Orthogonal Scores and Double/Debiased Machine Learning for Possibly Misspecified GMM

Byunghoon Kang*
Lancaster University

November 12, 2026

Abstract

We develop Neyman orthogonal scores for GMM models that remain valid when the moment conditions may fail to hold. Existing score constructions typically require the moment conditions to hold exactly, which may fail in over-identified models, where the number of moment functions exceeds the number of parameters. We construct *misspecification-robust* orthogonal score that has a closed form; it is the conventional orthogonal score plus adjustment terms, for the estimation of the Jacobian and of the weighting matrix, which are proportional to the population moment and vanish under correct specification. The moment functions may depend on nonparametric or machine learning first steps, and we show that double/debiased machine learning (DML) combined with the proposed scores and cross-fitting provides uniformly valid inference for a low-dimensional GMM estimand, while the conventional DML standard errors are inconsistent when the moment conditions fail. We illustrate the results in the partially linear instrumental variables (PLIV) model and in the local average treatment effects (LATE) model with multiple instruments.

Keywords: Neyman orthogonal moments, double/debiased machine learning, GMM, misspecification, semiparametric two-step estimation, cross-fitting.
JEL Classification: C13, C14, C26, C36.

1 Introduction

Many parameters of interest in economics are defined by moment conditions and estimated by the generalized method of moments (GMM, Hansen, 1982). The moment conditions often depend on nuisance parameters, which are unknown functions or high-dimensional vectors estimated in a first step by machine learning or nonparametric methods. Plugging first-step estimates into the moment function generally produces a first-order bias, because the regularization or smoothing bias of the first step affects the estimator of the parameter of interest. A large literature (Neyman, 1959; Newey, 1994; Chernozhukov et al., 2015b, 2018, 2022a) shows that this bias can be removed

*Email: b.kang1@lancaster.ac.uk.

by a Neyman orthogonal, or locally robust, moment function which is insensitive to the nuisance parameters. Double/debiased machine learning (DML) (Chernozhukov et al., 2015b, 2018) is now widely used in economics to estimate a low-dimensional structural parameter in such settings. Combined with cross-fitting that removes own-observation bias and avoids Donsker conditions on the first-step estimator, the resulting DML or debiased GMM estimators are $\sqrt{N}$-consistent and asymptotically normal under the rate conditions that many machine learning and nonparametric methods satisfy.

However, the orthogonal scores used in this literature are derived under correct specification: there exists a parameter value at which the population moments are exactly zero. In over-identified models, with more moment conditions than parameters, it is an assumption about the data generating process.[1] For instance, in the partially linear instrumental variables (PLIV) model, the moment conditions may fail when some of the instruments are invalid, or when the effect of the endogenous regressor is heterogeneous so that different instruments identify different local average treatment effects (LATE) (Imbens and Angrist, 1994).

When the moment conditions fail, the GMM estimator still converges to a well-defined limit, the GMM estimand or pseudo-true value, which solves the population first-order condition and depends on the weighting matrix. This estimand is often of interest in its own right; with heterogeneous treatment effects, for example, it is a weighted average of instrument-specific LATEs (Angrist and Imbens, 1995). However, the orthogonal score of Chernozhukov et al. (2015b) and Chernozhukov et al. (2018) for GMM, which combines the moment functions using the Jacobian and the weighting matrix, is no longer orthogonal when the population moments are not zero, and then the conventional DML variance estimator is inconsistent.

This paper constructs Neyman orthogonal/locally robust scores for over-identified GMM that remain valid when the moment conditions do not hold, which we label *misspecification-robust* orthogonal scores. We stack the GMM first-order condition with the moment equations that define the Jacobian and the weighting matrix, and write the over-identified GMM estimator as a just-identified estimator with an augmented parameter vector, as in Imbens (1997). The stacked moment conditions hold at the GMM estimand by construction, whether or not the original moment conditions hold. However, this stacked moment condition is *not* orthogonal to the Jacobian and the weighting matrix that are treated as additional finite-dimensional nuisance parameters. Then, applying the standard orthogonal score construction to the stacked system gives a score that is orthogonal with or without correct specification.

We show that the resulting score has a closed form. It equals the conventional orthogonal score plus two adjustment terms, one for the Jacobian and one for the weighting matrix, both proportional to the population moment at the GMM estimand. Under correct specification, the two terms vanish and the score reduces to the conventional one. With finite-dimensional nuisance parameters, we verify that these additional terms are those that Hall and Inoue (2003) and Hansen

[1] In just-identified models, where the number of moment conditions equals the number of parameters, the estimand can be defined as the unique solution to the population moment conditions. Chernozhukov et al. (2022a) note that their standard errors are robust to misspecification in the just-identified case.

and Lee (2021) obtain by direct expansion for the asymptotic variance of over-identified GMM under misspecification (see also Imbens (1997) and Hwang et al. (2022)). Although the construction does not give a new variance formula for GMM with finite-dimensional parameters, our result makes explicit the connection between the Neyman orthogonal score and the robust variance formula in misspecified GMM. Furthermore, the corrections appear here as terms of a Neyman orthogonal score, which, unlike a variance formula, can be made insensitive to a first-step estimate.

We extend our results to semiparametric two-step GMM, where the parameter of interest is finite-dimensional and the moment functions depend on unknown functions estimated in a first step, as in Newey (1994), Ackerberg et al. (2014) and Chernozhukov et al. (2022a). Starting from an orthogonal/locally robust moment function (e.g., obtained by adding the first-step influence function to the identifying moment function as in Chernozhukov et al. (2022a)), we derive a misspecification-robust score that is orthogonal jointly to the first-step functions, the Jacobian and the weighting matrix.

We then study the DML estimator with cross-fitting based on the misspecification-robust score. For scores that are linear in the parameter of interest with a one-step weighting matrix, we show that the estimator is asymptotically normal and the confidence intervals are uniformly valid for the GMM estimand over a class of data generating processes. The same confidence interval is valid whether or not the moment conditions hold. With linear moment functions and one-step weighting, the point estimate is unchanged and it coincides with the DML estimator of Chernozhukov et al. (2018) and the debiased GMM estimator of Chernozhukov et al. (2022a). So, a researcher who already reports DML estimates keeps the same estimate and replaces only the standard error with the one based on the misspecification-robust score. The rate conditions on the first step are similar to those of Chernozhukov et al. (2018), with the additional requirement that the first-step biases of the Jacobian and of the weighting matrix are small when the model is misspecified.

For clarity, our main asymptotic results focus on moment functions that are linear in the parameter of interest and a one-step weighting matrix that does not depend on it. Our general results on the Neyman orthogonal score constructions allow both one-step GMM (Theorem 1) and efficient GMM weighting (Theorem 5(i)). Efficient weighting requires an additional correction term, because the weighting matrix is then not orthogonal to the first step, and it makes the score non-linear in the parameter of interest even when the moment functions are linear. Non-linear moment functions and general weighting, including iterated GMM, are treated in Appendix B (Theorem 5; see also Remark 3).

We consider two illustrative examples. First, the over-identified PLIV model is a leading example. With an orthogonal moment function in the style of Robinson (1988), the nuisance parameters are the conditional expectations of the outcome, the endogenous regressor and the instruments given the controls. Then, the DML estimator is cross-fitted two-stage least squares on the residualized data, and the robust score adds correction terms that involve the same conditional expectations and sample moments of the residuals. Second, we consider the LATE model with multiple binary instruments, where over-identification arises from treatment effect heterogeneity rather than from

invalid instruments. With a diagonal weighting matrix the GMM estimand is a convex combination of the instrument-specific LATEs, and we provide valid inference for it. In a Monte Carlo study of the PLIV model, the robust standard error under misspecification is about 15–20% larger than the conventional one and close to the standard deviation of the estimator, and the conventional DML confidence interval undercovers at every sample size while the robust one stays close to the nominal level. Under correct specification, the two standard errors and confidence intervals are almost the same, so little is lost by using the robust score even when the moment conditions hold.

This paper is related to two strands of the literature. The first is the large literature on Neyman orthogonal or locally robust moment functions and semiparametric two-step estimation (Neyman, 1959; Newey, 1994; Chernozhukov et al., 2015b, 2018; Ichimura and Newey, 2022; Chernozhukov et al., 2022a,b). Ackerberg et al. (2014) study the efficiency of semiparametric two-step GMM. These papers assume that the moment conditions hold, or allow misspecification only in the just-identified case. We extend Chernozhukov et al. (2015b) and Chernozhukov et al. (2018) by providing the Neyman orthogonal score for over-identified GMM that is robust to misspecification of the moment conditions. Ai and Chen (2007) study sieve minimum distance estimation of possibly misspecified semiparametric conditional moment restriction models with a known weighting matrix; in their setting, the finite-dimensional parameter and the unknown functions are estimated jointly (Ai and Chen, 2003, 2012). Our paper considers the semiparametric two-step GMM based on the unconditional moment conditions, and the first step is a separate object, such as a conditional expectation, whose limit is well defined whether or not the moment conditions hold. This two-step structure is what allows the first step to be estimated by any machine learning method with cross-fitting. Chen and Santos (2018) study over-identification in a general class of regular models.

The second is the growing literature on model misspecification in general contexts (Kitamura et al., 2013; Armstrong and Kolesár, 2021; Bonhomme and Weidner, 2022; Andrews et al., 2025a,b), which goes back to White (1982) for maximum likelihood. For GMM with finite-dimensional parameters, Hall and Inoue (2003) derive the limit distribution of the estimator under misspecification, Hansen and Lee (2021) consider the iterated GMM estimator, and Lee (2018) considers two-stage least squares when instruments identify different LATEs; see also Imbens (1997) and Schennach (2007). We extend the misspecification-robust variance of this literature to models with nonparametric or machine learning first steps. Under correct specification, the adjustment terms in our score are of the order of the sample moment, and Hwang et al. (2022) show that the misspecification-robust variance, which takes these terms into account, acts as a finite-sample or higher-order correction in linear GMM. We consider first-order orthogonality; recent papers by Bonhomme et al. (2026) and Chetverikov et al. (2026) construct moment functions that are Neyman orthogonal to higher order in order to reduce higher-order bias.

The rest of the paper is organized as follows. Section 2 reviews the Neyman orthogonal score for GMM with finite-dimensional nuisance parameters, explains why it fails under misspecification, and constructs the misspecification-robust score. Section 3 extends the construction to functional nuisance parameters, defines the DML estimator with cross-fitting and establishes its asymptotic

properties. Section 4 applies the results to the over-identified PLIV model, and Section 5 to local average treatment effects identified by multiple instruments. Section 6 reports a Monte Carlo study, and Section 7 concludes. Appendix A collects the proofs, and Appendix B extends the results to non-linear scores and general weighting, including iterated GMM.

We use the following generic notation throughout the paper. $\Pr_P$ and $\mathrm{E}_P[\cdot]$ denote probability and expectation under the probability law P, with the subscript suppressed when no ambiguity arises. We use capital letters W to denote random elements and the corresponding lowercase w to denote fixed values W can take, and $(W_i)_{i=1}^N$ denotes a random sample of N independent copies of W. We equip the parameter spaces with their Borel σ-fields, and assume throughout that the moment functions and each component of the score functions are measurable. For $q \geq 1$ we write $\|f\|_{P,q} := \|f(W)\|_{P,q} := \big(\int |f(w)|^q \, dP(w)\big)^{1/q}$ for the $L^q(P)$ norm. For a column vector x, x' is its transpose. For a differentiable map $x \mapsto f(x)$ from $\mathbb{R}^d$ to $\mathbb{R}^k$, we denote $\partial_{x'} f$ as the matrix of partial derivatives $(\partial/\partial x')f$, and write $\partial_{x'} f(x_0)$ for $\partial_{x'} f(x)|_{x=x_0}$. $\mathrm{vec}(\cdot)$ denotes column-wise vectorization, and $\otimes$ denotes the Kronecker product.

2 Neyman orthogonal score/moment functions for GMM

In this section, we formally introduce the model and provide the definitions of Neyman orthogonality for a self-contained treatment. We closely follow the notation of Chernozhukov et al. (2018). Let $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ denote a finite-dimensional parameter vector of interest. We assume that the true value θ_0 satisfies the moment conditions

$$\mathrm{E}_P\big[\psi(W, \theta_0, \eta_0)\big] = 0, \tag{1}$$

where $\psi = (\psi_1, \ldots, \psi_{d_\theta})'$ is a known score function, W is a random element in a measurable space $(\mathcal{W}, \mathcal{A}_\mathcal{W})$ with law determined by a probability measure $P \in \mathcal{P}_N$, where $\mathcal{P}_N$ is a collection of probability measures, and a nuisance parameter η taking values in a convex subset T of a normed vector space, with true value $\eta_0 \in T$. To introduce the Neyman orthogonality condition, we define the pathwise (Gateaux) derivative map for $\eta \in T$,

$$D_r[\eta - \eta_0] := \partial_r \Big\{ \mathrm{E}_P\big[\psi\big(W, \theta_0, \eta_0 + r(\eta - \eta_0)\big)\big] \Big\}, \qquad r \in [0, 1),$$

whenever it exists, and write $\partial_\eta \mathrm{E}_P[\psi(W, \theta_0, \eta_0)][\eta - \eta_0] := D_0[\eta - \eta_0]$. Let $T_N \subset T$ be a nuisance realization set such that the estimators $\widehat{\eta}_0$ of η_0 specified below take values in this set with high probability.

Definition 1 (Neyman orthogonality; Definition 2.1 in Chernozhukov et al. (2018))**.** *The score ψ obeys the* orthogonality condition *at (θ_0, η_0) with respect to the nuisance realization set $T_N \subset T$ if* (1) *holds and the pathwise derivative map $D_r[\eta - \eta_0]$ exists for all $r \in [0, 1)$ and $\eta \in T_N$ and vanishes at $r = 0$:*

$$\partial_\eta \mathrm{E}_P\big[\psi(W, \theta_0, \eta_0)\big][\eta - \eta_0] = 0 \qquad \textit{for all } \eta \in T_N. \tag{2}$$

When η is a finite-dimensional vector, condition (2) holds with $T_N = T$ if the vector of partial derivatives $\partial_\eta \mathrm{E}_P[\psi_j(W, \theta_0, \eta)]|_{\eta=\eta_0} = 0$ for each $j = 1, \ldots, d_\theta$.

2.1 Construction of Neyman orthogonal scores under correct specification

Now, we provide the construction of the Neyman orthogonal scores for the GMM model. We first consider the GMM model with finite-dimensional nuisance parameters by closely following the same model in Chernozhukov et al. (2015b) and Chernozhukov et al. (2018) to describe why they do not apply directly to the misspecified GMM models, and then we consider the semiparametric GMM model with infinite-dimensional or functional nuisance parameters in Section 3.

Chernozhukov et al. (2015b) and Chernozhukov et al. (2018, Section 2.2.2) consider the following correctly specified GMM model, i.e., assume that the parameters θ and β have true values θ_0 and β_0 solving the moment conditions exactly

$$\mathrm{E}_P\big[m(W, \theta_0, \beta_0)\big] = 0, \tag{3}$$

where $m : \mathcal{W} \times \Theta \times \mathcal{B} \to \mathbb{R}^{d_m}$ is a known, vector-valued moment function, $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ is the parameter of interest and $\beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta}$ is a finite-dimensional nuisance parameter with $d_m \geq d_\theta + d_\beta$. We assume throughout that m is twice continuously differentiable in (θ, β) with dominated derivatives, so that differentiation and expectation can be interchanged.

Then, a Neyman orthogonal score function for GMM is

$$\psi(W, \theta, \eta) = \mu\, m(W, \theta, \beta), \qquad \eta = \big(\beta', \mathrm{vec}(\mu)'\big)' \in T = \mathcal{B} \times \mathbb{R}^{d_\theta d_m}, \tag{4}$$

where $\mu \in \mathbb{R}^{d_\theta \times d_m}$ is the orthogonalization matrix considered as an additional nuisance parameter, and whose true value is

$$\mu_0 := A'\Omega^{-1} - A'\Omega^{-1} G_{\beta,0}\big(G_{\beta,0}'\Omega^{-1}G_{\beta,0}\big)^{-1} G_{\beta,0}'\Omega^{-1} \ \in \mathbb{R}^{d_\theta \times d_m}, \tag{5}$$

where A is a $d_m \times d_\theta$ moment selection matrix, Ω is a $d_m \times d_m$ positive definite weighting matrix, and $G(\gamma)$ is the Jacobian matrix for $\gamma = (\theta', \beta')' \in \Theta \times \mathcal{B} \subset \mathbb{R}^{d_\gamma}, d_\gamma = d_\theta + d_\beta$,

$$G(\gamma) := \partial_{\gamma'}\, \mathrm{E}_P[m(W, \gamma)] = \big[\, G_\theta(\gamma) \ \ G_\beta(\gamma) \,\big] \in \mathbb{R}^{d_m \times d_\gamma},$$

and $G_0 = G(\gamma_0) = [\, G_{\theta,0} \ \ G_{\beta,0} \,]$ for $\gamma_0 = (\theta_0', \beta_0')'$. Both A and Ω may be chosen arbitrarily and

$$A = G_{\theta,0}, \qquad \Omega = \mathrm{Var}_P\big(m(W, \theta_0, \beta_0)\big) = \mathrm{E}_P\big[m(W, \theta_0, \beta_0) m(W, \theta_0, \beta_0)'\big] \tag{6}$$

yields the efficient score, in the sense of providing the estimator of θ_0 with the smallest variance in the class of GMM estimators (Hansen, 1982), and the semiparametrically efficient score (Chamberlain, 1987).

The following Lemma shows that the score in (4) satisfies the Neyman orthogonality condition.

Let $\eta_0 = (\beta_0', \mathrm{vec}(\mu_0)')'$ denote the true value of η.

Lemma 1 (Neyman orthogonal scores for GMM setup, Lemma 2.3 of Chernozhukov et al. (2018))**.** *If* (3) *holds,* G_0 *exists and* Ω*,* $G_{\beta,0}'\Omega^{-1}G_{\beta,0}$ *are invertible, then the score* ψ *in* (4) *is Neyman orthogonal at* (θ_0, η_0) *with respect to the nuisance realization set* $T_N = T$*.*

It is straightforward to show that the Neyman orthogonality condition in Lemma 1 requires the following:

$$\partial_\eta \mathrm{E}_P\big[\psi(W, \theta_0, \eta_0)\big] = \Big[\ \mu_0 G_{\beta,0}\ ,\ \ \mathrm{E}_P\big[m(W, \theta_0, \beta_0)\big]' \otimes I_{d_\theta}\ \Big] = 0. \tag{7}$$

Orthogonality with respect to β in the first block holds because $\mu_0 G_{\beta,0} = 0$ by construction in (5). However, the second block (partial derivatives with respect to μ) vanishes only when the moment condition $\mathrm{E}_P[m(W, \theta_0, \beta_0)] = 0$ holds at (θ_0, β_0). When the moment condition fails to hold, i.e., $\mathrm{E}_P[m(W, \theta, \beta)] \neq 0$ for all values of (θ, β), the score ψ does not satisfy the Neyman orthogonality condition as $\mathrm{E}_P[m(W, \theta_0, \beta_0)]' \otimes I_{d_\theta} \neq 0$ for any choice of A and Ω.

2.2 Neyman orthogonal score for possibly misspecified GMM

Now, we construct Neyman orthogonal scores/moment functions which are valid even when the moment condition (3) fails to hold, i.e., $\mathrm{E}_P[m(W, \theta, \beta)] \neq 0$ for all $\gamma \in \Theta \times \mathcal{B}$. Define the population moment as

$$\bar{m}(\gamma) := \mathrm{E}_P\big[m(W, \gamma)\big] \in \mathbb{R}^{d_m}, \qquad \gamma = (\theta', \beta')'.$$

The model is *correctly specified* if $\bar{m}(\gamma) = 0$ for some γ in the interior of $\Theta \times \mathcal{B}$. When the model is just-identified, $d_m = d_\gamma$, we define γ_0 as a solution to $\bar{m}(\gamma_0) = 0$. When the model is over-identified, $d_m > d_\gamma$, generically no $\gamma \in \Theta \times \mathcal{B}$ solves all d_m equations. When $d_m > d_\gamma$, what a GMM estimator estimates is a GMM estimand (or pseudo-true value), $\gamma_0 = (\theta_0', \beta_0')'$, defined as the solution of the population first-order condition

$$Q(\gamma_0) = 0, \qquad Q(\gamma) := G(\gamma)'\,\Omega(\gamma)^{-1}\,\bar{m}(\gamma) \in \mathbb{R}^{d_\gamma}, \tag{8}$$

where $\Omega(\gamma)$ is the symmetric, positive definite weight. We assume that a unique γ_0 exists and lies in the interior of $\Theta \times \mathcal{B}$. We define the weighting matrix $\Omega(\gamma) := \mathrm{E}_P[\zeta(W, \gamma)]$, where $\zeta(w, \gamma) = \zeta(w)$ for the one-step GMM, and $\zeta(w, \gamma) = m(w, \gamma)\, m(w, \gamma)'$ for the iterated GMM. For the one-step GMM, the weight $\Omega(\gamma) \equiv \Omega$ does not vary with γ, γ_0 is a minimizer of the population criterion $\bar{m}(\gamma)'\Omega^{-1}\bar{m}(\gamma)$, and (8) is the first-order condition. For the iterated GMM, (8) is instead a fixed point, and γ_0 is a minimizer of $\bar{m}(\gamma)'\Omega(\gamma_0)^{-1}\bar{m}(\gamma)$. Under correct specification, (8) holds for any weight, so γ_0 coincides with the true value $(\theta_0', \beta_0')'$ of Section 2.1. Under misspecification, γ_0 depends on the weighting matrix, and we suppress the dependence in the notation for simplicity.

