

Online Appendix

E Supporting Lemmata

E.1 Almost Sure Convergence of U-centred distance

Lemma E.1. *Under Assumptions 1(a), 1(b), 1(d), and 3(d),* $\sqrt{n}|\mathcal{Z}_{ij,n} - \mathcal{Z}_{ij}| = \mathcal{O}_p(1)$ *for* $j \neq i = 1, \ldots, n$.

Proof.

For any i, j, $1 \leq i, j \leq n$, $\mathbb{E}[\mathcal{Z}_{ij,n} - \mathcal{Z}_{ij}] = 0$ by the LIE. Moreover, it follows from Loève's c_r-inequality,[9] Assumption 1(a), the CS inequality, and Assumption 3(d) that

(E.1)

$$
\begin{aligned}
\mathbb{E}[|\mathcal{Z}_{ij,n} - \mathcal{Z}_{ij}|^2] &\leq \frac{3}{(n-2)^2}\mathbb{E}\Big[\Big(\sum_{k=1}^{n}(||\widetilde{Z}_{ik}|| - \mathbb{E}[(||\widetilde{Z}_{ik}||)|Z_i])\Big)^2\Big] \\
&+ \frac{3}{(n-2)^2}\mathbb{E}\Big[\Big(\sum_{k=1}^{n}(||\widetilde{Z}_{kj}|| - \mathbb{E}[(||\widetilde{Z}_{kj}||)|Z_j])\Big)^2\Big] \\
&+ \frac{3}{(n-1)^2(n-2)^2}\mathbb{E}\Big[\Big(\sum_{k=1}^{n}\sum_{l=1}^{n}(||\widetilde{Z}_{kl}|| - \mathbb{E}[||\widetilde{Z}_{kl}||])\Big)^2\Big] \\
&= \frac{3}{(n-2)^2}\sum_{k=1}^{n}\mathbb{E}[\mathrm{var}((||\widetilde{Z}_{ik}||)|Z_i)] + \frac{3}{(n-2)^2}\sum_{k=1}^{n}\mathbb{E}[\mathrm{var}((||\widetilde{Z}_{kj}||)|Z_j)] \\
&+ \frac{3}{(n-1)^2(n-2)^2}\sum_{k=1}^{n}\sum_{l=1}^{n}\mathrm{var}(||\widetilde{Z}_{kl}||) + \frac{6}{(n-1)^2(n-2)^2}\sum_{k=1}^{n}\sum_{l=1}^{n}\sum_{l'\neq l}\mathrm{cov}(||\widetilde{Z}_{kl}||, ||\widetilde{Z}_{kl'}||) \\
&\leq \frac{6n}{(n-2)^2}\mathbb{E}[||\widetilde{Z}||^2] + \frac{3n^2}{(n-1)^2(n-2)^2}\mathbb{E}[||\widetilde{Z}||^2] \\
&+ \frac{6n^2(n-1)}{(n-1)^2(n-2)^2}\sqrt{\mathbb{E}[||\widetilde{Z}_{kl}||^2]\cdot\mathbb{E}[||\widetilde{Z}_{kl'}||^2]} \\
&= \mathcal{O}(n^{-1}).
\end{aligned}
$$

Thus, $\mathbb{E}[|\mathcal{Z}_{ij,n} - \mathcal{Z}_{ij}|] \leq (\mathbb{E}[|\mathcal{Z}_{ij,n} - \mathcal{Z}_{ij}|^2])^{1/2} = \mathcal{O}(n^{-1/2})$ by Lyapunov's inequality. The

[9] As applied here, the inequality is $\mathbb{E}[(\xi_1 + \xi_2 + \xi_3)^2] \leq 3\mathbb{E}[\xi_1^2] + 3\mathbb{E}[\xi_2^2] + 3\mathbb{E}[\xi_3^2]$. See Davidson (1994, eqn. 9.62).

conclusion follows from Markov's inequality. □

E.2 The Hessian Matrix

Lemma E.2. *Under Assumptions 1(b), 2(a), 3(c), then the Hessian matrix is given by* $\mathcal{H} = 2\mathbb{E}\Big[f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(0)\mathcal{Z}\widetilde{X^g}'\widetilde{X^g}\Big]$.