To construct the Neyman orthogonal scores that are valid even when the moment condition fails to hold, we stack the GMM first-order condition together with moment equations that define the Jacobian and the weight matrix as additional nuisance parameters. To be specific, define the

stacked moment function as follows:

$$h(w,\theta,\lambda) = \begin{pmatrix} \mathrm{vec}\big(\Gamma - \partial_{\gamma'} m(w,\gamma)\big) \\ \mathrm{vec}\big(\Omega - \zeta(w,\gamma)\big) \\ \Gamma' \, \Omega^{-1} \, m(w,\gamma) \end{pmatrix} \in \mathbb{R}^{d_h}, \quad \lambda := \big(\beta',\ \mathrm{vec}(\Gamma)',\ \mathrm{vec}(\Omega)'\big)' \in \Lambda \subset \mathbb{R}^{d_\lambda} \tag{9}$$

where λ is the nuisance parameter with $d_\lambda = d_\beta + d_m d_\gamma + d_m^2$, $\Gamma \in \mathbb{R}^{d_m \times d_\gamma}$ is a Jacobian parameter, and $\Omega \in \mathbb{R}^{d_m \times d_m}$ is a weight parameter (symmetric positive definite). It is straightforward to check that the number of equations for the stacked moment function $h(w,\theta,\lambda)$ equals the number of parameters (θ,λ), so the system (9) is just-identified. Imbens (1997) considers the just-identified GMM representation with an augmented parameter vector for the standard over-identified GMM model (see also Newey and McFadden, 1994; Chamberlain and Imbens, 1995).[2] In one-step GMM, when $\zeta(W,\gamma) = \zeta(W)$ does not depend on γ, (9) has the same three components, with the outer product replaced by ζ, and for the iterated GMM, (9) is exactly the moment function in equation (4) of Imbens (1997).[3]

We can define the true value of the nuisance λ in (9), with γ_0 defined in (8), as

$$\lambda_0 := \big(\beta_0',\ \mathrm{vec}(G_0)',\ \mathrm{vec}(\Omega_0)'\big)', \qquad G_0 = \partial_{\gamma'} \mathrm{E}_P\big[m(W,\gamma)\big]\big|_{\gamma=\gamma_0}, \qquad \Omega_0 = \mathrm{E}_P\big[\zeta(W,\gamma_0)\big].$$

Then, the following moment condition holds by construction,

$$\mathrm{E}_P\big[h(W,\theta_0,\lambda_0)\big] = 0 \tag{10}$$

without assuming the original moment conditions $\mathrm{E}_P\big[m(W,\theta_0,\beta_0)\big] = 0$.

The following result extends the results of Chernozhukov et al. (2015b) and Chernozhukov et al. (2018) by providing the Neyman orthogonal construction of the score $\psi^*(w,\theta,\eta)$ for GMM that is robust to misspecification of the moment conditions (3).

Proposition 1 (Misspecification-robust Neyman orthogonal scores for GMM)**.** *Suppose that the moment condition* (10) *holds and $h(w,\theta,\lambda)$ is differentiable in (θ,λ) at (θ_0,λ_0). We consider the following score*

$$\psi^*(w,\theta,\eta) := \mu \, h(w,\theta,\lambda), \qquad \eta = \big(\lambda', \mathrm{vec}(\mu)'\big)' \in T = \Lambda \times \mathbb{R}^{d_\theta d_h}, \tag{11}$$

[2] Imbens (1997) considers vech(Ω) instead of vec(Ω). Because Ω is symmetric, the number of parameters is the same as the one considered in Imbens (1997) after eliminating duplicated $d_m(d_m-1)/2$ elements. To keep the matrix algebra transparent, we use vec(), and all discussions remain the same by replacing vec() with vech().

[3] The two-step GMM estimator also admits a just-identified representation of the form (9), with augmented parameter vectors of the first-step estimator, its own Jacobian and the first-order condition with the first-step weighting matrix. Thus, the orthogonal score for the two-step GMM requires an additional term relative to the one-step case. The main text focuses on the one-step and the iterated GMM estimators, and we do not treat the continuously updated estimator (CUE) of Hansen et al. (1996), in which the first-order condition contains an additional term from the derivative of the weighting matrix, which is different from (8). Under correct specification, the two-step, iterated and CUE estimators are first-order equivalent and asymptotically efficient. Under misspecification, they estimate possibly different estimands, and the textbook variance formula $(G_0'\Omega_0^{-1}G_0)^{-1}$ is invalid.

where Λ is convex, $\mu \in \mathbb{R}^{d_\theta \times d_h}$ is the orthogonalization matrix with $d_h = d_\theta + d_\lambda$, and whose true value is

$$\mu_0 := A_h'\Omega_h^{-1} - A_h'\Omega_h^{-1}J_\lambda\big(J_\lambda'\Omega_h^{-1}J_\lambda\big)^{-1}J_\lambda'\Omega_h^{-1} \in \mathbb{R}^{d_\theta \times d_h}, \qquad \eta_0 := \big(\lambda_0', \mathrm{vec}(\mu_0)'\big)', \tag{12}$$

where $A_h \in \mathbb{R}^{d_h \times d_\theta}$ is a moment selection matrix, Ω_h is a $d_h \times d_h$ positive definite matrix, and J is the Jacobian of the stacked system at the pseudo-true value,

$$J := \mathrm{E}_P\big[\partial_{(\theta',\lambda')}h(W,\theta_0,\lambda_0)\big] = \big[\, J_\theta \;\; J_\lambda \,\big] \in \mathbb{R}^{d_h \times d_h}, \qquad J_\theta \in \mathbb{R}^{d_h \times d_\theta}, \quad J_\lambda \in \mathbb{R}^{d_h \times d_\lambda},$$

and J_λ has full column rank d_λ.

Then $\mathrm{E}_P[\psi^*(W,\theta_0,\eta_0)] = 0$, and the score $\psi^*(w,\theta,\eta)$ of (11) is Neyman orthogonal at (θ_0,η_0) with respect to the unrestricted nuisance realization set $T_N = T$.

The Neyman orthogonality condition in Proposition 1 is robust against moment misspecification compared to the results in Chernozhukov et al. (2018, Lemma 2.3), in the sense that $\psi^*(w,\theta,\eta)$ is orthogonal at (θ_0,η_0) whether or not the original moment condition $\bar{m}(\gamma) = \bar{m}(\theta,\beta) = 0$ holds. Analogous to Lemma 1, the proof of Proposition 1 requires the following to hold for the stacked system

$$\partial_\eta \mathrm{E}_P\big[\psi^*(W,\theta_0,\eta_0)\big] = \Big[\; \mu_0 J_\lambda \;,\;\; \mathrm{E}_P\big[h(W,\theta_0,\lambda_0)\big]' \otimes I_{d_\theta} \;\Big] = 0, \tag{13}$$

where the first block holds by definition of μ_0 in (12), and the second condition holds because of (10), which holds by construction. For the original scores ψ in Lemma 1, it is equal to $\bar{m}(\gamma_0)' \otimes I_{d_\theta}$, which is possibly non-zero under misspecification for every choice of A and Ω.[4]

Remark 1. *It is important to note that the stacked moment function $h(w,\theta,\lambda)$ changes the representation, not the estimator. Write the sample moment and sample weight as*

$$\bar{m}_N(\gamma) := \frac{1}{N}\sum_{i=1}^N m(W_i,\gamma), \qquad \widehat{\Omega}_N(\gamma) := \frac{1}{N}\sum_{i=1}^N \zeta(W_i,\gamma).$$

Solving the sample analogue of (9), $N^{-1}\sum_i h(W_i,\widehat{\theta},\widehat{\lambda}) = 0$ provides the sample Jacobian $\widehat{\Gamma} = N^{-1}\sum_i \partial_{\gamma'} m(W_i,\widehat{\gamma})$, the sample weight $\widehat{\Omega} = \widehat{\Omega}_N(\widehat{\gamma})$, and the original GMM first-order condition $\widehat{\Gamma}'\widehat{\Omega}^{-1}\bar{m}_N(\widehat{\gamma}) = 0$. When $\zeta(w,\gamma) = \zeta(w)$ does not vary with γ, the M-estimator with the moment function $h(w,\theta,\lambda)$ is the same as the one-step GMM estimator

$$\widehat{\gamma}^{\mathrm{os}} := \operatorname*{arg\,min}_{\gamma \in \Theta \times \mathcal{B}} \; \bar{m}_N(\gamma)'\,\widehat{\Omega}_N^{-1}\,\bar{m}_N(\gamma),$$

and when $\zeta(w,\gamma) = m(w,\gamma)\,m(w,\gamma)'$, the estimator is the iterated GMM estimator, defined as the

[4] Similar to Chernozhukov et al. (2018, Definition 2.2), we can also use an approximate Neyman orthogonality condition. For example, when $J_\lambda'\Omega_h^{-1}J_\lambda$ is close to singular, the exact construction (12) can be replaced by those of Chernozhukov et al. (2018, Lemma 2.4). The resulting score is then Neyman λ_N near-orthogonal, i.e., (2) holds only up to a bound $\|\partial_\eta \mathrm{E}_P[\psi^*(W,\theta_0,\eta_0)][\eta-\eta_0]\| \le \lambda_N$ for all $\eta \in T_N$, for a sequence of positive constants $\lambda_N = o(N^{-1/2})$.

fixed point

$$\widehat{\gamma}^{\mathrm{it}} = \underset{\gamma\in\Theta\times\mathcal{B}}{\arg\min}\ \bar{m}_N(\gamma)'\,\widehat{\Omega}_N(\widehat{\gamma}^{\mathrm{it}})^{-1}\,\bar{m}_N(\gamma).$$

The same equivalence holds for the debiased GMM estimator (Chernozhukov et al., 2022a, Section 2.2), where β_0 is estimated in a first step and enters the moment functions. We will discuss the construction of the misspecification-robust score with functional nuisances β_0 in Section 3.

2.3 The robust orthogonal score in a closed form

Proposition 1 defines the robust orthogonal score function $\psi^*(w,\theta,\eta)$ implicitly, but it can be made explicit. The next Proposition shows that the robust score $\psi^*(w,\theta,\eta)$ is related to the terms used to construct the so-called misspecification-robust variance of Hall and Inoue (2003) and Hansen and Lee (2021) (see also Andrews et al. (2025b)), exhibiting the extra terms, relative to the textbook formula, that arise from the sampling variation of the estimated Jacobian and weight matrix.

The orthogonalization matrix μ_0 in (12) depends on the choice of (A_h,Ω_h), and so does the score ψ^* below. However, the influence function of the GMM estimator $\widehat{\theta}$ in Proposition 2 does not depend on (A_h,Ω_h).

Proposition 2 (The robust orthogonal score in closed form)**.** *Suppose the conditions of Proposition 1 hold, the Jacobian of the score $\mu_0 J_\theta = \partial_{\theta'}\mathrm{E}_P\big[\psi^*(W,\theta,\eta_0)\big]\big|_{\theta=\theta_0}$, H and $H_{\beta\beta}$ are nonsingular, where*

$$H := \partial_{\gamma'}Q(\gamma)\big|_{\gamma=\gamma_0} = \begin{pmatrix} H_{\theta\theta} & H_{\theta\beta} \\ H_{\beta\theta} & H_{\beta\beta}\end{pmatrix} \in \mathbb{R}^{d_\gamma\times d_\gamma}, H_{\theta\theta}\in\mathbb{R}^{d_\theta\times d_\theta}, H_{\beta\beta}\in\mathbb{R}^{d_\beta\times d_\beta},$$

and

$$J_0 := H_{\theta\theta} - H_{\theta\beta}H_{\beta\beta}^{-1}H_{\beta\theta}$$

is also nonsingular. Then the score (11) *at (θ_0,η_0) satisfies*

$$\psi^*(w,\theta_0,\eta_0) = \mu_0 J_\theta\, J_0^{-1}\big(s_\theta(w) - H_{\theta\beta}H_{\beta\beta}^{-1}s_\beta(w)\big), \tag{14}$$

where $s(w) = (s_\theta(w)', s_\beta(w)')'$ with $s_\theta(w)\in\mathbb{R}^{d_\theta}, s_\beta(w)\in\mathbb{R}^{d_\beta}$,

$$s(w) := G_0'\Omega_0^{-1}m(w,\gamma_0) + \big(\partial_{\gamma'}m(w,\gamma_0) - G_0\big)'\Omega_0^{-1}\bar{m}(\gamma_0) - G_0'\Omega_0^{-1}\big(\zeta(w,\gamma_0)-\Omega_0\big)\Omega_0^{-1}\bar{m}(\gamma_0). \tag{15}$$

Moreover, the GMM estimator $\widehat{\theta}$ of Remark 1 satisfies, under standard regularity conditions,

$$\begin{aligned}\sqrt{N}\big(\widehat{\theta}-\theta_0\big) &= -(\mu_0 J_\theta)^{-1}\,\frac{1}{\sqrt{N}}\sum_{i=1}^N \psi^*(W_i,\theta_0,\eta_0) + o_P(1) \\ &= -J_0^{-1}\,\frac{1}{\sqrt{N}}\sum_{i=1}^N\big(s_\theta(W_i) - H_{\theta\beta}H_{\beta\beta}^{-1}s_\beta(W_i)\big) + o_P(1) \ \xrightarrow{d}\ \mathcal{N}\big(0,\ J_0^{-1}\Sigma_0 J_0^{-1\prime}\big),\end{aligned} \tag{16}$$

where $\Sigma_0 = \mathrm{Var}_P\left(s_\theta(W) - H_{\theta\beta}H_{\beta\beta}^{-1}s_\beta(W)\right)$.

Misspecification is what activates the orthogonalization. When $\bar{m}(\gamma_0) = 0$, Proposition 2 reduces to Lemma 1, i.e., the score in (14) becomes (up to the factor $\mu_0 J_\theta J_0^{-1}$)

$$s_\theta(w) - H_{\theta\beta}H_{\beta\beta}^{-1}s_\beta(w) = \Big(G_{\theta,0}'\Omega_0^{-1} - G_{\theta,0}'\Omega_0^{-1}G_{\beta,0}\big(G_{\beta,0}'\Omega_0^{-1}G_{\beta,0}\big)^{-1}G_{\beta,0}'\Omega_0^{-1}\Big)m(w,\gamma_0),$$

which is exactly the score (4) of Lemma 1, with μ_0 from (5) evaluated at $A = G_{\theta,0}$ and $\Omega = \Omega_0$, because the Jacobian and weight matrix channels of (15) vanish, $s(w) = G_0'\Omega_0^{-1}m(w,\gamma_0)$, and H reduces to $G_0'\Omega_0^{-1}G_0$. When $\Omega_0 = \mathrm{E}_P[m(W,\gamma_0)m(W,\gamma_0)']$, this is the efficient choice (6).

Note that (Γ, Ω) are not nuisance parameters of the original GMM problem, but when $\bar{m}(\gamma_0) \neq 0$, the first-order condition in the stacked moment function $h(w,\theta,\lambda)$ is *not* orthogonal to (Γ,Ω), which is precisely why Jacobian and weight estimation then contribute to the orthogonal score, and thus the influence function and asymptotic variance of the GMM estimator under misspecification. The orthogonalization makes these contributions explicit: they are the Jacobian and weight matrix channels of (15), which are both proportional to $\bar{m}(\gamma_0)$.

Therefore, with the finite-dimensional nuisance parameters, it is not surprising that the influence function formula in Proposition 2 coincides with those of misspecified GMM literature. Hall and Inoue (2003) and Hansen and Lee (2021)) obtain the formula by direct expansion of the first-order condition with finite-dimensional nuisance parameters; the three terms in (15) correspond to those in Hall and Inoue (2003), and (16) reproduces the misspecification-robust asymptotic variance of Hall and Inoue (2003) for the one-step GMM, and that of Hansen and Lee (2021) for the iterated GMM. In our construction, our corrections appear as terms of a Neyman orthogonal score rather than as corrections to a variance formula. This distinction matters when a score (11) depends on the infinite-dimensional nuisance parameters β_0, because our robust-score construction can be made insensitive to a first-step estimate of β_0. Section 3 exploits this direction by considering semiparametric GMM setup, replacing the finite-dimensional β_0 by infinite-dimensional nuisance parameters β_0 estimated nonparametrically or by a machine learning first step.

3 Neyman orthogonal scores for GMM with infinite-dimensional nuisance parameters

Section 2 considers the moment condition model, where the nuisance parameter β is finite-dimensional. In this section, we consider the general semiparametric moment condition model, where the parameter of interest θ is still finite-dimensional, but the moment function depends on an unknown functional nuisance parameter β, estimated in a first step nonparametrically or by machine learning.

We closely follow the setup of Chernozhukov et al. (2022a) and allow the moment conditions to be misspecified. Let $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ be the finite-dimensional parameter of interest and let β be an unknown nonparametric or high-dimensional first-step function, such as a conditional expectation,

taking values in a convex subset $\mathcal{B}$ of a normed vector space with norm $\|\cdot\|_{\mathcal{B}}$. Let $\beta_0 := \beta(P)$ denote the probability limit of a first-step estimator $\widehat{\beta}$ when P is the true probability measure of the data W. This limit is defined under general misspecification, as in Newey (1994); for instance, if $\widehat{\beta}(x)$ is a nonparametric estimator of the conditional expectation, then $\beta(P)(x) = \mathrm{E}_P[Y|X = x]$, where P is unrestricted, except for regularity conditions.

We assume that there is a vector of known moment functions $m : \mathcal{W} \times \Theta \times \mathcal{B} \to \mathbb{R}^{d_m}$ with $d_m \geq d_\theta$. Under correct specification, Chernozhukov et al. (2022a) assume that θ_0 is identified by the moment conditions

$$\mathrm{E}_P[m(W, \theta_0, \beta_0)] = 0, \tag{17}$$

that is, θ_0 is the unique solution to $\mathrm{E}_P[m(W, \theta, \beta_0)] = 0$ over $\theta \in \Theta$.

Plugging an estimator of β_0 into $m(w, \theta, \beta)$ generally leads to a first-order bias from the first-step estimation. Thus, we consider an orthogonal, or locally robust, moment function $\psi(w, \theta, \beta, \alpha)$, where α is an additional unknown function with true value $\alpha_0 := \alpha(P)$, such that ψ is locally insensitive to the first-step estimation of (β, α). For example, Chernozhukov et al. (2022a) consider the first-step influence function (FSIF), $\phi(w, \theta, \beta, \alpha)$, which may depend on α and satisfies

$$\frac{d}{d\tau}\mathrm{E}_P\big[m\big(W, \theta, \beta(P_\tau)\big)\big]\Big|_{\tau=0} = \int \phi(w, \theta, \beta_0, \alpha_0)\, \tilde{P}(dw), \qquad \mathrm{E}_P\big[\phi(W, \theta, \beta_0, \alpha_0)\big] = 0,$$

for all $\theta \in \Theta$ and all alternative distributions $\tilde{P}$, where $P_\tau = (1 - \tau)P + \tau\tilde{P}$ for $\tau \in [0, 1]$ and $\tilde{P}$ is unrestricted, except for regularity conditions such as the existence of $\beta(P_\tau)$. Then, Chernozhukov et al. (2022a) define the orthogonal moment function as

$$\psi(w, \theta, \beta, \alpha) := m(w, \theta, \beta) + \phi(w, \theta, \beta, \alpha) \in \mathbb{R}^{d_m},$$

and show that ψ is locally insensitive to β and that $\mathrm{E}_P[\phi(W, \theta, \beta_0, \alpha)] = 0$ for all $\alpha \in \mathcal{A}$, the set of possible values of α_0 under regularity conditions. Different approaches to the construction of Neyman orthogonal moment functions ψ, such as those obtained by concentrating out or by projection, may be used; see Newey (1994), Chernozhukov et al. (2015b) and Chernozhukov et al. (2018), for example.

Example 1 (Partially linear IV model)**.** *Let $W = (Y, D, X, Z)$ with $D \in \mathbb{R}^{d_\theta}$ and $Z \in \mathbb{R}^{d_z}$, $d_z \geq d_\theta$, and consider*

$$Y = D'\theta + g(X) + U, \qquad \mathrm{E}_P[U \mid X, Z] = 0, \tag{18}$$

where Z is the vector of instrumental variables. With $\ell_0(X) := \mathrm{E}_P[Y|X]$ and $r_0(X) := \mathrm{E}_P[D|X]$, the model implies $g(X) = \ell_0(X) - r_0(X)'\theta$, and a natural identifying moment function is $m(w, \theta, \beta) = Z\big(Y - \ell(X) - (D - r(X))'\theta\big)$ with $\beta = (\ell, r)$. Its FSIF is $\phi(w, \theta, \beta, \alpha) = -z(X)\big(Y - \ell(X) - (D - r(X))'\theta\big)$ with the additional nuisance parameter $\alpha = z$, whose true value is $z_0(X) := \mathrm{E}_P[Z|X]$ (Newey, 1994; Chernozhukov et al., 2022a), so that

$$\psi(w, \theta, \beta, \alpha) = \big(Z - z(X)\big)\big(Y - \ell(X) - (D - r(X))'\theta\big), \tag{19}$$

which is the Robinson form of the orthogonal moment (Robinson, 1988; Chernozhukov et al., 2018). Let $\tilde{Z} := Z - z_0(X)$, $\tilde{Y} := Y - \ell_0(X)$ *and* $\tilde{D} := D - r_0(X)$. *Then, we have* $\mathrm{E}_P\big[m(W,\theta,\beta_0)\big] = \mathrm{E}_P\big[\psi(W,\theta,\beta_0,\alpha_0)\big] = \mathrm{E}_P[Z\tilde{Y}] - \mathrm{E}_P[Z\tilde{D}']\theta = \mathrm{E}_P[\tilde{Z}\tilde{Y}] - \mathrm{E}_P[\tilde{Z}\tilde{D}']\theta$.