Proof. Under the assumptions of Lemma A.1 and Property (d),

$$\mathbb{E}[\widehat{\mathcal{S}}_n(\theta)] = \mathbb{E}\Big[\frac{\partial Q_n(\theta)}{\partial\theta}\Big] = \frac{\partial\mathbb{E}[Q_n(\theta)]}{\partial\theta} = \frac{\partial Q(\theta)}{\partial\theta} \equiv \mathcal{S}(\theta)$$

by the dominated convergence theorem.

By the LIE and using $\widetilde{U}(\theta) = \widetilde{U} + \widetilde{X^g}(\bar{\theta})(\theta - \theta_o)$ by the MVT and Assumption 1(b),

$$\begin{aligned}
\mathbb{E}[\widehat{\mathcal{S}}_n(\theta)] = \mathcal{S}(\theta) &= \mathbb{E}\big[(1 - 2\mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}[\mathrm{I}\big(\widetilde{U}(\theta) < 0\big)]\big)\mathcal{Z}\widetilde{X^g}(\theta)\big] \\
&= \mathbb{E}\big[(1 - 2\mathbb{E}_{\widetilde{U}|\tilde{\sigma}(X,Z)}[\mathrm{I}(\widetilde{U} < -\widetilde{X^g}(\bar{\theta})(\theta - \theta_o))])\mathcal{Z}\widetilde{X^g}(\theta)\big] \\
&= \mathbb{E}\big[\big(1 - 2F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(-\widetilde{X^g}(\bar{\theta})(\theta - \theta_o))\big)\mathcal{Z}\widetilde{X^g}(\theta)\big].
\end{aligned}$$

Under the assumptions of Lemma E.3, the expectation and derivative are exchangeable by the dominated convergence theorem. The expression for $\mathcal{H}(\theta) \equiv \dfrac{\partial\mathbb{E}[\widehat{\mathcal{S}}_n(\theta)]}{\partial\theta} = \dfrac{\partial\mathcal{S}(\theta)}{\partial\theta}$ becomes

$$\begin{aligned}
\mathcal{H}(\theta) = 2\mathbb{E}\Big[&f_{\widetilde{U}|\tilde{\sigma}(X,Z)}\big(-\widetilde{X^g}(\bar{\theta})(\theta - \theta_o)\big)\mathcal{Z}\widetilde{X^g}(\theta)'\widetilde{X^g}(\bar{\theta})\Big] \\
&+ \mathbb{E}\Big[\big(1 - 2F_{\widetilde{U}|\tilde{\sigma}(X,Z)}\big(-\widetilde{X^g}(\bar{\theta})(\theta - \theta_o)\big)\big)\mathcal{Z}\frac{\partial\widetilde{X^g}(\theta)}{\partial\theta}\Big].
\end{aligned}$$

Evaluating $\mathcal{H}(\theta)$ at $\theta = \theta_o$ gives $\mathcal{H} = 2\mathbb{E}[f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(0)\mathcal{Z}\widetilde{X^g}'\widetilde{X^g}]$ because $\bar{\theta}$ satisfies $||\bar{\theta} - \theta_o|| \leq ||\theta - \theta_o||$ and $F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(0) = 1/2$ by Assumption 2(a). □

The result in the following lemma is essential to applying the dominated convergence

theorem in the proof of Lemma E.2. Define

$$\eta(\theta) \equiv \Big[2f_{\widetilde{U}|\tilde{\sigma}(X,Z)}\big(\widetilde{X^g}(\bar{\theta})(\theta-\theta_o)\big)\mathcal{Z}\widetilde{X^g}(\theta)'\widetilde{X^g}(\bar{\theta})\Big] + \Big[\Big(1-2F_{\widetilde{U}|\tilde{\sigma}(X,Z)}\big(\widetilde{X^g}(\bar{\theta})(\theta-\theta_o)\big)\Big)\mathcal{Z}\frac{\partial\widetilde{X^g}(\theta)}{\partial\theta}\Big]$$
$$\equiv \eta^A(\theta)+\eta^B(\theta)$$

The following additional condition is imposed.

Assumption 5. $\mathbb{E}\big[\sup_{\theta\in\Theta}\big|\big|\mathcal{Z}\frac{\partial\widetilde{X^g}(\theta)}{\partial\theta}\big|\big|^4\big] \le C.$

The following lemma derives a bound on each summand of $\eta(\theta)$.

Lemma E.3. *Under Assumptions 1(c), 3(c) and Assumption 5,* $\mathbb{E}\Big[\sup_{\theta\in\Theta}\Big|\Big|\eta^A(\theta)\Big|\Big|\Big] \le 2f_oC^{1/2}$ *and* $\mathbb{E}\Big[\sup_{\theta\in\Theta}\Big|\Big|\eta^B(\theta)\Big|\Big|\Big] \le C^{1/4}.$

Proof of Lemma E.3. For any $\theta\in\Theta$,

$$||\eta^A(\theta)|| = \Big|\Big|2\mathcal{Z}f_{\widetilde{U}|\tilde{\sigma}(X,Z)}\big(\widetilde{X^g}(\bar{\theta})(\theta-\theta_o)\big)\widetilde{X^g}(\theta)'\widetilde{X^g}(\bar{\theta})\Big|\Big| \le 2f_{\widetilde{U}|\tilde{\sigma}(X,Z)}\big(\widetilde{X^g}(\bar{\theta})(\theta-\theta_o)\big)||\mathcal{Z}\widetilde{X^g}(\theta)'\widetilde{X^g}(\bar{\theta})||$$
$$\le 2f_{\widetilde{U}|\tilde{\sigma}(X,Z)}\big(\widetilde{X^g}(\bar{\theta})(\theta-\theta_o)\big)\cdot\sup_{\theta\in\Theta}||\max\{|\mathcal{Z}|,1\}\widetilde{X^g}(\theta)||^2$$

by the Schwartz inequality and

$$||\eta^B(\theta)|| = \Big|1-2F_{\widetilde{U}|\tilde{\sigma}(X,Z)}\big(\widetilde{X^g}(\bar{\theta})(\theta-\theta_o)\big)\Big| \times \Big|\Big|\mathcal{Z}\frac{\partial\widetilde{X^g}(\theta)}{\partial\theta}\Big|\Big| \le \Big|\Big|\mathcal{Z}\frac{\partial\widetilde{X^g}(\theta)}{\partial\theta}\Big|\Big|$$

noting that $\Big|1-2F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(\cdot)\Big| \le 1.$

From the foregoing, the CS inequality, the Lyapunov inequality, and Assumptions 1(c), 3(c), and 5,

$$\mathbb{E}[\sup_{\theta\in\Theta}||\eta^A(\theta)||] \le 2\big(\mathbb{E}\big[\big(\sup_{\theta\in\Theta}f_{\widetilde{U}|\tilde{\sigma}(X,Z)}\big(\widetilde{X^g}(\bar{\theta})(\theta-\theta_o)\big)\big)^4\big]\big)^{1/4}\big(\mathbb{E}\big[\sup_{\theta\in\Theta}\big|\big|\max\{|\mathcal{Z}|,1\}\widetilde{X^g}(\theta)\big|\big|^4\big]\big)^{1/2}$$
$$\le 2f_o^{1/4}C^{1/2} \text{ and}$$
$$\mathbb{E}[\sup_{\theta\in\Theta}||\eta^B(\theta)||] \le \big(\mathbb{E}\big[\sup_{\theta\in\Theta}\big|\big|\mathcal{Z}\frac{\partial\widetilde{X^g}(\theta)}{\partial\theta}\big|\big|^4\big]\big)^{1/4} \le C^{1/4}.$$

□

E.3 The Stochastic Equi-continuity Condition

The following lemma verifies Assumption 3(e).

Lemma E.4. *Under Assumption 1((b) and (c)), Assumption 2(a), Assumption 3((a), (c) and (f)),* $\sup_{\theta\in\Theta_o} \frac{||v_n(\theta)-v_n(\theta_o)||}{1+\sqrt{n}||\mathcal{S}(\theta)||} = o_p(1)$ *in some open neighbourhood* Θ_o *of* θ_o.

Proof. The proof of this result proceeds by verifying the conditions of Honoré and Powell (1994, Lemma 2). Recall $\psi(W_i, W_j; \theta) \equiv \mathcal{Z}\big(1 - 2\mathrm{I}(\widetilde{U}(\theta) < 0)\big)\widetilde{X^g}(\theta)'$.

First, from Assumption 1(b), $\widetilde{U}(\theta)$ and $\widetilde{X^g}(\theta)$, are, respectively, measurable in $[U, U^\dagger, X, X^\dagger]$ and $[X, X^\dagger]$ for all $\theta \in \Theta$. It follows that for any θ_1, θ_2 in an open neighbourhood $\Theta_o \subset \Theta$ containing θ_o (Assumption 3(a)), $\sup_{||\theta_1-\theta_2||<d} ||\psi(W, W^\dagger; \theta_1) - \psi(W, W^\dagger; \theta_2)||$ is a measurable function of $W, W^\dagger$ for all d sufficiently small. Assumption N1 of Honoré and Powell (1994) is thus verified.