Throughout the paper, we define the population moment functions as follows:

$$\bar{m}(\theta) := \mathrm{E}_P\big[m(W,\theta,\beta_0)\big] = \mathrm{E}_P\big[\psi(W,\theta,\beta_0,\alpha_0)\big] \qquad \text{for every } \theta \in \Theta. \tag{20}$$

When we use the approach of Chernozhukov et al. (2022a), for example, (20) holds by Theorem 2 of Chernozhukov et al. (2022a), where the mean-zero property $\mathrm{E}_P\big[\phi(W,\theta,\beta_0,\alpha)\big] = 0$ holds for all $\theta \in \Theta$ and $\alpha \in \mathcal{A}$.[5]

We say that the model is correctly specified if $\bar{m}(\theta) = 0$ for some $\theta \in \Theta$, and misspecified otherwise. As in Section 2.2, the moment condition may fail, that is, $\bar{m}(\theta) \neq 0$ for all $\theta \in \Theta$. In Example 1, the moment condition $\bar{m}(\theta) = \mathrm{E}_P[ZU(\theta)] = \mathrm{E}_P[\tilde{Z}\,\mathrm{E}_P[U(\theta) \mid X, Z]]$, $U(\theta) := \tilde{Y} - \tilde{D}'\theta$, may fail when the conditional moment restriction in (18) fails to hold, for instance because some instruments are invalid or the effect of D is heterogeneous. In the just-identified case $d_z = d_\theta$ with $\mathrm{E}_P[\tilde{Z}\tilde{D}']$ nonsingular, $\bar{m}(\theta) = 0$ has the unique solution $\theta_0 = \mathrm{E}_P[\tilde{Z}\tilde{D}']^{-1}\mathrm{E}_P[\tilde{Z}\tilde{Y}]$ whether or not (18) holds. In the over-identified case $d_z > d_\theta$, it has a solution only when $\mathrm{E}_P[\tilde{Z}\tilde{Y}]$ lies in the column space of $\mathrm{E}_P[\tilde{Z}\tilde{D}']$, which does not hold in general. Chernozhukov et al. (2018) consider the just-identified setup, i.e., scalar Z and D, and thus the associated standard errors are robust to misspecification when the number of moment functions is the same as the number of parameters of interest as noted in Chernozhukov et al. (2022a).

Under misspecification, we define the GMM estimand θ_0 as the solution to the following population first-order condition

$$Q(\theta_0) = 0, \qquad Q(\theta) := G(\theta)'\,\Omega(\theta)^{-1}\,\bar{m}(\theta) \in \mathbb{R}^{d_\theta}, \tag{21}$$

where

$$G(\theta) := \partial_{\theta'}\bar{m}(\theta) \in \mathbb{R}^{d_m \times d_\theta}, \qquad \Omega(\theta) := \mathrm{E}_P\big[\zeta(W,\theta,\beta_0,\alpha_0)\big] \in \mathbb{R}^{d_m \times d_m}, \tag{22}$$

with $\zeta : \mathcal{W} \times \Theta \times \mathcal{B} \times \mathcal{A} \to \mathbb{R}^{d_m \times d_m}$ a known symmetric weight function. We use $\zeta(w,\theta,\beta,\alpha) = \zeta(w,\beta,\alpha)$ for one-step GMM and $\zeta(w,\theta,\beta,\alpha) = \psi(w,\theta,\beta,\alpha)\psi(w,\theta,\beta,\alpha)'$ for iterated GMM, and define $\bar{m}_0 := \bar{m}(\theta_0)$, $G_0 := G(\theta_0)$ and $\Omega_0 := \Omega(\theta_0)$.

Under regularity conditions, θ_0 is the probability limit of a version of the debiased GMM estimator of Chernozhukov et al. (2022a, equations (2.6)–(2.7))

$$\widehat{\theta}^{\mathrm{DGMM}} := \operatorname*{arg\,min}_{\theta \in \Theta}\ \widehat{\psi}_N(\theta)'\,\widehat{\Omega}_N^{\mathrm{cf}}(\widetilde{\theta})^{-1}\,\widehat{\psi}_N(\theta), \tag{23}$$

[5] Condition (20) holds for any construction that adds to m with a correction that has mean zero at (β_0, α_0) for every θ, such as the FSIF, but it can fail when the first-step limit depends on θ_0. Theorems 1–4 and the discussion below are unchanged with $\bar{m}(\theta)$ defined as $\mathrm{E}_P[\psi(W,\theta,\beta_0,\alpha_0)]$ by taking the orthogonal moment ψ as the primitive object.

where we take a K-fold partition $(I_k)_{k=1}^K$ of $\{1,\dots,N\}$, let $(\widehat{\beta}_k, \widehat{\alpha}_k)$ be estimators of (β_0, α_0) constructed from the observations not in I_k, and

$$\widehat{\psi}_N(\theta) := \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\psi(W_i,\theta,\widehat{\beta}_k,\widehat{\alpha}_k), \qquad \widehat{\Omega}_N^{\mathrm{cf}}(\theta) := \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\zeta(W_i,\theta,\widehat{\beta}_k,\widehat{\alpha}_k)$$

are the cross-fitted sample moment and weight. In the one-step case $\widehat{\Omega}_N^{\mathrm{cf}}$ is free of θ, and in the iterated case $\widetilde{\theta} = \widehat{\theta}^{\mathrm{DGMM}}$, so that (23) is defined as a fixed point.[6] Cross-fitting helps to remove the own-observation bias from the infinite-dimensional first step, and avoids the Donsker conditions for $(\widehat{\beta}_k, \widehat{\alpha}_k)$; see Newey and Robins (2018), Chernozhukov et al. (2018) and Chernozhukov et al. (2022a), for example.

3.1 The misspecification-robust score

The debiased GMM estimator (23) reduces the d_m orthogonal moments to d_θ dimensions through $\Gamma'\Omega^{-1}\psi$, with true values $\Gamma = G_0$ and $\Omega = \Omega_0$, and the associated conventional variance

$$V = \big(G_0'\Omega_0^{-1}G_0\big)^{-1}G_0'\Omega_0^{-1}\,\mathrm{E}_P\big[\psi(W,\theta_0,\beta_0,\alpha_0)\psi(W,\theta_0,\beta_0,\alpha_0)'\big]\,\Omega_0^{-1}G_0\big(G_0'\Omega_0^{-1}G_0\big)^{-1} \tag{24}$$

is that of Chernozhukov et al. (2022a, Theorem 9). Under correct specification the reduction $G_0'\Omega_0^{-1}\psi$ is Neyman orthogonal and (24) is valid. Under misspecification neither is true, because $\Gamma'\Omega^{-1}\psi$ is not orthogonal to the estimated Γ and Ω. As in Section 2.2, we therefore stack the moments together with Γ and Ω, taking the analogue of (9) built on ψ with the parameter of interest θ alone,

$$h(w,\theta,\lambda,\beta,\alpha) := \begin{pmatrix} \mathrm{vec}\big(\Gamma - \partial_{\theta'}\psi(w,\theta,\beta,\alpha)\big) \\ \mathrm{vec}\big(\Omega - \zeta(w,\theta,\beta,\alpha)\big) \\ \Gamma'\,\Omega^{-1}\,\psi(w,\theta,\beta,\alpha) \end{pmatrix} \in \mathbb{R}^{d_h}, \tag{25}$$

where $\lambda := (\mathrm{vec}(\Gamma)', \mathrm{vec}(\Omega)')' \in \Lambda \subset \mathbb{R}^{d_\lambda}$ with $\Gamma \in \mathbb{R}^{d_m\times d_\theta}$, $\Omega \in \mathbb{R}^{d_m\times d_m}$, $d_\lambda = d_m d_\theta + d_m^2$ and $d_h = d_\theta + d_\lambda$. With $\lambda_0 := (\mathrm{vec}(G_0)', \mathrm{vec}(\Omega_0)')'$, the first two blocks have mean zero by (22), and the third by (20) and (21), so

$$\mathrm{E}_P\big[h(W,\theta_0,\lambda_0,\beta_0,\alpha_0)\big] = 0 \tag{26}$$

by construction whether or not $\bar{m}_0 = 0$.

As in Remark 1, the stacking changes the representation, not the estimator. Solving the cross-fitted sample analogue of (26), $N^{-1}\sum_{k=1}^K\sum_{i\in I_k} h(W_i,\widehat{\theta},\widehat{\lambda},\widehat{\beta}_k,\widehat{\alpha}_k) = 0$, the first two blocks give

[6] There is a minor difference between the estimator (23) and those of Chernozhukov et al. (2022a). Chernozhukov et al. (2022a) evaluate the FSIF part of the cross-fitted moment at a preliminary estimator $\widetilde{\theta}_k$ of θ_0, $m(W_i,\theta,\widehat{\beta}_k) + \phi(W_i,\widetilde{\theta}_k,\widehat{\beta}_k,\widehat{\alpha}_k)$ in place of $\psi(W_i,\theta,\widehat{\beta}_k,\widehat{\alpha}_k)$.

$\widehat{\Gamma} = \partial_{\theta'} \widehat{\psi}_N(\widehat{\theta})$ and $\widehat{\Omega} = \widehat{\Omega}_N^{\mathrm{cf}}(\widehat{\theta})$, and the third block gives

$$\partial_{\theta'} \widehat{\psi}_N(\widehat{\theta})' \, \widehat{\Omega}_N^{\mathrm{cf}}(\widehat{\theta})^{-1} \, \widehat{\psi}_N(\widehat{\theta}) = 0,$$

which is the first-order condition of the debiased GMM estimator in (23). The stacked cross-fitted system therefore reproduces the debiased GMM estimator (23) exactly in finite samples, not only asymptotically.

We impose the following assumptions for the Neyman orthogonality based on the stacked moment functions.

Assumption 1 (Model and orthogonal moment function)**.** *Let $\mathcal{A}$ be a convex subset of a normed vector space with norm $\|\cdot\|_{\mathcal{A}}$ containing α_0, let $\mathcal{B}_N \subset \mathcal{B}$ and $\mathcal{A}_N \subset \mathcal{A}$ be nuisance realization sets containing β_0 and α_0. For all $P \in \mathcal{P}_N$:*

(a) θ_0 in (21) exists, is unique and lies in the interior of Θ, where Θ is compact and convex; $\bar{m}$ and Ω are twice continuously differentiable on Θ_0, a neighbourhood of θ_0. Ω_0 is positive definite, and $J_0 := \partial_{\theta'} Q(\theta_0)$ is nonsingular; $\theta \mapsto \psi(w, \theta, \beta_0, \alpha_0)$ is continuously differentiable on Θ_0 with $\sup_{\theta \in \Theta_0} \|\partial_{\theta'} \psi(W, \theta, \beta_0, \alpha_0)\| \leq \bar{d}(W)$ for some $\bar{d}$ with $\mathrm{E}_P[\bar{d}(W)] < \infty$.

(b) For each $\theta \in \Theta_0$, the maps $(\beta, \alpha) \mapsto \mathrm{E}_P[\psi(W, \theta, \beta, \alpha)]$, $(\beta, \alpha) \mapsto \mathrm{E}_P[\partial_{\theta'} \psi(W, \theta, \beta, \alpha)]$ and $(\beta, \alpha) \mapsto \mathrm{E}_P[\zeta(W, \theta, \beta, \alpha)]$ are twice continuously Gateaux differentiable on $\mathcal{B}_N \times \mathcal{A}_N$. Moreover, $\partial_{\theta'}$ and $\partial_{(\beta,\alpha)}$ can be interchanged at $(\theta_0, \beta_0, \alpha_0)$.

(c) The following holds for $\psi(w, \theta, \beta, \alpha)$:

$$\mathrm{E}_P\big[\psi(W, \theta, \beta_0, \alpha_0)\big] = \mathrm{E}_P\big[m(W, \theta, \beta_0)\big] \qquad \text{for all } \theta \in \Theta, \tag{27}$$

and

$$\partial_{(\beta,\alpha)} \mathrm{E}_P\big[\psi(W, \theta, \beta_0, \alpha_0)\big][\beta - \beta_0, \alpha - \alpha_0] = 0 \qquad \text{for all } \theta \in \Theta_0, (\beta, \alpha) \in \mathcal{B}_N \times \mathcal{A}_N. \tag{28}$$

(d) The weight function $\zeta(w, \theta, \beta, \alpha)$ satisfies the following:

$$\partial_{(\beta,\alpha)} \mathrm{E}_P\big[\zeta(W, \theta_0, \beta_0, \alpha_0)\big][\beta - \beta_0, \alpha - \alpha_0] = 0 \qquad \text{for all } (\beta, \alpha) \in \mathcal{B}_N \times \mathcal{A}_N. \tag{29}$$

Assumptions 1(a) and (b) are standard regularity conditions. Assumption 1(b) is similar to Chernozhukov et al. (2018, Assumption 3.1(c)), while Assumption 1(a) requires additional smoothness conditions for $\bar{m}$ and Ω in θ because the estimand θ_0 is defined by the first-order condition. The interchange of derivatives is imposed in Assumption 1(b), and this trivially holds when ψ is linear in θ.

Assumption 1(c) imposes two conditions on the orthogonal moment function ψ: that it has the same population mean as the identifying moment function m, and that ψ is Neyman orthogonal

with respect to (β, α) uniformly on Θ_0. Assumption 1(c) holds for $\psi = m + \phi$ constructed from the FSIF, by Theorems 1–3 of Chernozhukov et al. (2022a). Assumption 1(d) requires the Gateaux derivative of the weight matrix term to vanish. Assumption 1(d) holds for a weight function that does not depend on parameters, $\zeta(w, \theta, \beta, \alpha) = \zeta(w)$, but it also allows weight functions that depend on nuisance parameters. In Example 1, both $\zeta(w) = ZZ'$ and $\zeta(w, \alpha) = \big(Z - z(X)\big)\big(Z - z(X)\big)'$ satisfy (29). Under $\bar{m}_0 = 0$, the Jacobian and weight matrix channels vanish, and condition (29) is not needed, so that orthogonality at θ_0 suffices and the conditions reduce to those of Chernozhukov et al. (2018). Assumption 1(d) fails to hold for the efficient weight function $\zeta = \psi\psi'$, whose Gateaux derivative involves the moment residual, and the orthogonal moment construction involves the additional correction terms; see Remark 3 below and Appendix B for general weight function.

Remark 2 (Partially linear IV model (cont.))**.** *In the partially linear IV model in Example 1, one can also consider the identifying moment function* $m(w, \theta, \beta) = Z\big(Y - D'\theta - g(X)\big)$ *with* $\beta = g$*, and the orthogonal moment function* $\psi(w, \theta, \beta, \alpha) = \big(Z - z(X)\big)\big(Y - D'\theta - g(X)\big)$ *with* $\alpha = z$ *(Chernozhukov et al., 2018, equation (4.7)). However, its true value* $g_0(X) = \mathrm{E}_P[Y - D'\theta_0|X]$ *depends on* θ_0 *and hence, under misspecification, on the weighting matrix. Furthermore, this alternative form of* ψ *does not satisfy Assumption 1(c); it is orthogonal at* θ_0 *but not uniformly in* θ*: at* $\theta \neq \theta_0$*, its Gateaux derivative in the direction* $z - z_0$ *is* $-\mathrm{E}_P\big[(z - z_0)(X)\, r_0(X)'\big](\theta_0 - \theta) \neq 0$ *in general. The Robinson form* (19) *has advantages because its nuisance parameters* $\beta = (\ell, r)$ *and* $\alpha = z$ *are conditional expectations defined without reference to* θ_0 *or to the model, so Assumption 1(c) above holds by construction.*

The next result provides the Neyman orthogonal moment $\psi^*(w, \theta, \eta)$ that is robust to misspecification of the moment conditions (17), i.e., without imposing $\mathrm{E}_P\big[m(W, \theta_0, \beta_0)\big] = 0$.

Theorem 1 (Misspecification-robust Neyman orthogonal scores with infinite-dimensional nuisance parameters)**.** *Suppose Assumption 1 holds, and consider the score*

$$\psi^*(w, \theta, \eta) := \mu\, h(w, \theta, \lambda, \beta, \alpha), \qquad \eta := \big(\lambda', \mathrm{vec}(\mu)', \beta, \alpha\big) \in T := \Lambda \times \mathbb{R}^{d_\theta d_h} \times \mathcal{B} \times \mathcal{A}, \tag{30}$$

where $\mu \in \mathbb{R}^{d_\theta \times d_h}$ *has true value* μ_0 *given by* (12) *with*

$$J := \mathrm{E}_P\big[\partial_{(\theta', \lambda')} h(W, \theta_0, \lambda_0, \beta_0, \alpha_0)\big] = \big[\, J_\theta \;\; J_\lambda \,\big] \in \mathbb{R}^{d_h \times d_h},$$

and (A_h, Ω_h) *chosen such that* $\mu_0 J_\theta = J_0$*, and* $\eta_0 := (\lambda_0', \mathrm{vec}(\mu_0)', \beta_0, \alpha_0)$*. Then*

$$\mathrm{E}_P\big[\psi^*(W, \theta_0, \eta_0)\big] = 0, \tag{31}$$

and ψ^* *is Neyman orthogonal at* (θ_0, η_0) *with respect to* $T_N := \Lambda \times \mathbb{R}^{d_\theta d_h} \times \mathcal{B}_N \times \mathcal{A}_N$*. Furthermore, for all* $\theta \in \Theta$*,*

$$\begin{aligned} \psi^*(w, \theta, \eta_0) = G_0' \Omega_0^{-1} \psi(w, \theta, \beta_0, \alpha_0) + \big(\partial_{\theta'} \psi(w, \theta, \beta_0, \alpha_0) - G_0\big)' \Omega_0^{-1} \bar{m}_0 \\ - G_0' \Omega_0^{-1} \big(\zeta(w, \theta, \beta_0, \alpha_0) - \Omega_0\big) \Omega_0^{-1} \bar{m}_0, \end{aligned} \tag{32}$$

and $\partial_{\theta'} \mathrm{E}_P[\psi^*(W, \theta, \eta_0)]|_{\theta=\theta_0} = J_0$*, where*

$$J_0 = G_0' \Omega_0^{-1} G_0 + \left(I_{d_\theta} \otimes \bar{m}_0' \Omega_0^{-1}\right) \partial_{\theta'} \operatorname{vec} G(\theta_0) - \left(\bar{m}_0' \Omega_0^{-1} \otimes G_0' \Omega_0^{-1}\right) \partial_{\theta'} \operatorname{vec} \Omega(\theta_0). \tag{33}$$

Theorem 1 generalizes Propositions 1 and 2, and also extends the misspecified GMM literature (e.g., Hall and Inoue (2003), Hansen and Lee (2021)) to functional nuisance parameters; it shows that (32) is insensitive to the first-step estimation of β_0. The first term of (32) is the standard orthogonal score ψ, and the second and third are the Jacobian and weight matrix orthogonalization parts, both proportional to $\bar{m}_0$. Under correct specification, $\mu_0 = [\,0 \;\; 0 \;\; I_{d_\theta}\,]$, $J_0 = G_0' \Omega_0^{-1} G_0$ and $\psi^*(w, \theta, \eta_0) = G_0' \Omega_0^{-1} \psi(w, \theta, \beta_0, \alpha_0)$, so that the asymptotic variance based on ψ^* coincides with (24).

3.2 DML estimation and inference

We now define the DML estimator of Chernozhukov et al. (2018, Definition 3.2 (DML2)) with the misspecification-robust score ψ^*, using the cross-fitted sample moments and weights defined below, and develop general asymptotic theory for the DML estimator.

Definition 2 (DML with the misspecification-robust score)**.** *(a) Take a K-fold random partition $(I_k)_{k=1}^K$ of $[N] = \{1, \ldots, N\}$ such that the size of each fold I_k is $n = N/K$, and for each $k \in [K]$ define $I_k^c := [N] \setminus I_k$.*

(b) For each $k \in [K]$, construct estimators $\widehat{\beta}_k$ and $\widehat{\alpha}_k$ of β_0 and α_0 from $(W_i)_{i \in I_k^c}$.

(c) Let $\widetilde{\theta}$ be a preliminary estimator of θ_0, for instance the debiased GMM estimator $\widehat{\theta}^{\mathrm{DGMM}}$ of (23)*. Construct the cross-fitted Jacobian, weight matrix and moment,*

$$\widehat{\Gamma} = \frac{1}{N} \sum_{k=1}^K \sum_{i \in I_k} \partial_{\theta'} \psi(W_i, \widetilde{\theta}, \widehat{\beta}_k, \widehat{\alpha}_k), \quad \widehat{\Omega} = \frac{1}{N} \sum_{k=1}^K \sum_{i \in I_k} \zeta(W_i, \widetilde{\theta}, \widehat{\beta}_k, \widehat{\alpha}_k) = \widehat{\Omega}_N^{\mathrm{cf}}(\widetilde{\theta}),$$
$$\widehat{\bar{m}} = \frac{1}{N} \sum_{k=1}^K \sum_{i \in I_k} \psi(W_i, \widetilde{\theta}, \widehat{\beta}_k, \widehat{\alpha}_k),$$

and set $\widehat{\lambda} := (\operatorname{vec}(\widehat{\Gamma})', \operatorname{vec}(\widehat{\Omega})')'$,

$$\widehat{\mu} := \Big[- \left(I_{d_\theta} \otimes \widehat{\bar{m}}' \widehat{\Omega}^{-1}\right) \quad \widehat{\bar{m}}' \widehat{\Omega}^{-1} \otimes \widehat{\Gamma}' \widehat{\Omega}^{-1} \quad I_{d_\theta} \Big], \qquad \widehat{\eta}_k := \left(\widehat{\lambda}', \operatorname{vec}(\widehat{\mu})', \widehat{\beta}_k, \widehat{\alpha}_k\right).$$

(d) Construct $\check{\theta}$ as the solution of

$$\frac{1}{N} \sum_{k=1}^K \sum_{i \in I_k} \psi^*(W_i, \check{\theta}, \widehat{\eta}_k) = 0, \tag{34}$$

where, as in Theorem 1,

$$\psi^*(W_i,\theta,\widehat{\eta}_k) = \widehat{\mu}\, h(W_i,\theta,\widehat{\lambda},\widehat{\beta}_k,\widehat{\alpha}_k) = \widehat{\Gamma}'\widehat{\Omega}^{-1}\psi(W_i,\theta,\widehat{\beta}_k,\widehat{\alpha}_k) + \big(\partial_{\theta'}\psi(W_i,\theta,\widehat{\beta}_k,\widehat{\alpha}_k) - \widehat{\Gamma}\big)'\widehat{\Omega}^{-1}\widehat{\bar{m}} - \widehat{\Gamma}'\widehat{\Omega}^{-1}\big(\zeta(W_i,\theta,\widehat{\beta}_k,\widehat{\alpha}_k) - \widehat{\Omega}\big)\widehat{\Omega}^{-1}\widehat{\bar{m}}. \tag{35}$$

The DML estimator of Chernozhukov et al. (2018) based on the original orthogonal moment function ψ derived under correct specification, as in Lemma 1, instead solves

$$\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\widehat{\Gamma}'\widehat{\Omega}^{-1}\psi(W_i,\theta,\widehat{\beta}_k,\widehat{\alpha}_k) = 0, \tag{36}$$

whose estimating function is the first term of (35). The two estimating equations differ by the sample means of the Jacobian and weight matrix channels, which are proportional to $\widehat{\bar{m}}$. Under correct specification $\widehat{\bar{m}} = O_P(N^{-1/2})$ and the two estimators are first-order equivalent. Under misspecification the reduction $\widehat{\Gamma}'\widehat{\Omega}^{-1}\psi$ is not orthogonal to (Γ,Ω), and the conventional variance (24) attached to (36) is invalid.

Next, we focus on the moment condition models with linear scores same as in Chernozhukov et al. (2018, Section 3.2), and with a one-step weight function ζ.

Assumption 2 (Linear scores and one-step weighting)**.** *For all $w \in \mathcal{W}$, $\theta \in \Theta$, $\beta \in \mathcal{B}$ and $\alpha \in \mathcal{A}$,*

$$\psi(w,\theta,\beta,\alpha) = \psi^b(w,\beta,\alpha) + \psi^a(w,\beta,\alpha)\,\theta, \tag{37}$$

and $\zeta(w,\theta,\beta,\alpha) = \zeta(w,\beta,\alpha)$.

Under Assumption 2, $\partial_{\theta'}\psi(w,\theta,\beta,\alpha) = \psi^a(w,\beta,\alpha)$, and substituting (37) into (32) gives $\psi^*(w,\theta,\eta_0) = \psi^{*a}(w,\eta_0)\,\theta + \psi^{*b}(w,\eta_0)$ with $\psi^{*a}(w,\eta_0) = G_0'\Omega_0^{-1}\psi^a(w,\beta_0,\alpha_0)$. Also the estimated score (35) is linear in θ,

$$\psi^*(W_i,\theta,\widehat{\eta}_k) = \psi^{*a}(W_i,\widehat{\eta}_k)\,\theta + \psi^{*b}(W_i,\widehat{\eta}_k),$$

where

$$\begin{aligned}\psi^{*a}(W_i,\widehat{\eta}_k) &:= \widehat{\Gamma}'\widehat{\Omega}^{-1}\psi^a(W_i,\widehat{\beta}_k,\widehat{\alpha}_k),\\ \psi^{*b}(W_i,\widehat{\eta}_k) &:= \widehat{\Gamma}'\widehat{\Omega}^{-1}\psi^b(W_i,\widehat{\beta}_k,\widehat{\alpha}_k) + \big(\psi^a(W_i,\widehat{\beta}_k,\widehat{\alpha}_k) - \widehat{\Gamma}\big)'\widehat{\Omega}^{-1}\widehat{\bar{m}}\\ &\quad - \widehat{\Gamma}'\widehat{\Omega}^{-1}\big(\zeta(W_i,\widehat{\beta}_k,\widehat{\alpha}_k) - \widehat{\Omega}\big)\widehat{\Omega}^{-1}\widehat{\bar{m}},\end{aligned}$$

and here $\widehat{\Gamma} = N^{-1}\sum_{k=1}^{K}\sum_{i\in I_k}\psi^a(W_i,\widehat{\beta}_k,\widehat{\alpha}_k)$ and $\widehat{\Omega}$ of Definition 2 do not depend on $\widetilde{\theta}$. Further,

$$J_0 = \partial_{\theta'}\mathrm{E}_P[\psi^*(W,\theta,\eta_0)]|_{\theta=\theta_0} = \mathrm{E}_P\big[\psi^{*a}(W,\eta_0)\big] = G_0'\Omega_0^{-1}G_0,$$

because $\mathrm{E}_P[\psi^a(W,\beta_0,\alpha_0)] = G_0$. The conventional variance (24) and the robust variance of Theorem 2 below share the same Jacobian J_0, and differ only in their middle terms, $G_0'\Omega_0^{-1}\mathrm{E}_P[\psi\psi']\Omega_0^{-1}G_0$ against $\mathrm{E}_P[\psi^*\psi^{*\prime}]$, which coincide when $\bar{m}_0 = 0$, because $\psi^* = G_0'\Omega_0^{-1}\psi$.

Under Assumption 2, the estimating equation (34) has a closed-form solution

$$\check{\theta} = -\big(\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\psi^{*a}(W_i,\widehat{\eta}_k)\big)^{-1}\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\psi^{*b}(W_i,\widehat{\eta}_k) = -\big(\widehat{\Gamma}'\widehat{\Omega}^{-1}\widehat{\Gamma}\big)^{-1}\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\psi^{*b}(W_i,\widehat{\eta}_k).$$

Also, under Assumption 2,

$$\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\partial_{\theta'}\psi(W_i,\theta,\widehat{\beta}_k,\widehat{\alpha}_k) = \widehat{\Gamma} \quad \text{for all } \theta, \qquad \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\zeta(W_i,\widehat{\beta}_k,\widehat{\alpha}_k) = \widehat{\Omega},$$

so the sample means of the last two terms in (35) vanish and (34) reduces to (36). Hence DML based on ψ^*, DML based on the first term of ψ^* ($\widehat{\Gamma}'\widehat{\Omega}^{-1}\psi$), and the debiased GMM estimator coincide identically in finite samples. The misspecification-robust score construction therefore leaves the estimator exactly as it is and changes only the variance estimator; a practitioner who already reports DML/debiased GMM keeps the same point estimate and replaces the conventional standard error with the robust standard error. The conventional variance (24) omits the Jacobian and weight matrix parts of ψ^* in (32); these are of the same order as the first term and vanish only when $\bar{m}_0 = 0$, so the conventional variance is inconsistent when $\bar{m}_0 \neq 0$. This is the semiparametric counterpart of the results in Hall and Inoue (2003) and Hansen and Lee (2021). If the score is non-linear in θ, or the weight matrix depends on θ, the DML estimator $\check{\theta}$ differs from both the DML estimator solving (36) and the debiased GMM estimator.[7]

Assumption 3 (Regularity conditions)**.** *Let $q > 2$, $0 < c_0 \leq c_1$ and $C < \infty$ be constants, and let $\{\delta_N\}$ and $\{\Delta_N\}$ be sequences of positive constants converging to zero with $\delta_N \geq N^{-[(1-2/q)\wedge 1/2]}$. For all $N \geq 3$ and $P \in \mathcal{P}_N$, the following conditions hold.*

(a) With probability at least $1-\Delta_N$, $(\widehat{\beta}_k,\widehat{\alpha}_k) \in \mathcal{B}_N \times \mathcal{A}_N$ for all $k \in [K]$.

(b) The following conditions hold uniformly over $(\beta,\alpha) \in \mathcal{B}_N \times \mathcal{A}_N$,

$$\begin{aligned}
&\big(\mathrm{E}_P\|\psi(W,\theta_0,\beta,\alpha)\|^q\big)^{1/q} \vee \big(\mathrm{E}_P\|\psi^a(W,\beta,\alpha)\|^q\big)^{1/q} \vee \big(\mathrm{E}_P\|\zeta(W,\beta,\alpha)\|^q\big)^{1/q} \leq C,\\
&\big\|\psi(W,\theta_0,\beta,\alpha) - \psi(W,\theta_0,\beta_0,\alpha_0)\big\|_{P,2} \vee \big\|\psi^a(W,\beta,\alpha) - \psi^a(W,\beta_0,\alpha_0)\big\|_{P,2}\\
&\qquad\qquad \vee \big\|\zeta(W,\beta,\alpha) - \zeta(W,\beta_0,\alpha_0)\big\|_{P,2} \leq \delta_N.
\end{aligned}$$

[7] In Definition 2, the first-step estimators $(\widehat{\beta}_k,\widehat{\alpha}_k)$ are cross-fitted as in Chernozhukov et al. (2018, Definition 3.2), whereas the finite-dimensional components $(\widehat{\Gamma},\widehat{\Omega},\widehat{m})$, and hence $\widehat{\mu}$, are averages over all folds. The nuisance parameters $\widehat{\eta}_k$ could instead be constructed for each fold k from $(W_i)_{i\in I_k^c}$ alone, as in Chernozhukov et al. (2018). We keep the above definition in the main paper for its computational simplicity and its relationship with the debiased GMM estimator (23), but we consider the usual fold-specific DML definition in Appendix B with non-linear scores and efficient weights.

(c) For all $(\beta,\alpha)\in\mathcal{B}_N\times\mathcal{A}_N$,

$$\begin{aligned}
&\big\|\mathrm{E}_P\big[\psi(W,\theta_0,\beta,\alpha)\big]-\mathrm{E}_P[\psi(W,\theta_0,\beta_0,\alpha_0)]\big\|\\
&\quad+\|\mathrm{E}_P[\psi(W,\theta_0,\beta_0,\alpha_0)]\|\Big(\big\|\mathrm{E}_P\big[\psi^a(W,\beta,\alpha)\big]-\mathrm{E}_P[\psi^a(W,\beta_0,\alpha_0)]\big\|\\
&\quad+\big\|\mathrm{E}_P\big[\zeta(W,\beta,\alpha)\big]-\mathrm{E}_P[\zeta(W,\beta_0,\alpha_0)]\big\|\Big)\le\delta_N N^{-1/2},\\
&\big\|\mathrm{E}_P\big[\psi^a(W,\beta,\alpha)\big]-\mathrm{E}_P[\psi^a(W,\beta_0,\alpha_0)]\big\|+\big\|\mathrm{E}_P\big[\zeta(W,\beta,\alpha)\big]-\mathrm{E}_P[\zeta(W,\beta_0,\alpha_0)]\big\|\le\delta_N.
\end{aligned}$$

(d) The singular values of G_0 *and the eigenvalues of* Ω_0 *lie in* $[c_0,c_1]$*, and all eigenvalues of* $\mathrm{E}_P[\psi^*(W,\theta_0,\eta_0)\psi^*(W,\theta_0,\eta_0)']$ *are bounded below by* c_0.

Assumption 3 imposes primitive conditions on (ψ,ζ) and the first-step estimators $(\widehat{\beta}_k,\widehat{\alpha}_k)$. Assumption 3(a) is Assumption 3.2(a) of Chernozhukov et al. (2018). Assumption 3(b) corresponds to Assumption 3.2(b) and (c) of Chernozhukov et al. (2018) on the moment and the rate conditions, extended to the Jacobian and the weight matrix parts of (32), which contribute to the limit distribution.

Assumption 3(c) is required to control biases of the Jacobian and the weight matrix with the factor $\|\bar{m}_0\|$. Under correct specification, only the bias of ψ at θ_0 is restricted, while under misspecification the first-order biases of the Jacobian and of the weight matrix must also be $o(N^{-1/2})$. A sufficient condition for (c) is that the maps $(\beta,\alpha)\mapsto\mathrm{E}_P[\psi(W,\theta_0,\beta,\alpha)]$, $\mathrm{E}_P[\psi^a(W,\beta,\alpha)]$ and $\mathrm{E}_P[\zeta(W,\beta,\alpha)]$ have bounded second Gateaux derivatives and that the first step $(\widehat{\beta}_k,\widehat{\alpha}_k)$ converges at the familiar rate $o(N^{-1/4})$. Assumption 3(d) corresponds to Assumptions 3.1(e) and 3.2(d) of Chernozhukov et al. (2018), with the non-degeneracy condition imposed on the variance of ψ^*, which includes the Jacobian and weight matrix, instead of ψ.

Theorem 2 (DML inference for misspecified semiparametric GMM with linear scores)**.** *Suppose Assumptions 1, 2 and 3 hold. Then, the DML estimator* $\check{\theta}$ *of Definition 2 concentrates in a* $1/\sqrt{N}$ *neighbourhood of* θ_0 *and is approximately linear and centred Gaussian,*

$$\sqrt{N}\,\sigma^{-1}\big(\check{\theta}-\theta_0\big)=-\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sigma^{-1}J_0^{-1}\psi^*(W_i,\theta_0,\eta_0)+O_P(\delta_N)\ \xrightarrow{d}\ \mathcal{N}\big(0,I_{d_\theta}\big),\tag{38}$$

uniformly over $P\in\mathcal{P}_N$*, where* $\sigma^2:=J_0^{-1}\mathrm{E}_P[\psi^*(W,\theta_0,\eta_0)\psi^*(W,\theta_0,\eta_0)']J_0^{-1\prime}$*, and* $J_0=G_0'\Omega_0^{-1}G_0$*. Moreover, let*

$$\widehat{\sigma}^2:=\widehat{J}_0^{-1}\Big(\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\psi^*(W_i,\check{\theta},\widehat{\eta}_k)\,\psi^*(W_i,\check{\theta},\widehat{\eta}_k)'\Big)\widehat{J}_0^{-1\prime},\qquad\widehat{J}_0:=\widehat{\Gamma}'\widehat{\Omega}^{-1}\widehat{\Gamma}.\tag{39}$$

If $\|\widetilde{\theta}-\theta_0\|=O_P(\delta_N)$ *uniformly over* $P\in\mathcal{P}_N$*, then* $\widehat{\sigma}^2=\sigma^2+O_P(\delta_N)$*, and, for a nominal level* $a\in(0,1)$ *and any* $\ell\in\mathbb{R}^{d_\theta}$ *with* $\|\ell\|=1$*, the confidence interval*

$$\mathrm{CI}_N:=\Big[\,\ell'\check{\theta}\pm\Phi^{-1}(1-a/2)\sqrt{\ell'\widehat{\sigma}^2\ell/N}\,\Big]\tag{40}$$

is uniformly valid for $\ell'\theta_0$ over $P \in \mathcal{P}_N$,

$$\sup_{P\in\mathcal{P}_N} \left| \Pr_P\left(\ell'\theta_0 \in \mathrm{CI}_N\right) - (1-a)\right| \longrightarrow 0,$$

where Φ is the standard normal distribution function.

Remark 3 (Efficient weighting and non-linear scores)**.** *Theorem 2 focuses on one-step weighting, $\zeta(w,\theta,\beta,\alpha) = \zeta(w,\beta,\alpha)$, while Theorem 1 only requires the weight function to satisfy Assumption 1(d). Under misspecification, Assumption 1(d) fails to hold for the efficient weight function $\zeta = \psi\psi'$. However, the orthogonal moment can be constructed without Assumption 1(d), by adding a correction term $\kappa(w,\theta,\eta)$ to ψ^*, where η is enlarged to include any further nuisances, such that $\mathrm{E}_P[\kappa(W,\theta_0,\eta_0)] = 0$, the derivatives of $\mathrm{E}_P[\kappa]$ vanish at the true values, and*

$$\partial_{(\beta,\alpha)}\mathrm{E}_P\big[\kappa(W,\theta_0,\eta_0)\big][\beta-\beta_0,\alpha-\alpha_0] = G_0'\Omega_0^{-1}\partial_{(\beta,\alpha)}\mathrm{E}_P[\zeta(W,\theta_0,\beta_0,\alpha_0)][\beta-\beta_0,\alpha-\alpha_0]\,\Omega_0^{-1}\bar{m}_0. \quad (41)$$

Then, the proof of Theorem 1 applies to $\psi^* + \kappa$. The right-hand side of (41) is also proportional to $\bar{m}_0$, so $\kappa = 0$ under correct specification or under Assumption 1(d).

Theorem 2 not only excludes the efficient weight $\zeta = \psi\psi'$, which is quadratic in θ, but also excludes moment functions that are non-linear in θ. In both cases, we can construct the Neyman orthogonal moment ψ^* and obtain valid inference for the DML estimator that solves the cross-fitted estimating equation for ψ^*, and Appendix B provides the general results (Theorem 5).

4 The partially linear IV model

We now consider our main example, the partially linear IV model in Example 1. To keep the notation simple, throughout this section we consider a scalar endogenous regressor case, $d_\theta = 1$. Let $W = (Y, D, X, Z)$, where $D \in \mathbb{R}$ is the endogenous regressor, X is a vector of controls, and $Z \in \mathbb{R}^{d_z}$ is a vector of instrumental variables with $d_z > 1$, so that the model is over-identified.

The correctly specified model is

$$Y = D\,\theta_0 + g_0(X) + U, \qquad \mathrm{E}_P[U \mid X, Z] = 0, \quad (42)$$

for some true values $\theta_0 \in \Theta$ and g_0, which we do not impose: the conditional moment restriction in (42) may fail, for instance because some instruments are invalid or because the effect of D is heterogeneous.

The nuisance parameters are the conditional expectations

$$\beta_0 := (\ell_0, r_0), \quad \alpha_0 := z_0, \qquad \ell_0(X) := \mathrm{E}_P[Y|X], \quad r_0(X) := \mathrm{E}_P[D|X], \quad z_0(X) := \mathrm{E}_P[Z|X], \quad (43)$$

with residuals $\tilde{Y} = Y - \ell_0(X) \in \mathbb{R}$, $\tilde{D} = D - r_0(X) \in \mathbb{R}$ and $\tilde{Z} = Z - z_0(X) \in \mathbb{R}^{d_z}$. Every component of (43) is defined without reference to θ_0 or to the model (42), as described in Remark 2. The

identifying moment function is $m(w,\theta,\beta) = Z\big(Y-\ell(X)-(D-r(X))\theta\big)$. The model is misspecified when there is no solution to

$$\mathrm{E}_P[m(W,\theta,\beta_0)] = \mathrm{E}_P\big[Z\big(Y-\ell_0(X)-(D-r_0(X))\theta\big)\big] = 0$$

over $\theta \in \Theta$. Adding the first-step influence function to m gives the orthogonal moment function (Chernozhukov et al., 2018, equation (4.8)),

$$\psi(w,\theta,\beta,\alpha) = \big(Z-z(X)\big)\big(Y-\ell(X)-(D-r(X))\theta\big) \in \mathbb{R}^{d_z}, \tag{44}$$

which is linear in θ in the sense of (37) with

$$\psi^b(w,\beta,\alpha) = \big(Z-z(X)\big)\big(Y-\ell(X)\big), \qquad \psi^a(w,\beta,\alpha) = -\big(Z-z(X)\big)\big(D-r(X)\big). \tag{45}$$

Let $\Pi := \mathrm{E}_P[\tilde{Z}\tilde{D}] \in \mathbb{R}^{d_z}$. Then

$$\bar{m}(\theta) = \mathrm{E}_P\big[\tilde{Z}\tilde{Y}\big] - \Pi\,\theta, \qquad G(\theta) \equiv G_0 = \mathrm{E}_P\big[\psi^a(W,\beta_0,\alpha_0)\big] = -\Pi.$$

We consider the usual 2SLS weight matrix

$$\zeta(w,\alpha) = \big(Z-z(X)\big)\big(Z-z(X)\big)', \qquad \Omega_0 = S := \mathrm{E}_P\big[\tilde{Z}\tilde{Z}'\big] \in \mathbb{R}^{d_z\times d_z}. \tag{46}$$

Then, θ_0 solves the population first-order condition (21),

$$\theta_0 = \big(\Pi' S^{-1}\Pi\big)^{-1}\Pi' S^{-1}\,\mathrm{E}_P\big[\tilde{Z}\tilde{Y}\big], \qquad \bar{m}_0 = \mathrm{E}_P\big[\tilde{Z}\tilde{U}\big], \quad \tilde{U} := \tilde{Y} - \tilde{D}\theta_0. \tag{47}$$

As noted in Example 1, $\bar{m}(\theta) = 0$ for some θ if and only if $\mathrm{E}_P[\tilde{Z}\tilde{Y}]$ is proportional to Π, which for $d_z > 1$ is a genuine restriction: provided every $\Pi_j \neq 0$, it requires that the d_z instrument-specific IV estimands $\mathrm{E}_P[\tilde{Z}_j\tilde{Y}]/\mathrm{E}_P[\tilde{Z}_j\tilde{D}]$, $j = 1,\ldots,d_z$, are all the same. If (42) holds, then $\bar{m}_0 = \mathrm{E}_P[\tilde{Z}\,\mathrm{E}_P[U|X,Z]] = 0$ and the pseudo-true value (47) is the structural θ_0 of (42).