Second, Assumption 1(b), Assumption 2(a), and Assumption 3(f) imply Assumption N2 of Honoré and Powell (1994).

Third, for the purpose of verifying Honoré and Powell (1994, Assumption N3), this paper follows Honoré and Powell (1994) is complementing the condition of d being sufficiently small with the upper bound $d \leq d_o = 1$. By the triangle inequality,

$$
\begin{aligned}
&||\psi(W, W^\dagger; \theta_1) - \psi(W, W^\dagger; \theta_2)|| \\
&= ||\mathcal{Z}(\widetilde{X^g}(\theta_1) - \widetilde{X^g}(\theta_2)) - 2(\mathrm{I}(\widetilde{U}(\theta_1) < 0)\widetilde{X^g}(\theta_1) - \mathrm{I}(\widetilde{U}(\theta_2) < 0)\widetilde{X^g}(\theta_2))|| \\
&\leq ||\mathcal{Z}(\widetilde{X^g}(\theta_1) - \widetilde{X^g}(\theta_2))|| + 2||\mathrm{I}(\widetilde{U}(\theta_1) < 0)\widetilde{X^g}(\theta_1) - \mathrm{I}(\widetilde{U}(\theta_2) < 0)\widetilde{X^g}(\theta_2)||.
\end{aligned}
$$

For the second summand, note that by triangle and Schwartz inequalities,

$$
\begin{aligned}
&||\mathrm{I}(\widetilde{U}(\theta_1) < 0)\widetilde{X^g}(\theta_1) - \mathrm{I}(\widetilde{U}(\theta_2) < 0)\widetilde{X^g}(\theta_2)|| \\
&= ||\mathrm{I}(\widetilde{U}(\theta_1) < 0)\widetilde{X^g}(\theta_1) - \mathrm{I}(\widetilde{U}(\theta_1) < 0)\widetilde{X^g}(\theta_2) + \mathrm{I}(\widetilde{U}(\theta_2) < 0)\widetilde{X^g}(\theta_1) - \mathrm{I}(\widetilde{U}(\theta_2) < 0)\widetilde{X^g}(\theta_2)|| \\
&\leq |\mathrm{I}(\widetilde{U}(\theta_1) < 0) - \mathrm{I}(\widetilde{U}(\theta_2) < 0)| \cdot ||\widetilde{X^g}(\theta_1)|| + \mathrm{I}(\widetilde{U}(\theta_2) < 0)||\widetilde{X^g}(\theta_1) - \widetilde{X^g}(\theta_2)|| \\
&\leq |\mathrm{I}(\widetilde{U}(\theta_1) < 0) - \mathrm{I}(\widetilde{U}(\theta_2) < 0)| \cdot ||\widetilde{X^g}(\theta_1)|| + ||\widetilde{X^g}(\theta_1) - \widetilde{X^g}(\theta_2)||.
\end{aligned}
$$

Re-arranging both summands,

$$
\begin{aligned}
&||\psi(W_i, W_j; \theta_1) - \psi(W_i, W_j; \theta_2)|| \\
&\leq 3||\max\{|\mathcal{Z}|, 1\}(\widetilde{X^g}(\theta_1) - \widetilde{X^g}(\theta_2))|| + 2|\mathrm{I}(\widetilde{U}(\theta_1) < 0) - \mathrm{I}(\widetilde{U}(\theta_2) < 0)| \cdot ||\widetilde{X^g}(\theta_1)||.
\end{aligned} \tag{E.2}
$$

Consider the first summand of (E.2). By the Assumption 1(b), the MVT, and the Schwartz inequality,

$$
\sup_{||\theta_1 - \theta_2|| < d} ||\widetilde{X^g}(\theta_1) - \widetilde{X^g}(\theta_2)|| \leq \sup_{||\theta_1 - \theta_2|| < d} ||\frac{\partial \widetilde{X^g}(\theta)}{\partial \theta}\Big|_{\theta = \bar{\theta}_{12}}|| \times ||\theta_1 - \theta_2|| < d \sup_{\theta \in \Theta} ||\frac{\partial \widetilde{X^g}(\theta)}{\partial \theta}||.
$$