It is straightforward to check the conditions in Assumption 1 for the PLIV model. Assumption 1(a) holds when $\Pi \neq 0$, the relevance condition, and θ_0 is interior to Θ. Assumption 1(b) is also immediate. Because (44) is the identifying moment function plus its FSIF, it satisfies the uniform orthogonality condition of Assumption 1(c) by Chernozhukov et al. (2022a, Theorems 1–3), at every θ; see also Chernozhukov et al. (2018, Section 4.2). The Gateaux derivatives in the directions of $\beta = (\ell, r)$ and $\alpha = z$ are computed in the proof of Theorem 3; they vanish because $\mathrm{E}_P[\tilde{Z}|X] = 0$ and $\mathrm{E}_P[\tilde{Y}|X] = \mathrm{E}_P[\tilde{D}|X] = 0$ by the law of iterated expectations, and do not use the conditional moment restriction in (42). Assumption 1(d) also holds with (46).

4.1 Robust scores and the implementation

Substituting (45)–(46) into (32) gives the misspecification-robust score

$$\psi^*(w,\theta_0,\eta_0) = -\,\Pi' S^{-1}\tilde{Z}\tilde{U} - \big(\tilde{Z}\tilde{D} - \Pi\big)' S^{-1}\bar{m}_0 + \Pi' S^{-1}\big(\tilde{Z}\tilde{Z}' - S\big)S^{-1}\bar{m}_0, \tag{48}$$

with $J_0 = G_0'\Omega_0^{-1}G_0 = \Pi' S^{-1}\Pi$. The standard (correct-specification) Neyman orthogonal score of ψ is

$$G_0'\Omega_0^{-1}\psi(w,\theta_0,\beta_0,\alpha_0) = -\,\Pi' S^{-1}\tilde{Z}\tilde{U}. \tag{49}$$

For comparison, it is useful to see (48) and (49) explicitly with full conditional expectations:

$$\begin{gathered}\Pi = \mathrm{E}_P\big[\big(Z - z_0(X)\big)\big(D - r_0(X)\big)\big], \qquad S = \mathrm{E}_P\big[\big(Z - z_0(X)\big)\big(Z - z_0(X)\big)'\big], \\ \bar{m}_0 = \mathrm{E}_P\Big[\big(Z - z_0(X)\big)\big(Y - \ell_0(X) - (D - r_0(X))\theta_0\big)\Big],\end{gathered} \tag{50}$$

and the two orthogonal scores,

$$\begin{aligned} G_0'\Omega_0^{-1}\psi(w,\theta_0,\beta_0,\alpha_0) &= -\,\Pi' S^{-1}\big(Z - z_0(X)\big)\big(Y - \ell_0(X) - (D - r_0(X))\theta_0\big), \\ \psi^*(w,\theta_0,\eta_0) &= -\,\Pi' S^{-1}\big(Z - z_0(X)\big)\big(Y - \ell_0(X) - (D - r_0(X))\theta_0\big) \\ &\quad - \Big[\big(Z - z_0(X)\big)\big(D - r_0(X)\big) - \Pi\Big]' S^{-1}\bar{m}_0 \\ &\quad + \Pi' S^{-1}\Big[\big(Z - z_0(X)\big)\big(Z - z_0(X)\big)' - S\Big]S^{-1}\bar{m}_0. \end{aligned} \tag{51}$$

Every nuisance function appearing in (50)–(51) is a conditional expectation, directly estimable by machine learning whether or not the model is correctly specified, and Π, S and $\bar{m}_0$ are simple population moments of the residuals. Both correction terms in (51) are multiplied by $\bar{m}_0$, so it reduces to the standard orthogonal score under $\bar{m}_0 = 0$.

Now, we define the DML estimator in the PLIV setup following Definition 2. Let $\widehat{\beta}_k = (\widehat{\ell}_k, \widehat{r}_k)$ and $\widehat{\alpha}_k = \widehat{z}_k$ be machine learning or nonparametric estimators of the regressions of Y, D and Z on X, fitted on $(W_i)_{i\in I_k^c}$, and for $i \in I_k$ form the cross-fitted residuals

$$\widehat{\tilde{Y}}_i := Y_i - \widehat{\ell}_k(X_i), \qquad \widehat{\tilde{D}}_i := D_i - \widehat{r}_k(X_i), \qquad \widehat{\tilde{Z}}_i := Z_i - \widehat{z}_k(X_i), \qquad i \in I_k.$$

The cross-fitted Jacobian, weight matrix and moment in Definition 2(c) are the fold averages

$$\widehat{\Pi} := \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\widehat{\tilde{Z}}_i\widehat{\tilde{D}}_i, \qquad \widehat{S} := \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\widehat{\tilde{Z}}_i\widehat{\tilde{Z}}_i', \qquad \overline{\widehat{\tilde{Z}}\widehat{\tilde{Y}}} := \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\widehat{\tilde{Z}}_i\widehat{\tilde{Y}}_i.$$

The DML estimator $\check{\theta}$ solves (34), $N^{-1}\sum_k\sum_{i\in I_k}\psi^*(W_i,\theta,\widehat{\eta}_k) = 0$ with ψ^* given by (35). Because $\widehat{\Gamma}$ and $\widehat{\Omega}$ are the pooled averages of ψ^a and ζ above, the last two terms of (35) vanish, and

(34) reduces to (36), $\widehat{\Gamma}'\widehat{\Omega}^{-1}N^{-1}\sum_k\sum_{i\in I_k}\psi(W_i,\theta,\widehat{\beta}_k,\widehat{\alpha}_k)=0$. This gives

$$\check{\theta}=\big(\widehat{\Pi}'\widehat{S}^{-1}\widehat{\Pi}\big)^{-1}\widehat{\Pi}'\widehat{S}^{-1}\,\overline{\widehat{\tilde{Z}}\tilde{Y}}, \tag{52}$$

which is exactly cross-fitted two-stage least squares on the residualized data.

The estimated scores and Jacobian are

$$\begin{aligned}\widehat{\psi}_i^{*}&=-\widehat{\Pi}'\widehat{S}^{-1}\widehat{\tilde{Z}}_i\widehat{\tilde{U}}_i-\big(\widehat{\tilde{Z}}_i\widehat{\tilde{D}}_i-\widehat{\Pi}\big)'\widehat{S}^{-1}\widehat{\bar{m}}+\widehat{\Pi}'\widehat{S}^{-1}\big(\widehat{\tilde{Z}}_i\widehat{\tilde{Z}}_i'-\widehat{S}\big)\widehat{S}^{-1}\widehat{\bar{m}},\qquad i\in I_k,\\ \widehat{J}_0&=\widehat{\Pi}'\widehat{S}^{-1}\widehat{\Pi},\qquad \widehat{\sigma}^2=\widehat{J}_0^{-1}\,\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\big(\widehat{\psi}_i^{*}\big)^2\widehat{J}_0^{-1\prime},\end{aligned} \tag{53}$$

where $\widehat{\bar{m}}=\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\widehat{\tilde{Z}}_i\widehat{\tilde{U}}_i$ is the cross-fitted moment, and $\widehat{\tilde{U}}_i=\widehat{\tilde{Y}}_i-\widehat{\tilde{D}}_i\check{\theta}$ is a residual.

The misspecification-robust orthogonal score does not change the estimator and changes only the asymptotic variance, $\widehat{\sigma}^2$. The conventional DML standard error is obtained by dropping the last two terms of $\widehat{\psi}_i^{*}$ in (53); the difference between the two standard errors is entirely due to $\widehat{\bar{m}}$, and is asymptotically negligible when the moment conditions hold.

4.2 Asymptotic distribution

Assumption 4 (Regularity conditions for the PLIV)**.** *Let $q>4$ and $0<c_0\le c_1$, $C<\infty$ be constants and $\{\delta_N\},\{\Delta_N\}$ sequences of positive constants converging to zero. For all $P\in\mathcal{P}_N$, the following conditions hold.*

(a) θ_0 in (47) exists, is unique and lies in the interior of Θ, where $\Theta\subset\mathbb{R}$ is a compact interval.

(b) $\|Y\|_{P,q}+\|D\|_{P,q}+\|Z\|_{P,q}\le C$.

(c) (β_0,α_0) is given by (43) and the conditional second moments are bounded: $\|\mathrm{E}_P[\tilde{Y}^2|X]\|_{P,\infty}\vee\|\mathrm{E}_P[\tilde{D}^2|X]\|_{P,\infty}\vee\|\mathrm{E}_P[\|\tilde{Z}\|^2|X]\|_{P,\infty}\le C$.

(d) $c_0\le\|\Pi\|\le c_1$, the eigenvalues of S lie in $[c_0,c_1]$, and $\mathrm{E}_P\big[\psi^*(W,\theta_0,\eta_0)^2\big]\ge c_0$ for ψ^* of (48).

(e) With P-probability at least $1-\Delta_N$ the cross-fitted estimators satisfy $(\widehat{\beta}_k,\widehat{\alpha}_k)\in\mathcal{B}_N\times\mathcal{A}_N$ for all k, where every $\beta=(\ell,r)\in\mathcal{B}_N$ and $\alpha=z\in\mathcal{A}_N$ obey

$$\|\ell-\ell_0\|_{P,q}\vee\|r-r_0\|_{P,q}\vee\|z-z_0\|_{P,q}\le C,\qquad \|\ell-\ell_0\|_{P,2}\vee\|r-r_0\|_{P,2}\vee\|z-z_0\|_{P,2}\le\delta_N,$$

and the rate conditions

$$\|z-z_0\|_{P,2}\big(\|\ell-\ell_0\|_{P,2}+\|r-r_0\|_{P,2}\big)\le\delta_N N^{-1/2}\qquad\text{and}\qquad\|z-z_0\|_{P,2}^2\le\delta_N N^{-1/2}. \tag{54}$$

Assumption 4 differs from Chernozhukov et al. (2018, Assumption 4.2) in several aspects. First of all, we do not impose the conditional moment restriction (42), imposed in their Assumption 4.2(a)

as their equation (4.5). We consider the over-identified model with a vector Z, $d_z > d_\theta$, and θ_0 is the 2SLS estimands (47) with residualization. The moment conditions in Assumption 4(b) are their Assumption 4.2(b), and the bounded conditional second moments in Assumption 4(c) are similar to their Assumption 4.2(d). Assumption 4(d) is similar to their Assumption 4.2(c), with the non-degeneracy here imposed on the variance of the misspecification-robust score (48) instead. The two norm bounds on the first step in Assumption 4(e) are their Assumption 4.2(e). The first product rate in (54) is exactly Chernozhukov et al. (2018, Assumption 4.2(e)(ii)), while the second rate condition is new, and is required by the weight matrix channel and can be dropped when $\bar{m}_0 = 0$.

Theorem 3 (DML inference in the misspecified partially linear IV model)**.** *Suppose Assumption 4 holds. Then the DML estimator $\check{\theta}$ in* (52) *satisfies*

$$\sqrt{N}\,\sigma^{-1}\big(\check{\theta} - \theta_0\big) \xrightarrow{d} \mathcal{N}(0,1), \qquad \sigma^2 = \big(\Pi' S^{-1}\Pi\big)^{-1}\mathrm{E}_P\big[\psi^*(W,\theta_0,\eta_0)^2\big]\big(\Pi' S^{-1}\Pi\big)^{-1},$$

uniformly over $P \in \mathcal{P}_N$. Moreover $\widehat{\sigma}^2$ of (53) *satisfies $\widehat{\sigma}^2 = \sigma^2 + o_P(1)$, and, for a nominal level $a \in (0,1)$, the confidence interval based on $\check{\theta}$ is uniformly asymptotically valid for θ_0,*

$$\sup_{P\in\mathcal{P}_N}\Big|\,\mathrm{Pr}_P\left(\theta_0 \in \check{\theta} \pm \Phi^{-1}(1-a/2)\,\widehat{\sigma}/\sqrt{N}\right) - (1-a)\Big| \longrightarrow 0.$$

Theorem 3 extends the result of Chernozhukov et al. (2018, Theorem 4.2) to be robust to misspecification of the moment condition (42). Under the just-identified setup or under correct specification, $\bar{m}_0 = 0$, so that $\psi^*(w,\theta_0,\eta_0) = -\Pi' S^{-1}\tilde{Z}\tilde{U}$ is the standard orthogonal score (49), and the conclusion reduces to Chernozhukov et al. (2018, Theorem 4.2).

5 Local average treatment effects with multiple instruments

In this section, we consider estimation and inference on the local average treatment effects (LATEs) with a binary treatment variable and multiple binary instruments. This section parallels Chernozhukov et al. (2018, Section 5.2), who treat a single binary instrument.

The data are $W = (Y, D, X, Z)$ with outcome $Y \in \mathbb{R}$, binary treatment $D \in \{0,1\}$, covariates X, and a vector of binary instruments $Z = (Z_1, \ldots, Z_{d_z})' \in \{0,1\}^{d_z}$ with $d_z > 1$. Consider the functions $\beta_0 := \big(\ell_{0j}, \nu_{0j}\big)_{j\le d_z}, \alpha_0 := \big(p_{0j}\big)_{j\le d_z}$,

$$\ell_{0j}(z,X) := \mathrm{E}_P[Y \mid Z_j = z, X], \quad \nu_{0j}(z,X) := \mathrm{E}_P[D \mid Z_j = z, X], \quad p_{0j}(X) := \mathrm{Pr}_P(Z_j = 1 \mid X), \tag{55}$$

which are equations (5.6)–(5.8) of Chernozhukov et al. (2018). These are again conditional expectations, which exist and are well defined for every $P \in \mathcal{P}_N$ with finite second moments, and are estimable by machine learning regression.

Under the assumptions of Imbens and Angrist (1994) and Frölich (2007),[8] each instrument Z_j

[8] Let $D(z_1, \ldots, z_{d_z})$ denote the potential treatment, and $Y(d)$ the potential outcome under treatment d, so that $D =$

identifies

$$\theta^{(j)} := \mathrm{E}_P\big[Y(1) - Y(0) \mid D_j(1) > D_j(0)\big] = \frac{\mathrm{E}_P\big[\ell_{0j}(1,X) - \ell_{0j}(0,X)\big]}{\mathrm{E}_P\big[\nu_{0j}(1,X) - \nu_{0j}(0,X)\big]},$$

the LATE for the compliers of instrument Z_j. Taking this as the identifying moment function, $m_j(w,\theta,\beta) = \big[\ell_j(1,X) - \ell_j(0,X)\big] - \theta\big[\nu_j(1,X) - \nu_j(0,X)\big]$ for each instrument $j = 1,\dots,d_z$, and adding its first-step influence function, which is the inverse propensity score correction with additional nuisance parameter $\alpha = (p_j)_j$, gives

$$\psi = (\psi_1,\dots,\psi_{d_z})' \in \mathbb{R}^{d_z}, \qquad \psi_j(w,\theta,\beta,\alpha) = \psi_j^Y(w,\beta,\alpha) - \theta\,\psi_j^D(w,\beta,\alpha), \tag{56}$$

where, as in Chernozhukov et al. (2018, Section 5.2),

$$\begin{aligned}
\psi_j^Y(w,\beta,\alpha) &= \ell_j(1,X) - \ell_j(0,X) + \frac{Z_j\big(Y - \ell_j(1,X)\big)}{p_j(X)} - \frac{(1-Z_j)\big(Y - \ell_j(0,X)\big)}{1-p_j(X)},\\
\psi_j^D(w,\beta,\alpha) &= \nu_j(1,X) - \nu_j(0,X) + \frac{Z_j\big(D - \nu_j(1,X)\big)}{p_j(X)} - \frac{(1-Z_j)\big(D - \nu_j(0,X)\big)}{1-p_j(X)}.
\end{aligned} \tag{57}$$

By construction (56) is linear in θ in the sense of (37), with

$$\psi^b(w,\beta,\alpha) = \psi^Y(w,\beta,\alpha), \qquad \psi^a(w,\beta,\alpha) = -\psi^D(w,\beta,\alpha), \tag{58}$$

and the population moment and Jacobian are

$$\bar m(\theta) = \tau^Y - \tau^D\theta, \qquad G(\theta) \equiv G_0 = \mathrm{E}_P\big[\psi^a(W,\beta_0,\alpha_0)\big] = -\tau^D, \tag{59}$$

where $\tau^Y := \mathrm{E}_P\big[\psi^Y(W,\beta_0,\alpha_0)\big] \in \mathbb{R}^{d_z}$, $\tau^D := \mathrm{E}_P\big[\psi^D(W,\beta_0,\alpha_0)\big] \in \mathbb{R}^{d_z}$. Here $\tau_j^Y = \mathrm{E}_P[\ell_{0j}(1,X) - \ell_{0j}(0,X)]$ and $\tau_j^D = \mathrm{E}_P[\nu_{0j}(1,X) - \nu_{0j}(0,X)]$, so $\theta^{(j)} = \tau_j^Y/\tau_j^D$.

With a single instrument, $d_z = 1$, the system is just-identified, $\bar m(\theta) = 0$ always has a solution, and this is the case of Chernozhukov et al. (2018, Section 5.2), in which misspecification in our sense does not arise. With $d_z > 1$, the over-identifying restriction is that $\bar m(\theta) = 0$ for some θ, i.e., τ^Y is proportional to τ^D, which imposes

$$\theta^{(1)} = \theta^{(2)} = \cdots = \theta^{(d_z)}.$$

In general, different instruments identify different groups of compliers, and the average treatment effect need not be the same in each complier group. This is the source of misspecification here; it is not a failure of instrument validity.

The population GMM first-order condition in (21) is $\tau^{D\prime}\Omega_0^{-1}(\tau^Y - \tau^D\theta_0) = 0$, so that $\theta_0 =$

$D(Z)$ and $Y = Y(D)$, and write $D_j(z) := D(z, Z_{-j})$ for the potential treatment when Z_j is set to z. We can consider the following LATE conditions, which are imposed for each instrument Z_j, $j = 1,\dots,d_z$: 1) $\big(Y(1), Y(0), D(\cdot)\big) \perp Z \mid X$, and $Z_1,\dots,Z_{d_z}$ are mutually independent conditional on X; 2) $\epsilon \le p_{0j}(X) = \Pr_P(Z_j = 1 \mid X) \le 1 - \epsilon$ almost surely, for some $\epsilon \in (0, 1/2)$; 3) $D(1, z_{-j}) \ge D(0, z_{-j})$ almost surely for every value z_{-j} of the other instruments; 4) $\Pr_P\big(D_j(1) > D_j(0)\big)$ is bounded away from zero.

$\frac{\tau^{D\prime}\Omega_0^{-1}\tau^Y}{\tau^{D\prime}\Omega_0^{-1}\tau^D}$ with the weighting matrix Ω_0. In order for the target parameter to be a convex combination of LATEs, we use the following diagonal weight matrix that does not depend on θ,

$$\zeta(w) = \Omega_0 = \mathrm{diag}(\varpi_1, \ldots, \varpi_{d_z}), \qquad \varpi_j > 0. \tag{60}$$

The weights ϖ_j are assumed to be known throughout, with the identity matrix $\Omega_0 = I_{d_z}$ being one example. Another natural choice is $\zeta(w,\beta,\alpha) = \mathrm{diag}\left(\psi_j^D(w,\beta,\alpha)\right)$, for which $\Omega_0 = \mathrm{diag}\left(\mathrm{E}_P[\psi_j^D(W,\beta_0,\alpha_0)]\right) = \mathrm{diag}(\tau_j^D)$ and each instrument is weighted by its complier share, so that an instrument with a stronger first stage receives more weight. This weight satisfies Assumption 1(d), because each ψ_j^D is an orthogonal score. With this choice, the GMM estimand is

$$\theta_0 = \frac{\tau^{D\prime}\Omega_0^{-1}\tau^Y}{\tau^{D\prime}\Omega_0^{-1}\tau^D} = \sum_{j=1}^{d_z} \omega_j\, \theta^{(j)}, \qquad \omega_j = \frac{\varpi_j^{-1}\left(\tau_j^D\right)^2}{\sum_{k=1}^{d_z} \varpi_k^{-1}\left(\tau_k^D\right)^2}, \qquad \bar{m}_0 = \tau^Y - \tau^D\theta_0, \tag{61}$$

with $\bar{m}_{0,j} = \tau_j^D\left(\theta^{(j)} - \theta_0\right)$. Because Ω_0 is diagonal and positive, the weights in (61) are *non-negative and sum to one*: the pseudo-true value is a convex combination of the instrument-specific LATEs, an average treatment effect for a mixture of the complier groups, with instrument j receiving weight proportional to its complier share when $\varpi_j \propto \tau_j^D$. Chow and Kasahara (2026) also characterize the estimand of over-identified IV-GMM as a combination of the instrument-specific Wald estimands whose weights sum to one, and show that the diagonality of the weighting matrix keeps every weight non-negative. For a general weighting matrix, a weight can be negative, which is why we do not consider the efficient weight or the two-stage least squares weight; with an additively separable specification in the covariates X, it need not be a convex combination of covariate-specific effects either (Angrist and Imbens, 1995; Mogstad et al., 2021; Sloczynski, 2021; Blandhol et al., 2022). In our setup, the scores (57) are already nonparametric as the covariates X enter only through the conditional expectations (55), and can be estimated by machine learning with cross-fitting.[9]

Substituting (58)–(60) into (32), the misspecification-robust score is

$$\psi^*(w,\theta_0,\eta_0) = -\tau^{D\prime}\Omega_0^{-1}\psi(w,\theta_0,\beta_0,\alpha_0) - \left(\psi^D(w,\beta_0,\alpha_0) - \tau^D\right)'\Omega_0^{-1}\bar{m}_0, \tag{62}$$

with $J_0 = G_0'\Omega_0^{-1}G_0 = \tau^{D\prime}\Omega_0^{-1}\tau^D$ and the Jacobian term in (62) remains whenever $\bar{m}_0 \neq 0$ with known weights. The standard Neyman orthogonal score of ψ is

$$G_0'\Omega_0^{-1}\psi(w,\theta_0,\beta_0,\alpha_0) = -\tau^{D\prime}\Omega_0^{-1}\psi(w,\theta_0,\beta_0,\alpha_0) \in \mathbb{R}.$$

The DML estimator of LATEs takes the following explicit form. For each $k \in [K]$ and each instrument j, let $\widehat{\ell}_{j,k}$, $\widehat{\nu}_{j,k}$ and $\widehat{p}_{j,k}$ be machine-learning or nonparametric estimators of ℓ_{0j}, ν_{0j} and

[9] The diagonal of the two-stage least squares weight, $\Omega_0 = \mathrm{diag}\left(\mathrm{E}_P[\psi_j^D(W,\beta_0,\alpha_0)^2]\right)$ can be considered here. The weights ω_j remain positive; however, Assumption 1(d) fails and the correction of Remark 3 is required. See Appendix B for general results under efficient weights.

p_{0j} based on $(W_i)_{i\in I_k^c}$, and for $i \in I_k$ form the cross-fitted scores

$$\widehat{\psi}_i^Y := \psi^Y\big(W_i, \widehat{\beta}_k, \widehat{\alpha}_k\big), \qquad \widehat{\psi}_i^D := \psi^D\big(W_i, \widehat{\beta}_k, \widehat{\alpha}_k\big), \qquad i \in I_k,$$

with $\widehat{\beta}_k = (\widehat{\ell}_{j,k}, \widehat{\nu}_{j,k})_j$ and $\widehat{\alpha}_k = (\widehat{p}_{j,k})_j$. The cross-fitted Jacobian and moment of the DML estimator are the fold averages

$$\widehat{\tau}^Y := \frac{1}{N}\sum_{k=1}^K \sum_{i\in I_k} \widehat{\psi}_i^Y, \qquad \widehat{\tau}^D := \frac{1}{N}\sum_{k=1}^K \sum_{i\in I_k} \widehat{\psi}_i^D, \qquad \widehat{\Gamma} = -\widehat{\tau}^D, \qquad \widehat{\Omega} = \Omega_0,$$

and we have

$$\check{\theta} = \frac{\widehat{\tau}^{D\prime}\Omega_0^{-1}\widehat{\tau}^Y}{\widehat{\tau}^{D\prime}\Omega_0^{-1}\widehat{\tau}^D} = \sum_{j=1}^{d_z} \widehat{\omega}_j\, \widehat{\theta}^{(j)}, \qquad \widehat{\theta}^{(j)} = \frac{\widehat{\tau}_j^Y}{\widehat{\tau}_j^D}, \tag{63}$$

where $\widehat{\omega}_j$ is ω_j of (61) with $\widehat{\tau}^D$ in place of τ^D. The estimated scores and Jacobian are

$$\begin{aligned} \widehat{\psi}_i^* &= -\widehat{\tau}^{D\prime}\Omega_0^{-1}\big(\widehat{\psi}_i^Y - \check{\theta}\,\widehat{\psi}_i^D\big) - \big(\widehat{\psi}_i^D - \widehat{\tau}^D\big)'\Omega_0^{-1}\widehat{\bar{m}}, \qquad i \in I_k, \\ \widehat{J}_0 &= \widehat{\tau}^{D\prime}\Omega_0^{-1}\widehat{\tau}^D, \qquad \widehat{\sigma}^2 = \widehat{J}_0^{-1}\,\frac{1}{N}\sum_{k=1}^K \sum_{i\in I_k} \big(\widehat{\psi}_i^*\big)^2 \widehat{J}_0^{-1\prime}, \end{aligned} \tag{64}$$

where $\widehat{\bar{m}} := \frac{1}{N}\sum_{k=1}^K \sum_{i\in I_k}(\widehat{\psi}_i^Y - \check{\theta}\widehat{\psi}_i^D)$ is the cross-fitted moment.

For each j, $\widehat{\theta}^{(j)}$ in (63) is the single-instrument DML estimator of Chernozhukov et al. (2018, Theorem 5.2), and the point estimate $\check{\theta}$ is a convex combination of these, with non-negative weights $\widehat{\omega}_j$.

5.1 Asymptotic distribution

Assumption 5 (Regularity conditions for the LATE)**.** *Let $q > 2$, $\epsilon \in (0, 1/2)$ and $0 < c_0 \le c_1$, $C < \infty$ be constants and $\{\delta_N\}, \{\Delta_N\}$ sequences of positive constants converging to zero with $\delta_N \ge N^{-[(1-2/q)\wedge 1/2]}$. For all $P \in \mathcal{P}_N$, the following conditions hold.*

(a) $D \in \{0,1\}$, $Z \in \{0,1\}^{d_z}$, $\|Y\|_{P,q} \le C$, and $\epsilon \le p_{0j}(X) \le 1-\epsilon$ almost surely for every j.

(b) (β_0, α_0) is given by (55) and $\|\mathrm{E}_P[Y^2 \mid Z, X]\|_{P,\infty} \le C$.

(c) For every j, τ_j^D and ϖ_j lie in $[c_0, c_1]$; and $\mathrm{E}_P\big[\psi^*(W, \theta_0, \eta_0)^2\big] \ge c_0$.

(d) θ_0 in (61) lies in the interior of Θ, where $\Theta \subset \mathbb{R}$ is a compact interval.

(e) With P-probability at least $1 - \Delta_N$, the cross-fitted estimators satisfy $(\widehat{\beta}_k, \widehat{\alpha}_k) \in \mathcal{B}_N \times \mathcal{A}_N$ for all k, where all $\beta = (\ell_j, \nu_j)_j \in \mathcal{B}_N$ and $\alpha = (p_j)_j \in \mathcal{A}_N$ satisfy, for all j and $z \in \{0,1\}$,

$$\epsilon \le p_j(X) \le 1-\epsilon \quad \text{and} \quad \nu_j(z,X) \in [0,1] \ \ a.s., \qquad \|\ell_j(z,\cdot) - \ell_{0j}(z,\cdot)\|_{P,q} \le C,$$

$$\|\ell_j(z,\cdot) - \ell_{0j}(z,\cdot)\|_{P,2} \vee \|\nu_j(z,\cdot) - \nu_{0j}(z,\cdot)\|_{P,2} \vee \|p_j - p_{0j}\|_{P,2} \le \delta_N,$$

and

$$\|p_j - p_{0j}\|_{P,2}\Big(\sum_{z\in\{0,1\}}\|\ell_j(z,\cdot)-\ell_{0j}(z,\cdot)\|_{P,2}+\sum_{z\in\{0,1\}}\|\nu_j(z,\cdot)-\nu_{0j}(z,\cdot)\|_{P,2}\Big)\le\delta_N N^{-1/2}. \tag{65}$$

Assumption 5 is Chernozhukov et al. (2018, Assumption 5.2) imposed on each instrument. With $d_z > 1$, the system is over-identified and θ_0 is the pseudo-true value (61) rather than a common LATE. Accordingly, the non-degeneracy in Assumption 5(c) is imposed on the misspecification-robust score (62) rather than the conventional score ψ. The rate condition (65) is the same as the product rate of Chernozhukov et al. (2018, Assumption 5.2(g)). The additional terms from the Jacobian and weight under misspecification are products of two different first-step errors, so the same product rate condition (65) suffices.

Theorem 4 (DML inference on a convex combination of LATEs)**.** *Suppose Assumption 5 holds. Then the DML estimator* $\check{\theta}$ *of* (63) *satisfies*

$$\sqrt{N}\,\sigma^{-1}\big(\check{\theta}-\theta_0\big)\ \xrightarrow{d}\ \mathcal{N}(0,1),\qquad \sigma^2=\frac{\mathrm{E}_P\big[\psi^*(W,\theta_0,\eta_0)^2\big]}{\big(\tau^{D\prime}\Omega_0^{-1}\tau^D\big)^2},$$

uniformly over $P\in\mathcal{P}_N$*. Moreover* $\widehat{\sigma}^2$ *of* (64) *satisfies* $\widehat{\sigma}^2=\sigma^2+o_P(1)$*, and, for a nominal level* $a\in(0,1)$*, the confidence interval* $\mathrm{CI}_N=[\,\check{\theta}\pm\Phi^{-1}(1-a/2)\,\widehat{\sigma}/\sqrt{N}\,]$ *is uniformly asymptotically valid for* θ_0*,*

$$\sup_{P\in\mathcal{P}_N}\Big|\,\mathrm{Pr}_P\left(\theta_0\in\mathrm{CI}_N\right)-(1-a)\Big|\longrightarrow 0.$$

Under the over-identifying restriction $\theta^{(1)}=\cdots=\theta^{(d_z)}$, we have $\bar{m}_0=0$, the second term of (62) drops out, and Theorem 4 reduces to a pooled version of Chernozhukov et al. (2018, Theorem 5.2), where $\check{\theta}$ still estimates the interpretable causal parameter (61) at rate $\sqrt{N}$ with valid inference.

6 Simulations

This section illustrates the main results and the DML inference in the partially linear IV model of Section 4 using Monte Carlo simulations. Our simulation designs closely follow the linear instrumental variables model of Chernozhukov et al. (2015a),

$$Y=\theta^*D+X'\theta_y+Z'\delta+\varepsilon,\qquad D=X'\theta_d+Z'\gamma+V, \tag{66}$$

where θ^* is the structural coefficient on the endogenous regressor D, X is the vector of controls, Z is the vector of instruments, and γ is the first stage. Here, some instruments are allowed to have a direct effect δ on the outcome Y. The instruments are valid when $\delta=0$, in which case (66) is the partially linear IV model (42) with $g_0(X)=X'\theta_y$.

The controls, instruments and errors are drawn as follows:

$$X \sim \mathcal{N}(0, \Sigma_x), \quad \Sigma_{x,kj} = 0.5^{|j-k|}, \qquad Z = (X_1, \ldots, X_{d_z})' + \xi, \quad \xi \sim \mathcal{N}\big(0, 0.25\, I_{d_z}\big),$$
$$(\varepsilon, V)' \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right),$$

with $X \in \mathbb{R}^{p_x}$, $p_x = 200$, $d_z = 5$, and ξ, (ε, V) and X mutually independent. We set the structural coefficient as $\theta^* = 1$, control coefficients $\theta_y = \theta_d = 1/j^2$ for $j = 1, \ldots, p_x$, and the first stage as $\gamma_j = 1/j$.

For Design A (invalid instruments) in (66), the GMM estimand is $\theta_0 = \theta^* + \gamma'\delta/\gamma'\gamma$. We report $\delta = 0$, the correctly specified benchmark, and $\delta = (0,\, 0.6,\, -0.6,\, 0,\, 0)'$, for which some instruments are invalid. We also consider Design B (heterogeneous effects with valid instruments). This design keeps $\delta = 0$ in (66), so all instruments are exogenous and excluded. Instead, we let $G \in \{A, B\}$ be a latent type with $\Pr(G = B) = 1/2$, drawn independently of (X, ξ, ε, V), and replace θ^* and γ in (66) by θ_G and γ_G, with $\theta_A = \theta^* - \tau/2$, $\theta_B = \theta^* + \tau/2$ and $\gamma_A = 0.6\,(1, 1, 1, 0, 0)'$, $\gamma_B = 0.6\,(0, 0, 0, 1, 1)'$. We report $\tau = 0$, the correctly specified benchmark, which imposes a homogeneous treatment effect, $\theta_A = \theta_B$, and $\tau = 2$, the heterogeneous case, for which $\theta_A = 0$ and $\theta_B = 2$. Each instrument identifies the effect for the subpopulation it moves, a local average treatment effect, with $\theta^{(j)} = \theta_A$ for $j \leq 3$ and $\theta^{(j)} = \theta_B$ for $j \geq 4$. The pseudo-true value is the convex combination $\theta_0 = (3\theta_A + 2\theta_B)/5 = 0.8$, and the over-identifying restrictions fail because those subpopulations differ, not because any instrument is invalid.

6.1 Results

In both designs the DML2 estimator (52) is the same cross-fitted two-stage least squares estimator, with the cross-fitted residuals $(\widehat{\tilde{Y}}_i, \widehat{\tilde{D}}_i, \widehat{\tilde{Z}}_i)$. We report the two different standard errors $\widehat{\sigma}/\sqrt{N}$ with $\widehat{\sigma}^2$ from (53); the misspecification-robust standard error uses all three terms of $\widehat{\psi}_i^*$, and the conventional one uses the first term alone. The nuisance parameters are the conditional expectations (43) that are estimated by the post-LASSO estimator of Belloni and Chernozhukov (2013) with the data-dependent penalty of Belloni et al. (2012) and $K = 5$ folds. We report 2000 replications with sample sizes $N \in \{500, 1000, 2000\}$ and nominal level 95%.[10]

For each design and sample size, Table 1 reports the median absolute deviation of $\check{\theta}$ from θ_0, the standard deviation of $\check{\theta}$, the average of each standard error, and the coverage of the corresponding 95% confidence interval for θ_0.

First of all, the robust standard errors based on the misspecification-robust orthogonal scores work well even under correct specification. In both designs, the two standard errors and the coverage frequencies are almost the similar. Second, under misspecification, the robust standard

[10] In an unreported simulation exercise, we also compare the parametric 2SLS estimator, either with a naive choice of controls among the $p_x = 200$ candidates or without controls, with the 2SLS estimator with the first-step LASSO considered here. The parametric estimator is inconsistent for θ_0, and its coverage is essentially zero even with the misspecification-robust standard errors.

Table 1: Monte Carlo results for the partially linear IV model.

			average s.e.		coverage	
N	MAD	SD($\check{\theta}$)	conventional	robust	conventional	robust
Design A, $\delta = 0$ *(correctly specified),* $\theta_0 = 1$						
500	0.0517	0.0759	0.0744	0.0752	0.927	0.928
1000	0.0366	0.0530	0.0525	0.0528	0.946	0.947
2000	0.0263	0.0381	0.0370	0.0371	0.942	0.942
Design A, $\delta \neq 0$ *(invalid instruments),* $\theta_0 = 1.068$						
500	0.0656	0.0959	0.0786	0.0946	0.885	0.936
1000	0.0457	0.0667	0.0551	0.0664	0.890	0.952
2000	0.0323	0.0476	0.0389	0.0467	0.893	0.945
Design B, $\tau = 0$ *(correctly specified),* $\theta_0 = 1$						
500	0.0910	0.1342	0.1302	0.1382	0.924	0.933
1000	0.0660	0.0949	0.0931	0.0959	0.938	0.948
2000	0.0456	0.0667	0.0664	0.0674	0.949	0.953
Design B, $\tau = 2$ *(heterogeneous effects),* $\theta_0 = 0.8$						
500	0.2631	0.3847	0.3416	0.4033	0.938	0.962
1000	0.1837	0.2742	0.2425	0.2812	0.921	0.957
2000	0.1367	0.2021	0.1724	0.1980	0.909	0.949

Notes. For each design and sample size, we report the median of $|\check{\theta} - \theta_0|$ (MAD), and the standard deviation of the cross-fitted 2SLS estimator. We consider the two standard errors of (53), and report average SE across simulations, and the coverage of the confidence intervals. 2000 replications, $K = 5$ folds, nominal level 95%. The first-step estimation of the nuisance parameters is based on the post-LASSO.

error is about 15–20% larger than the conventional one. However, the conventional standard error is much smaller than the standard deviation of the DML estimator $\check{\theta}$ while the robust one matches it very closely. Consequently, the conventional confidence interval undercovers at all sample sizes: in Design A, at 0.885, 0.890 and 0.893 against the robust one 0.936, 0.952 and 0.945; in Design B, at 0.938, 0.921 and 0.909, while the robust coverage stays close to the nominal level, 0.962, 0.957 and 0.949.

Figure 1 plots the density of the studentized statistic $(\check{\theta} - \theta_0)/\widehat{\text{se}}$ against the standard normal, separately for the two standard errors (sample size $N = 2000$). In both designs, the t-statistics with robust standard error closely follow the reference normal density, while the conventional one is more dispersed. This is in line with the coverage results in Table 1.

7 Conclusion

This paper develops Neyman orthogonal scores for GMM that remain valid when the moment conditions fail to hold. Misspecification-robust orthogonal score has a closed form; the conventional orthogonal score plus adjustment terms for the Jacobian and weighting matrix, both proportional to the population moment at the GMM estimand and therefore negligible under correct specification.

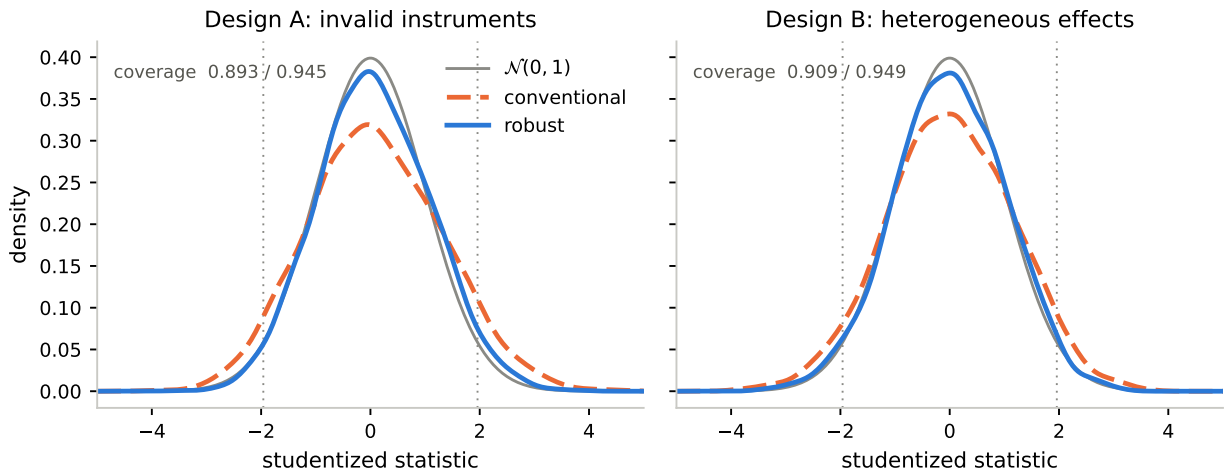


Figure 1: Comparison of conventional and robust standard errors for the DML estimator. The density of the studentized statistic $(\check{\theta} - \theta_0)/\widehat{\text{se}}$ at $N = 2000$ under misspecification, over 2000 replications. Left: Design A, invalid instruments. Right: Design B, heterogeneous effects. The dashed line: the conventional standard error; solid line: the misspecification-robust standard error of Theorem 3.

The construction extends to semiparametric moment condition models, where the moment functions depend on nonparametric or machine learning first steps, and the DML estimator with cross-fitting based on the proposed score provides uniformly valid inference for the GMM estimand. For linear scores with a one-step weighting matrix, the point estimate is unchanged, and only the standard error has to be recomputed. In the over-identified partially linear IV model, the estimator is cross-fitted two-stage least squares on the residualized data, and the robust standard error uses only the same conditional expectations and sample moments of the residuals; in the simulations the robust standard error is 15–20% larger than the conventional one, and the conventional confidence interval undercovers. In the local average treatment effects model with several instruments, over-identification comes from treatment-effect heterogeneity rather than from invalid instruments, a diagonal weighting matrix makes the estimand a convex combination of the instrument-specific LATEs, and the robust variance makes a confidence interval uniformly valid.