Assumption 5 and the foregoing imply

$$
\mathbb{E}\Big[\sup_{||\theta_1 - \theta_2|| < d} \max\{|\mathcal{Z}|, 1\}||\widetilde{X^g}(\theta_1) - \widetilde{X^g}(\theta_2)||\Big] \leq C^{1/4} d.
$$

Consider the element $|\mathrm{I}(\widetilde{U}(\theta_1) < 0) - \mathrm{I}(\widetilde{U}(\theta_2) < 0)|$ in the second summand of (E.2).

$$
\begin{aligned}
&\sup_{||\theta_1 - \theta_2|| < d} |\mathrm{I}(\widetilde{U}(\theta_1) < 0) - \mathrm{I}(\widetilde{U}(\theta_2) < 0)| \cdot ||\widetilde{X^g}(\theta_1)|| \\
&\leq \sup_{||\theta_1 - \theta_2|| \leq d} |\mathrm{I}(\widetilde{U}(\theta_1) < 0) - \mathrm{I}(\widetilde{U}(\theta_2) < 0)| \cdot ||\widetilde{X^g}(\theta_1)|| \\
&\equiv |\mathrm{I}(\widetilde{U}(\theta_1^*) < 0) - \mathrm{I}(\widetilde{U}(\theta_2^*) < 0)| \cdot ||\widetilde{X^g}(\theta_1^*)||
\end{aligned}
$$

for some θ_1^*, θ_2^* that satisfy $||\theta_1^* - \theta_2^*|| \leq d \leq 1$. By Assumption 1(b) and the MVT, $\widetilde{U}(\theta_2^*) =$

$\widetilde{U}(\theta_1^*) + \widetilde{X^g}(\bar{\theta}_{12})(\theta_2^* - \theta_1^*)$ hence

$$
\begin{aligned}
|\mathrm{I}(\widetilde{U}(\theta_1^*) < 0) - \mathrm{I}(\widetilde{U}(\theta_2^*) < 0)| &= |\mathrm{I}(\widetilde{U}(\theta_1^*) < -\widetilde{X^g}(\bar{\theta}_{12})(\theta_2^* - \theta_1^*)) - \mathrm{I}(\widetilde{U}(\theta_1^*) < 0)| \\
&= \mathrm{I}(0 \leq \widetilde{U} < -\widetilde{X^g}(\bar{\theta}_{12})(\theta_2^* - \theta_1^*)) + \mathrm{I}(-\widetilde{X^g}(\bar{\theta}_{12})(\theta_2^* - \theta_1^*) \leq \widetilde{U} < 0).
\end{aligned}
$$

From the LIE, Assumption 1(b), the MVT, and the Schwartz inequality,

$$
\begin{aligned}
&\mathbb{E}[|\mathrm{I}(\widetilde{U}(\theta_1^*) < 0) - \mathrm{I}(\widetilde{U}(\theta_2^*) < 0)| \cdot ||\widetilde{X^g}(\theta_1^*)||] \\
&= \mathbb{E}[|F_{\widetilde{U}(\theta_1^*)|\widetilde{X^g},\mathcal{Z}}(-\widetilde{X^g}(\bar{\theta}_{12})(\theta_2^* - \theta_1^*)) - F_{\widetilde{U}(\theta_1^*)|\widetilde{X^g},\mathcal{Z}}(0)| \cdot ||\widetilde{X^g}(\theta_1^*)||] \\
&\leq \mathbb{E}[f_{\widetilde{U}(\theta_1^*)|\widetilde{X^g},\mathcal{Z}}(-\lambda\widetilde{X^g}(\bar{\theta}_{12})(\theta_2^* - \theta_1^*)) \cdot ||\widetilde{X^g}(\bar{\theta}_{12})|| \cdot ||\widetilde{X^g}(\theta_1^*)|| \cdot ||\theta_2^* - \theta_1^*||] \\
&\leq d\mathbb{E}[f_{\widetilde{U}(\theta_1^*)|\widetilde{X^g},\mathcal{Z}}(-\lambda\widetilde{X^g}(\bar{\theta}_{12})(\theta_2^* - \theta_1^*)) \cdot \sup_{\theta\in\Theta} ||\widetilde{X^g}(\theta)||^2]
\end{aligned}
$$

for some $\lambda \in (0,1)$. To complete this part, it remains to show that $\mathbb{E}[f_{\widetilde{U}(\theta_1^*)|\widetilde{X^g},\mathcal{Z}}(-\lambda\widetilde{X^g}(\bar{\theta}_{12})(\theta_2^* - \theta_1^*)) \cdot \sup_{\theta\in\Theta} ||\widetilde{X^g}(\theta)||^2] < \infty$. By Assumption 1(b) and the MVT,