References

Ackerberg, D., X. Chen, J. Hahn, and Z. Liao (2014). Asymptotic efficiency of semiparametric two-step GMM. *Review of Economic Studies*, 81(3), 919–943.

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

Ai, C. and X. Chen (2007). Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, 141(1), 5–43.

Ai, C. and X. Chen (2012). The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, 170(2), 442–457.

Andrews, I., N. Barahona, M. Gentzkow, A. Rambachan, and J. M. Shapiro (2025a). Structural estimation under misspecification: Theory and implications for practice. *Quarterly Journal of Economics*, 140(3), 1801–1855.

Andrews, I., J. Chen, and O. Tecchio (2025b). The purpose of an estimator is what it does: Misspecification, estimands, and over-identification. Forthcoming, 2025 ESWC Monograph, *arXiv:2508.13076*.

Angrist, J. D. and G. W. Imbens (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430), 431–442.

Armstrong, T. B. and M. Kolesár (2021). Sensitivity analysis using approximate moment condition models. *Quantitative Economics*, 12(1), 77–108.

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

Belloni, A. and V. Chernozhukov (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2), 521–547.

Blandhol, C., J. Bonney, M. Mogstad, and A. Torgovitsky (2022). When is TSLS actually LATE? *Review of Economic Studies*, forthcoming.

Bonhomme, S., K. Jochmans, W. K. Newey, and M. Weidner (2026). Higher-order Neyman orthogonality in moment-condition models. *arXiv:2605.10842*.

Bonhomme, S. and M. Weidner (2022). Minimizing sensitivity to model misspecification. *Quantitative Economics*, 13(3), 907–954.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3), 305–334.

Chamberlain, G. and G. W. Imbens (1995). Semiparametric applications of Bayesian inference. Harvard Institute of Economic Research Working Paper.

Chen, X. and A. Santos (2018). Overidentification in regular models. *Econometrica*, 86(5), 1771–1817.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1), C1–C68.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2022). Locally robust semiparametric estimation. *Econometrica*, 90(4), 1501–1535.

Chernozhukov, V., C. Hansen, and M. Spindler (2015). Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review: Papers & Proceedings*, 105(5), 486–490.

Chernozhukov, V., C. Hansen, and M. Spindler (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics*, 7, 649–688.

Chernozhukov, V., W. K. Newey, and R. Singh (2022b). Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3), 967–1027.

Chetverikov, D., J. R.-V. Sørensen, and A. Tsyvinski (2026). Triple/double-debiased Lasso. Working paper.

Chow, C. P. and H. Kasahara (2026). When is GMM actually LATE? Weighting matrices and causal interpretation in overidentified IV. *arXiv:2604.07131*.

Frölich, M. (2007). Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics*, 139(1), 35–75.

Hall, A. R. and A. Inoue (2003). The large sample behaviour of the generalized method of moments estimator in misspecified models. *Journal of Econometrics*, 114(2), 361–394.

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

Hansen, L. P., J. Heaton, and A. Yaron (1996). Finite-sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics*, 14(3), 262–280.

Hansen, B. E. and S. Lee (2021). Inference for iterated GMM under misspecification. *Econometrica*, 89(3), 1419–1447.

Hwang, J., B. Kang, and S. Lee (2022). A doubly corrected robust variance estimator for linear GMM. *Journal of Econometrics*, 229(2), 276–298.

Ichimura, H. and W. K. Newey (2022). The influence function of semiparametric estimators. *Quantitative Economics*, 13(1), 29–61.

Imbens, G. W. (1997). One-step estimators for over-identified generalized method of moments models. *Review of Economic Studies*, 64(3), 359–383.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2), 467–476.

Kitamura, Y., T. Otsu, and K. Evdokimov (2013). Robustness, infinitesimal neighborhoods, and moment restrictions. *Econometrica*, 81(3), 1185–1201.

Lee, S. (2018). A consistent variance estimator for 2SLS when instruments identify different LATEs. *Journal of Business & Economic Statistics*, 36(3), 400–410.

Mogstad, M., A. Torgovitsky, and C. R. Walters (2021). The causal interpretation of two-stage least squares with multiple instrumental variables. *American Economic Review*, 111(11), 3663–3698.

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

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. In R. F. Engle and D. L. McFadden (Eds.), *Handbook of Econometrics*, Volume 4, 2111–2245. Amsterdam: North-Holland.

Newey, W. K. and J. R. Robins (2018). Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv:1801.09138*.

Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. In U. Grenander (Ed.), *Probability and Statistics: The Harald Cramér Volume*, 213–234. New York: Wiley.

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica*, 56(4), 931–954.

Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *Annals of Statistics*, 35(2), 634–672.

Słoczyński, T. (2021). When should we (not) interpret linear IV estimands as LATE? *arXiv:2011.06695*.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1), 1–25.

A Proofs

A.1 Proof of Lemma 1

Proof. The argument is the same as that of Chernozhukov et al. (2018, Lemma 2.3) and we provide it here for completeness. Since the nuisance parameters $\eta = (\beta', \mathrm{vec}(\mu)')'$ are finite-dimensional, it suffices to show that the partial derivatives of $\eta \mapsto \mathrm{E}_P[\psi(W, \theta_0, \eta)]$ vanish at η_0. We note that $\mathrm{E}_P[\psi(W, \theta_0, \beta, \mu_0)] = \mu_0 \, \bar{m}(\gamma)$ with $\gamma = (\theta_0', \beta')'$, so

$$\partial_{\beta'} \mathrm{E}_P[\psi(W, \theta_0, \eta_0)] = \mu_0 G_{\beta,0} = A'\Omega^{-1}G_{\beta,0} - A'\Omega^{-1}G_{\beta,0}\big(G_{\beta,0}'\Omega^{-1}G_{\beta,0}\big)^{-1}G_{\beta,0}'\Omega^{-1}G_{\beta,0} = 0$$

by (5). Further, $\partial_{\mathrm{vec}(\mu)'}\mathrm{E}_P[\psi(W,\theta_0,\eta_0)] = \mathrm{E}_P[m(W,\theta_0,\beta_0)]' \otimes I_{d_\theta} = 0$, and this gives (7). Since the derivatives are exact zeros for arbitrary finite-dimensional deviations, orthogonality holds with the unrestricted realization set $T_N = T$. □

A.2 Proof of Proposition 1

Proof. The just-identified stacked system (9)–(10) is a GMM problem of exactly the form considered in Section 2.1 with a target θ, a finite-dimensional nuisance λ, and a known moment function $h(w,\theta,\lambda)$ whose moment conditions hold at the (pseudo-)true parameter values. The stacked system is always just-identified rather than possibly over-identified for the original moment function m, which Lemma 1 also allows.

First, $\mathrm{E}_P[\psi^*(W,\theta_0,\eta_0)] = \mu_0\mathrm{E}_P[h(W,\theta_0,\lambda_0)] = 0$ by (10). The argument is similar to that of Lemma 1, with the finite-dimensional nuisance $\eta = (\lambda', \mathrm{vec}(\mu)')'$. For the λ-block, $\mathrm{E}_P[\psi^*(W,\theta_0,\lambda,\mu_0)] = \mu_0\mathrm{E}_P[h(W,\theta_0,\lambda)]$, so $\partial_{\lambda'}\mathrm{E}_P[\psi^*(W,\theta_0,\eta_0)] = \mu_0 J_\lambda$, and by (12)

$$\mu_0 J_\lambda = A_h'\Omega_h^{-1}J_\lambda - A_h'\Omega_h^{-1}J_\lambda\big(J_\lambda'\Omega_h^{-1}J_\lambda\big)^{-1}J_\lambda'\Omega_h^{-1}J_\lambda = 0,$$

where $J_\lambda'\Omega_h^{-1}J_\lambda$ is nonsingular because J_λ has full column rank and Ω_h is positive definite. For the μ-block, $\mu \mapsto \mu h$ is linear, so $\partial_{\mathrm{vec}(\mu)'}\mathrm{E}_P[\psi^*(W,\theta_0,\eta_0)] = \mathrm{E}_P[h(W,\theta_0,\lambda_0)]' \otimes I_{d_\theta} = 0$ by (10). These give (13) and complete the proof. □

A.3 Proof of Proposition 2

Proof. By using the identity $\mathrm{vec}(ABC) = (C' \otimes A)\,\mathrm{vec}(B)$, the first-order condition (8) has the following representations

$$Q(\gamma) = \big(I_{d_\gamma} \otimes \bar m(\gamma)'\Omega(\gamma)^{-1}\big)\,\mathrm{vec}\,G(\gamma) = \big(\bar m(\gamma)' \otimes G(\gamma)'\big)\,\mathrm{vec}\,\big(\Omega(\gamma)^{-1}\big).$$

Then, using the chain rule and $\partial\,\mathrm{vec}(\Omega^{-1}) = -(\Omega^{-1}\otimes\Omega^{-1})\,\partial\,\mathrm{vec}\,\Omega$ gives

$$H = \partial_{\gamma'}Q(\gamma)\big|_{\gamma=\gamma_0} = G_0'\Omega_0^{-1}G_0 + \big(I_{d_\gamma}\otimes \bar m_0'\Omega_0^{-1}\big)\dot G - \big(\bar m_0'\Omega_0^{-1}\otimes G_0'\Omega_0^{-1}\big)\dot\Omega, \tag{67}$$

where $\bar m_0 = \bar m(\gamma_0)$, $\dot G := \partial_{\gamma'}\,\mathrm{vec}\,G(\gamma)|_{\gamma_0} \in \mathbb{R}^{d_m d_\gamma \times d_\gamma}$ and $\dot\Omega := \partial_{\gamma'}\,\mathrm{vec}\,\Omega(\gamma)|_{\gamma_0} \in \mathbb{R}^{d_m^2\times d_\gamma}$.

Since $\mathrm{E}_P[h(W,\theta,\lambda)] = \big(\mathrm{vec}(\Gamma - G(\gamma))',\ \mathrm{vec}(\Omega-\Omega(\gamma))',\ (\Gamma'\Omega^{-1}\bar m(\gamma))'\big)'$, the partial derivatives with respect to $\big(\gamma',\ \mathrm{vec}(\Gamma)',\ \mathrm{vec}(\Omega)'\big)'$ give

$$J = \mathrm{E}_P\big[\partial_{(\theta',\lambda')}h(W,\theta_0,\lambda_0)\big] = \big[\,J_\theta\ \ J_\lambda\,\big] = \begin{pmatrix} -\dot G & I_{d_m d_\gamma} & 0 \\ -\dot\Omega & 0 & I_{d_m^2} \\ G_0'\Omega_0^{-1}G_0 & I_{d_\gamma}\otimes\bar m_0'\Omega_0^{-1} & -\big(\bar m_0'\Omega_0^{-1}\otimes G_0'\Omega_0^{-1}\big) \end{pmatrix}. \tag{68}$$

Next, we set

$$M := \Big[-\big(I_{d_\gamma} \otimes \bar{m}_0' \Omega_0^{-1}\big) \quad \bar{m}_0' \Omega_0^{-1} \otimes G_0' \Omega_0^{-1} \quad I_{d_\gamma} \Big], \qquad P := \big[I_{d_\theta} \;\; -H_{\theta\beta} H_{\beta\beta}^{-1} \big], \qquad \mu_0^\star := PM. \tag{69}$$

By (67) and (68), $MJ = [\, H \;\; 0 \;\; 0\,]$. Let $H = [H_\theta \quad H_\beta]$. Then, $\mu_0^\star J_\theta = PH_\theta = H_{\theta\theta} - H_{\theta\beta} H_{\beta\beta}^{-1} H_{\beta\theta} = J_0$ and $\mu_0^\star J_\lambda = [\, PH_\beta \;\; 0 \;\; 0\,] = 0$, since $PH_\beta = H_{\theta\beta} - H_{\theta\beta} H_{\beta\beta}^{-1} H_{\beta\beta} = 0$.

Now let μ_0 be given by (12), so that $\mu_0 J_\lambda = 0$. With $\mu_0 J_\theta$ nonsingular, $(\mu_0 J_\theta)^{-1} \mu_0 J = (\mu_0 J_\theta)^{-1} \mu_0 [J_\theta \; J_\lambda] = [\, I_{d_\theta} \;\; 0\,] = J_0^{-1} \mu_0^\star J$, and since J is nonsingular,

$$(\mu_0 J_\theta)^{-1} \mu_0 = J_0^{-1} \mu_0^\star, \tag{70}$$

where both sides are the first d_θ rows of J^{-1}. Note that $\mu_0^\star$ is itself of the form (12) with $\Omega_h = I_{d_h}$ and $A_h = \mu_0^{\star\prime}$, because $\mu_0^\star \big(I_{d_h} - J_\lambda (J_\lambda' J_\lambda)^{-1} J_\lambda'\big) = \mu_0^\star$ by $\mu_0^\star J_\lambda = 0$. For this choice $\mu_0 = \mu_0^\star$ and $\mu_0 J_\theta = J_0$, and the score is $s_\theta(w) - H_{\theta\beta} H_{\beta\beta}^{-1} s_\beta(w)$ itself.

By the definition of M in (69), and $h(w, \theta, \lambda)$ evaluated at (θ_0, λ_0), we have

$$M\, h(w, \theta_0, \lambda_0) = G_0' \Omega_0^{-1} m(w, \gamma_0) + \big(\partial_{\gamma'} m(w, \gamma_0) - G_0\big)' \Omega_0^{-1} \bar{m}_0 - G_0' \Omega_0^{-1} \big(\zeta(w, \gamma_0) - \Omega_0\big) \Omega_0^{-1} \bar{m}_0 = s(w),$$

so $\mu_0^\star h(w, \theta_0, \lambda_0) = P\, s(w) = s_\theta(w) - H_{\theta\beta} H_{\beta\beta}^{-1} s_\beta(w)$. Further, $\mathrm{E}_P[s(W)] = M\, \mathrm{E}_P[h(W, \theta_0, \lambda_0)] = 0$ by (10). By (70), $\psi^*(w, \theta_0, \eta_0) = \mu_0 h(w, \theta_0, \lambda_0) = (\mu_0 J_\theta) J_0^{-1} \mu_0^\star h(w, \theta_0, \lambda_0)$, which gives (14).

The stacked estimator solves the sample just-identified system $N^{-1} \sum_{i=1}^N h(W_i, \widehat{\theta}, \widehat{\lambda}) = 0$ (Remark 1), so under standard regularity conditions (Newey and McFadden, 1994, Theorem 3.4) $\sqrt{N}((\widehat{\theta}', \widehat{\lambda}')' - (\theta_0', \lambda_0')') = -J^{-1} N^{-1/2} \sum_{i=1}^N h(W_i, \theta_0, \lambda_0) + o_P(1)$. By (70), its first d_θ rows are $-(\mu_0 J_\theta)^{-1} N^{-1/2} \sum_{i=1}^N \psi^*(W_i, \theta_0, \eta_0) + o_P(1) = -J_0^{-1} N^{-1/2} \sum_{i=1}^N \big(s_\theta(W_i) - H_{\theta\beta} H_{\beta\beta}^{-1} s_\beta(W_i)\big) + o_P(1)$, which is (16) and it does not depend on the choice of (A_h, Ω_h). □

A.4 Proof of Theorem 1

Proof. The stacked equation (25) is (9) with γ replaced by θ, m replaced by $\psi(\cdot, \cdot, \beta, \alpha)$ and no finite-dimensional β in λ for any $\beta \in \mathcal{B}$ and $\alpha \in \mathcal{A}$. Hence the calculations in the proof of Proposition 2 apply with d_γ replaced by d_θ, $P = I_{d_\theta}$, and $\bar{m}$, G, Ω and Q in (21) and (22), where $\mathrm{E}_P[\psi(W, \theta, \beta_0, \alpha_0)] = \bar{m}(\theta)$ holds by Assumption 1(c). In particular, J is given by (68), so that J_λ has full column rank and μ_0 of the form (12) is well defined. $J_0 = \partial_{\theta'} Q(\theta_0)$ is similarly calculated as in (67), which is (33). Further $\mu_0^\star = M$ of (69) satisfies $MJ = [\, J_0 \;\; 0\,]$; and, as in the proof of Proposition 2, (70) gives $\mu_0 = \mu_0 J_\theta J_0^{-1} M$ for any μ_0 of the form (12) with $\mu_0 J_\theta$ nonsingular. The normalization $\mu_0 J_\theta = J_0$ yields $\mu_0 = M$; M itself is of the form (12) with $A_h = M'$ and $\Omega_h = I_{d_h}$, and more generally, since (12) is linear in A_h', replacing A_h' by $J_0 (\mu_0 J_\theta)^{-1} A_h'$ premultiplies μ_0 by $J_0 (\mu_0 J_\theta)^{-1}$. For any other choice of (A_h, Ω_h), ψ^* is premultiplied by the nonsingular matrix $\mu_0 J_\theta J_0^{-1}$, which changes neither the solution of the estimating equation nor the influence function $-(\mu_0 J_\theta)^{-1} \psi^*$. Moreover, a calculation similar to that in Appendix A.3 gives, for all $\theta \in \Theta$, $\beta \in \mathcal{B}$

and $\alpha \in \mathcal{A}$,

$$\begin{aligned} M\, h(w,\theta,\lambda_0,\beta,\alpha) = G_0'\Omega_0^{-1}\psi(w,\theta,\beta,\alpha) + \big(\partial_{\theta'}\psi(w,\theta,\beta,\alpha) - G_0\big)'\Omega_0^{-1}\bar{m}_0 \\ - G_0'\Omega_0^{-1}\big(\zeta(w,\theta,\beta,\alpha) - \Omega_0\big)\Omega_0^{-1}\bar{m}_0. \end{aligned} \tag{71}$$

Setting $(\beta,\alpha) = (\beta_0,\alpha_0)$ gives (32), and $\partial_{\theta'}\mathrm{E}_P[\psi^*(W,\theta,\eta_0)]|_{\theta_0} = \mu_0 J_\theta = J_0$. Further, (31) follows from $\mathrm{E}_P[\psi^*(W,\theta_0,\eta_0)] = \mu_0\mathrm{E}_P[h(W,\theta_0,\lambda_0,\beta_0,\alpha_0)] = 0$ by (26).

For orthogonality, the derivatives with respect to the finite-dimensional components (λ,μ) are $[\,\mu_0 J_\lambda,\ \mathrm{E}_P[h(W,\theta_0,\lambda_0,\beta_0,\alpha_0)]' \otimes I_{d_\theta}\,] = 0$ exactly as in the proof of Proposition 1, by (12) and (26). For (β,α), write $D_f := \partial_{(\beta,\alpha)}\mathrm{E}_P[f(W,\theta_0,\beta_0,\alpha_0)][\beta-\beta_0,\alpha-\alpha_0]$ for $(\beta,\alpha) \in \mathcal{B}_N \times \mathcal{A}_N$, which exists by Assumption 1(b). By (71),

$$\partial_{(\beta,\alpha)}\mathrm{E}_P\big[\psi^*(W,\theta_0,\lambda_0,\mu_0,\beta_0,\alpha_0)\big][\beta-\beta_0,\alpha-\alpha_0] = G_0'\Omega_0^{-1}D_\psi + D'_{\partial_{\theta'}\psi}\,\Omega_0^{-1}\bar{m}_0 - G_0'\Omega_0^{-1}D_\zeta\,\Omega_0^{-1}\bar{m}_0.$$

Here $D_\psi = 0$ by Assumption 1(c) at θ_0, and $D_\zeta = 0$ by Assumption 1(d). For the Jacobian channel, the interchange of derivatives in Assumption 1(b) states that, for all $(\beta,\alpha) \in \mathcal{B}_N \times \mathcal{A}_N$,

$$\begin{aligned} &\partial_{(\beta,\alpha)}\mathrm{E}_P\big[\partial_{\theta'}\psi(W,\theta_0,\beta_0,\alpha_0)\big][\beta-\beta_0,\alpha-\alpha_0] \\ &\quad = \partial_{\theta'}\Big(\partial_{(\beta,\alpha)}\mathrm{E}_P\big[\psi(W,\theta,\beta_0,\alpha_0)\big][\beta-\beta_0,\alpha-\alpha_0]\Big)\Big|_{\theta=\theta_0}. \end{aligned} \tag{72}$$

Since (28) holds for every $\theta \in \Theta_0$ and θ_0 is interior to Θ_0, the inner derivative on the right-hand side of (72) vanishes identically on a neighbourhood of θ_0 and

$$D_{\partial_{\theta'}\psi} = \partial_{\theta'}\Big(\partial_{(\beta,\alpha)}\mathrm{E}_P\big[\psi(W,\theta,\beta_0,\alpha_0)\big][\beta-\beta_0,\alpha-\alpha_0]\Big)\Big|_{\theta=\theta_0} = 0.$$

This shows that derivatives vanish for every $\eta \in T_N$, so ψ^* is Neyman orthogonal at (θ_0,η_0) with respect to T_N. □