$$
\begin{aligned}
f_{\widetilde{U}(\theta_1^*)|\widetilde{X^g},\mathcal{Z}}(\lambda) &= \frac{\partial\mathbb{E}[\mathrm{I}(\widetilde{U}(\theta_1^*) < \lambda)]}{\partial\lambda} = \frac{\partial\mathbb{E}[\mathrm{I}(\widetilde{U} + \widetilde{X^g}(\bar{\theta}_1^*)(\theta_1^* - \theta_o) < \lambda)]}{\partial\lambda} \\
&= \frac{\partial F_{\widetilde{U}|\tilde{\sigma}(X,Z)}(-\widetilde{X^g}(\bar{\theta}_1^*)(\theta_1^* - \theta_o) + \lambda)}{\partial\lambda} = f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(-\widetilde{X^g}(\bar{\theta}_1^*)(\theta_1^* - \theta_o) + \lambda)
\end{aligned}
$$

From the foregoing, Assumption 1(c), and Assumption 3(c),

$$
\begin{aligned}
&\mathbb{E}[f_{\widetilde{U}(\theta_1^*)|\widetilde{X^g},\mathcal{Z}}(-\lambda\widetilde{X^g}(\bar{\theta}_{12})(\theta_2^* - \theta_1^*)) \cdot \sup_{\theta\in\Theta} ||\widetilde{X^g}(\theta)||^2] \\
&= \mathbb{E}[f_{\widetilde{U}|\tilde{\sigma}(X,Z)}(-\widetilde{X^g}(\bar{\theta}_1^*)(\theta_1^* - \theta_o) - \lambda\widetilde{X^g}(\bar{\theta}_{12})(\theta_2^* - \theta_1^*)) \cdot \sup_{\theta\in\Theta} ||\widetilde{X^g}(\theta)||^2] \\
&\leq f_o^{1/4}\mathbb{E}[\sup_{\theta\in\Theta} ||\widetilde{X^g}(\theta)||^2] \leq f_o^{1/4}C^{1/2}.
\end{aligned}
$$

Thus,

$$
\mathbb{E}[\sup_{||\theta_1-\theta_2||<d} ||\psi(W_i, W_j; \theta_1) - \psi(W_i, W_j; \theta_2)||] \leq (3C^{1/4} + 2C^{1/2})d.
$$

By the c_r-inequality and (E.2),

$$
\begin{aligned}
&||\psi(W_i, W_j; \theta_1) - \psi(W_i, W_j; \theta_2)||^2 \\
&\leq 18||\max\{|\mathcal{Z}|, 1\}(\widetilde{X^g}(\theta_1) - \widetilde{X^g}(\theta_2))||^2 + 8|\mathrm{I}(\widetilde{U}(\theta_1) < 0) - \mathrm{I}(\widetilde{U}(\theta_2) < 0)| \cdot ||\widetilde{X^g}(\theta_1)||^2.
\end{aligned}
$$

Observe that $|\mathrm{I}(\widetilde{U}(\theta_1) < 0) - \mathrm{I}(\widetilde{U}(\theta_2) < 0)|^2 = |\mathrm{I}(\widetilde{U}(\theta_1) < 0) - \mathrm{I}(\widetilde{U}(\theta_2) < 0)|$. Using arguments analogous to the above and recalling that for all $d \leq d_o = 1$, $d^2 \leq d$,

$$
\mathbb{E}[\sup_{||\theta_1 - \theta_2|| < d} ||\psi(W_i, W_j; \theta_1) - \psi(W_i, W_j; \theta_2)||^2] \leq (18C^{1/2} + 8f_o^{1/2}C)d^2 < (18C^{1/2} + 8f_o^{1/2}C)d.
$$

Honoré and Powell (1994, Assumption N3) is thus verified.

Finally, $\mathbb{E}[||\psi(W, W^\dagger)||^2] \leq C^{1/2}$ from (C.5). This completes the proof. □

References

[1] Davidson, James. *Stochastic limit theory: An introduction for econometricians*. Oxford University Press, 1994.

[2] Honoré, Bo E and James L Powell. "Pairwise difference estimators of censored and truncated regression models". *Journal of Econometrics* 64.1-2 (1994), pp. 241–278.