A.5 Proof of Theorem 2

Proof. Fix any sequence $\{P_N\}_{N\geq 1}$ with $P_N \in \mathcal{P}_N$ for every N. Since the sequence is arbitrary, it suffices to establish the conclusions under P_N, as in the proof of Chernozhukov et al. (2018, Theorem 3.1). We write O_P for O_{P_N} throughout, and all $O_P(\cdot)$ statements below hold uniformly over $P \in \mathcal{P}_N$, because the bounds depend only on the constants in Assumptions 1–3, which hold for every $P \in \mathcal{P}_N$. Let $\mathcal{E}_N$ be the event, of probability at least $1-\Delta_N$, on which $(\widehat{\beta}_k,\widehat{\alpha}_k) \in \mathcal{B}_N \times \mathcal{A}_N$ for all $k \in [K]$. Because $\Delta_N \to 0$, $\Pr_P(\mathcal{E}_N) \to 1$, so every bound established below on $\mathcal{E}_N$ holds unconditionally up to a term that is $o_P(1)$. Throughout, f denotes any one of the three functions $f(w,\beta,\alpha) \in \big\{\,\psi(w,\theta_0,\beta,\alpha), \psi^a(w,\beta,\alpha), \zeta(w,\beta,\alpha)\,\big\}$, and we have that

$$\mathrm{E}_P\big[\psi(W,\theta_0,\beta_0,\alpha_0)\big] = \bar{m}_0, \qquad \mathrm{E}_P\big[\psi^a(W,\beta_0,\alpha_0)\big] = G_0, \qquad \mathrm{E}_P\big[\zeta(W,\beta_0,\alpha_0)\big] = \Omega_0.$$

We also define the first-step bias and the centred cross-fitted average as

$$b_f(\beta,\alpha) := \mathrm{E}_P\big[f(W,\beta,\alpha)\big] - \mathrm{E}_P\big[f(W,\beta_0,\alpha_0)\big],$$
$$\Delta_f := \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k} f(W_i,\widehat{\beta}_k,\widehat{\alpha}_k) - \mathrm{E}_P\big[f(W,\beta_0,\alpha_0)\big],$$

and abbreviate Δ_ψ, Δ_a and Δ_ζ for the different choices of f, respectively. By Assumption 2, $\partial_{\theta'}\psi(w,\theta,\beta,\alpha) = \psi^a(w,\beta,\alpha)$ does not depend on θ and $\psi(w,\theta,\beta,\alpha) = \psi(w,\theta_0,\beta,\alpha)+\psi^a(w,\beta,\alpha)(\theta-\theta_0)$, so the cross-fitted quantities of Definition 2 satisfy

$$\widehat{\Gamma} = G_0 + \Delta_a, \qquad \widehat{\Omega} = \Omega_0 + \Delta_\zeta, \qquad \widehat{\bar{m}} = \bar{m}_0 + \Delta_\psi + \widehat{\Gamma}\big(\widetilde{\theta}-\theta_0\big).$$

The sample means of the last two terms of (35) vanish and so the DML2 estimator $\check{\theta}$ solves (36) and, whenever $\widehat{\Gamma}'\widehat{\Omega}^{-1}\widehat{\Gamma}$ is nonsingular,

$$\check{\theta}-\theta_0 = -\big(\widehat{\Gamma}'\widehat{\Omega}^{-1}\widehat{\Gamma}\big)^{-1}\widehat{\Gamma}'\widehat{\Omega}^{-1}\big(\bar{m}_0+\Delta_\psi\big). \tag{73}$$

For each $f \in \big\{\psi(w,\theta_0,\beta,\alpha), \psi^a(w,\beta,\alpha), \zeta(w,\beta,\alpha)\big\}$,

$$\Delta_f = \frac{1}{N}\sum_{i=1}^{N}\Big(f(W_i,\beta_0,\alpha_0) - \mathrm{E}_P\big[f(W,\beta_0,\alpha_0)\big]\Big) + \bar{b}_f + O_P\big(\delta_N N^{-1/2}\big), \qquad \bar{b}_f := \frac{1}{K}\sum_{k=1}^{K} b_f\big(\widehat{\beta}_k,\widehat{\alpha}_k\big). \tag{74}$$

The first term is $O_P(N^{-1/2})$ by Chebyshev's inequality and Assumption 3(b). The remainder is $N^{-1}\sum_{i\in I_k}\{f(W_i,\widehat{\beta}_k,\widehat{\alpha}_k) - f(W_i,\beta_0,\alpha_0) - b_f(\widehat{\beta}_k,\widehat{\alpha}_k)\}$. Because $(\widehat{\beta}_k,\widehat{\alpha}_k)$ is constructed from $(W_j)_{j\in I_k^c}$ alone, conditionally on $(W_j)_{j\in I_k^c}$, its summands are independent with mean zero and, conditional on $\mathcal{E}_N$, second moments $\|f(W,\widehat{\beta}_k,\widehat{\alpha}_k) - f(W,\beta_0,\alpha_0)\|_{P,2}^2 \le \delta_N^2$ by Assumption 3(b). Conditionally on $(W_j)_{j\in I_k^c}$, the remainder term in (74) is $O_P(\delta_N N^{-1/2})$ by Chebyshev's inequality, and unconditionally by Lemma 6.1 of Chernozhukov et al. (2018).

Assumptions 3(b) and (c) give, on $\mathcal{E}_N$,

$$\begin{aligned} &\|\Delta_a\| + \|\Delta_\zeta\| = O_P(\delta_N), \qquad \|\Delta_\psi\| + \|\bar{m}_0\|\big(\|\Delta_a\| + \|\Delta_\zeta\|\big) = O_P(N^{-1/2}),\\ &\big\|\bar{b}_\psi\big\| + \|\bar{m}_0\|\big(\|\bar{b}_{\psi^a}\| + \|\bar{b}_\zeta\|\big) \le \delta_N N^{-1/2}, \end{aligned} \tag{75}$$

where $\|\bar{m}_0\| \le C$ by Assumption 3(b). By (75) and Assumption 3(d), $\widehat{\Omega}$ is nonsingular with probability approaching one and $\widehat{\Omega}^{-1} = \Omega_0^{-1} - \Omega_0^{-1}\Delta_\zeta\Omega_0^{-1} + O_P(\|\Delta_\zeta\|^2)$. Expanding and using $G_0'\Omega_0^{-1}\bar{m}_0 = 0$ from (21),

$$\widehat{\Gamma}'\widehat{\Omega}^{-1}\big(\bar{m}_0+\Delta_\psi\big) = G_0'\Omega_0^{-1}\Delta_\psi + \Delta_a'\Omega_0^{-1}\bar{m}_0 - G_0'\Omega_0^{-1}\Delta_\zeta\Omega_0^{-1}\bar{m}_0 + R, \tag{76}$$

where the remainder term $R = O_P(\delta_N)O_P(N^{-1/2})$ by (75).

Substituting (74) into (76), we have

$$\widehat{\Gamma}'\widehat{\Omega}^{-1}\big(\bar{m}_0 + \Delta_\psi\big) = \frac{1}{N}\sum_{i=1}^{N} \psi^*(W_i, \theta_0, \eta_0) + O_P\big(\delta_N N^{-1/2}\big),$$

where the bias parts are bounded by (75) together with Assumption 3(d). Also $\widehat{\Gamma}'\widehat{\Omega}^{-1}\widehat{\Gamma} = J_0 + O_P(\delta_N)$, whose singular values are bounded away from zero by Assumption 3(d). Chebyshev's inequality and the second-moment bound established below give $N^{-1/2}\sum_i \psi^*(W_i, \theta_0, \eta_0) = O_P(1)$. Therefore (73) gives

$$\sqrt{N}\big(\check{\theta} - \theta_0\big) = -J_0^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \psi^*(W_i, \theta_0, \eta_0) + O_P(\delta_N).$$

By (32), $\psi^*(\cdot, \theta_0, \eta_0)$ is a linear combination of $\psi(\cdot, \theta_0, \beta_0, \alpha_0)$, $\psi^a(\cdot, \beta_0, \alpha_0)$ and $\zeta(\cdot, \beta_0, \alpha_0)$ with coefficients bounded by Assumption 3(d), and thus $\psi^*(\cdot, \theta_0, \eta_0)$ has bounded q-th moment with $q > 2$ by Assumption 3(b), while the eigenvalues of $\mathrm{E}_P[\psi^*\psi^{*\prime}]$ are bounded below by Assumption 3(d). Then the Lindeberg-Feller central limit theorem, whose condition is implied by Lyapunov's condition, together with the Cramér-Wold device, yields (38) uniformly over $P \in \mathcal{P}_N$.

Now, we show the consistency of the variance estimator. First of all, $\widehat{J}_0 = \widehat{\Gamma}'\widehat{\Omega}^{-1}\widehat{\Gamma} = J_0 + O_P(\delta_N)$ and $\widehat{\bar{m}} - \bar{m}_0 = \Delta_\psi + \widehat{\Gamma}(\widetilde{\theta} - \theta_0) = O_P(\delta_N)$ by $\|\widetilde{\theta} - \theta_0\| = O_P(\delta_N)$ and $\delta_N \geq N^{-1/2}$. Let $\psi^*_{0,i} := \psi^*(W_i, \theta_0, \eta_0)$ and $\widehat{\psi}^*_i := \psi^*(W_i, \check{\theta}, \widehat{\eta}_k)$ for $i \in I_k$. By (35) and $\psi(W_i, \check{\theta}, \widehat{\beta}_k, \widehat{\alpha}_k) = \psi(W_i, \theta_0, \widehat{\beta}_k, \widehat{\alpha}_k) + \psi^a(W_i, \widehat{\beta}_k, \widehat{\alpha}_k)(\check{\theta} - \theta_0)$, each term of $\widehat{\psi}^*_i - \psi^*_{0,i}$ is either one of the differences $\widehat{\Gamma} - G_0$, $\widehat{\Omega}^{-1} - \Omega_0^{-1}$, $\widehat{\bar{m}} - \bar{m}_0$, $\check{\theta} - \theta_0$, all $O_P(\delta_N)$, multiplied by $f(W_i, \widehat{\beta}_k, \widehat{\alpha}_k)$, or a bounded matrix multiplied by $f(W_i, \widehat{\beta}_k, \widehat{\alpha}_k) - f(W_i, \beta_0, \alpha_0)$, for f in $\big\{\psi(w, \theta_0, \beta, \alpha), \psi^a(w, \beta, \alpha), \zeta(w, \beta, \alpha)\big\}$. Conditional on $\mathcal{E}_N$, the second moments of $f(W_i, \widehat{\beta}_k, \widehat{\alpha}_k)$ and of $f(W_i, \widehat{\beta}_k, \widehat{\alpha}_k) - f(W_i, \beta_0, \alpha_0)$ are bounded by C^2 and by δ_N^2 by Assumption 3(b), so the conditional Markov inequality and Lemma 6.1 of Chernozhukov et al. (2018) give $N^{-1}\sum_{i=1}^N \|\widehat{\psi}^*_i - \psi^*_{0,i}\|^2 = O_P(\delta_N^2)$. Since $N^{-1}\sum_{i=1}^N \|\psi^*_{0,i}\|^2 = O_P(1)$, the Cauchy–Schwarz inequality gives $N^{-1}\sum_{i=1}^N (\widehat{\psi}^*_i\widehat{\psi}^{*\prime}_i - \psi^*_{0,i}\psi^{*\prime}_{0,i}) = O_P(\delta_N)$, and

$$\frac{1}{N}\sum_{i=1}^{N} \psi^*_{0,i}\psi^{*\prime}_{0,i} - \mathrm{E}_P\big[\psi^*\psi^{*\prime}\big] = O_P\big(N^{-[(1-2/q)\wedge 1/2]}\big) = O_P(\delta_N),$$

by Chebyshev's inequality and the Hölder inequality with $q \geq 4$, and by the von Bahr-Esseen inequality with $p = q/2$ when $q \in (2, 4)$, similarly as in the proof of Chernozhukov et al. (2018, Theorem 3.2, equation (A.25)). Hence $\widehat{\sigma}^2 = \sigma^2 + O_P(\delta_N)$, and uniform validity of the confidence intervals follows from (38). □

A.6 Proof of Theorem 3

Proof. It suffices to verify Assumptions 1, 2 and 3; the conclusions then follow from Theorems 1 and 2. First of all, Assumption 2 holds as in (45), and (46) does not depend on θ. For $\beta = (\ell, r) \in \mathcal{B}_N$

and $\alpha = z \in \mathcal{A}_N$ write

$$\delta_\ell := \ell - \ell_0, \qquad \delta_r := r - r_0, \qquad \delta_z := z - z_0, \qquad e(\theta) := \delta_\ell - \delta_r\theta \tag{77}$$

for the deviations from the true values (43), and put $U(\theta) := \tilde{Y} - \tilde{D}\theta$ and $e := e(\theta_0)$, so that $\tilde{U} = U(\theta_0)$. We have

$$\begin{aligned} \psi(w,\theta,\beta,\alpha) &= \big(Z - z(X)\big)\big(Y - \ell(X) - (D - r(X))\theta\big) = (\tilde{Z} - \delta_z)\big(U(\theta) - e(\theta)\big), \\ \psi^a(w,\beta,\alpha) &= -\big(Z - z(X)\big)\big(D - r(X)\big) = -(\tilde{Z} - \delta_z)(\tilde{D} - \delta_r), \\ \zeta(w,\alpha) &= \big(Z - z(X)\big)\big(Z - z(X)\big)' = (\tilde{Z} - \delta_z)(\tilde{Z} - \delta_z)', \end{aligned} \tag{78}$$

which are equal at (β_0, α_0) to $\tilde{Z}U(\theta)$, $-\tilde{Z}\tilde{D}$ and $\tilde{Z}\tilde{Z}'$, respectively.

For every $\theta \in \Theta$,

$$\begin{aligned} &\mathrm{E}_P\big[\psi(W,\theta,\beta,\alpha)\big] = \bar{m}(\theta) + \mathrm{E}_P\big[\delta_z\, e(\theta)\big], \qquad \mathrm{E}_P\big[\psi^a(W,\beta,\alpha)\big] = G_0 - \mathrm{E}_P\big[\delta_z\delta_r\big], \\ &\mathrm{E}_P\big[\zeta(W,\alpha)\big] = \Omega_0 + \mathrm{E}_P\big[\delta_z\delta_z'\big], \end{aligned} \tag{79}$$

since (77) are functions of X, $\mathrm{E}_P[\tilde{Z}|X] = 0$ and $\mathrm{E}_P[U(\theta)|X] = 0$ by (43), so every cross term in (78) has mean zero. Assumption 1(b) follows from (79), since the three maps are (at most) quadratic polynomials in the deviations with second derivatives bounded under Assumption 4(b)–(c), and its interchange condition holds because ψ is linear in θ. Assumption 1(a) holds by Assumption 4(a), from the definition of $\bar{m}(\theta)$ and Ω, and by Assumption 4(d). Its dominance condition also holds by $\mathrm{E}_P\|\tilde{Z}\tilde{D}\| < \infty$ under Assumption 4(b) and the Cauchy–Schwarz inequality.

For Assumption 1(c), (27) holds because $\bar{m}(\theta) = \mathrm{E}_P[Z\tilde{Y}] - \mathrm{E}_P[Z\tilde{D}]\theta = \mathrm{E}_P[\tilde{Z}\tilde{Y}] - \mathrm{E}_P[\tilde{Z}\tilde{D}]\theta$. Along the segment from (β_0, α_0) to (β, α), indexed by $s \in [0, 1)$, the three expansions read $\bar{m}(\theta) + s^2\mathrm{E}_P[\delta_z e(\theta)]$, $G_0 - s^2\mathrm{E}_P[\delta_r\delta_z]$ and $\Omega_0 + s^2\mathrm{E}_P[\delta_z\delta_z']$, whose derivatives at $s = 0$ all vanish. Since θ and the direction were arbitrary, these give (28) and (29). Thus, Assumption 1(c) and (d) hold. Assumption 4(e) implies Assumption 3(a). For Assumption 3(b), the three functions in (78) are products of two factors with bounded $L^q(P)$ norms by Assumption 4(b) and (e), hence have bounded $L^{q/2}(P)$ norms with $q/2 > 2$ by Hölder's inequality.

Expanding (78) at $\theta = \theta_0$, their deviations from the values at (β_0, α_0) are

$$-\delta_z\tilde{U} - \tilde{Z}e + \delta_z e, \qquad \delta_z\tilde{D} + \tilde{Z}\delta_r - \delta_z\delta_r, \qquad -\delta_z\tilde{Z}' - \tilde{Z}\delta_z' + \delta_z\delta_z',$$

respectively. Each linear term in the deviations δ has bounded conditional second moments by Assumption 4(c). The bound for $\tilde{U}$ follows from $\mathrm{E}_P[\tilde{U}^2|X] \le 2\mathrm{E}_P[\tilde{Y}^2|X] + 2\theta_0^2\mathrm{E}_P[\tilde{D}^2|X]$ and the compactness of Θ. Each quadratic term is a product of two deviations δ, bounded by Lyapunov's inequality, $\|\delta\|_{P,4} \le \|\delta\|_{P,2}^{t}\|\delta\|_{P,q}^{1-t}$ with $t = (q-4)/(2q-4) \in (0, \frac{1}{2})$, which gives $\|\delta_z\delta_r\|_{P,2} \le \|\delta_z\|_{P,4}\|\delta_r\|_{P,4} \le C\delta_N^{2t}$ and likewise for $\delta_z e$ and $\delta_z\delta_z'$. Assumption 3(b) therefore holds with $q/2$ in

place of q and

$$\bar{\delta}_N = C\,\delta_N^{2t} \vee N^{-[(1-4/q)\wedge 1/2]}$$

in place of δ_N. Note that $\bar{\delta}_N \geq \delta_N$, and $\bar{\delta}_N \to 0$ because $q > 4$, and so Assumption 3(c) and (d) continue to hold with $\bar{\delta}_N$.

Finally, (79) and the Cauchy–Schwarz inequality give

$$\big\|\mathrm{E}_P[\delta_z e]\big\| + \|\bar{m}_0\|\Big(\big\|\mathrm{E}_P[\delta_z\delta_r]\big\| + \big\|\mathrm{E}_P[\delta_z\delta_z']\big\|\Big) \leq C\Big(\|\delta_z\|_{P,2}\big(\|\delta_\ell\|_{P,2} + \|\delta_r\|_{P,2}\big) + \|\bar{m}_0\|\,\|\delta_z\|_{P,2}^2\Big),$$

which is at most $C\delta_N N^{-1/2}$ by Assumption 4(e), while $\|\mathrm{E}_P[\delta_z\delta_r]\| + \|\mathrm{E}_P[\delta_z\delta_z']\| \leq 2\delta_N^2 \leq \delta_N$. These give Assumption 3(c). Assumption 3(d) is Assumption 4(d), with $G_0 = -\Pi$ and $\Omega_0 = S$.

Then, Theorem 2 gives the limit distribution for the DML estimator $\check{\theta}$, and (53) uses $\widetilde{\theta} = \check{\theta}$, which satisfies $\|\widetilde{\theta} - \theta_0\| = O_P(N^{-1/2})$, so that $\widehat{\sigma}^2$ is consistent and the intervals are uniformly valid. □

A.7 Proof of Theorem 4

Proof. It suffices to verify Assumptions 1, 2 and 3, and then the conclusions follow from Theorems 1 and 2. For $z \in \{0,1\}$, let

$$\delta_{\ell j}^{(z)} := \ell_j(z,\cdot) - \ell_{0j}(z,\cdot), \qquad \delta_{\nu j}^{(z)} := \nu_j(z,\cdot) - \nu_{0j}(z,\cdot), \qquad \delta_{pj} := p_j - p_{0j},$$

denote the deviations of $\beta = (\ell_j, \nu_j)_j \in \mathcal{B}_N$ and $\alpha = (p_j)_j \in \mathcal{A}_N$ from the true values (55).

For the third term in the score ψ_j^Y of (57),

$$\mathrm{E}_P\Big[\frac{Z_j\big(Y - \ell_j(1,X)\big)}{p_j(X)}\,\Big|\,X\Big] = \frac{p_{0j}}{p_j}\big(\ell_{0j}(1,X) - \ell_j(1,X)\big) = -\frac{p_{0j}}{p_j}\,\delta_{\ell j}^{(1)},$$

by definition of (55), and $\mathrm{E}_P[Z_j \mid X] = p_{0j}(X)$ and $\mathrm{E}_P[Z_j Y \mid X] = p_{0j}(X)\ell_{0j}(1,X)$ by the law of iterated expectations. Similarly, the last term of ψ_j^Y, which involves $1 - Z_j$, has conditional mean $\big((1-p_{0j})/(1-p_j)\big)\delta_{\ell j}^{(0)}$. Adding these to $\ell_j(1,X) - \ell_j(0,X) = \ell_{0j}(1,X) - \ell_{0j}(0,X) + \delta_{\ell j}^{(1)} - \delta_{\ell j}^{(0)}$, we have

$$\mathrm{E}_P\big[\psi_j^Y(W,\beta,\alpha)\big] = \tau_j^Y + b_j^Y, \qquad b_j^Y := \mathrm{E}_P\Big[\frac{\delta_{\ell j}^{(1)}\,\delta_{pj}}{p_j}\Big] + \mathrm{E}_P\Big[\frac{\delta_{\ell j}^{(0)}\,\delta_{pj}}{1-p_j}\Big]. \tag{80}$$

The same calculation for the score ψ_j^D with (Y, ℓ_j, ℓ_{0j}) replaced by (D, ν_j, ν_{0j}) gives $\mathrm{E}_P[\psi_j^D(W,\beta,\alpha)] = \tau_j^D + b_j^D$, where b_j^D is b_j^Y of (80) with $\delta_{\nu j}^{(z)}$ in place of $\delta_{\ell j}^{(z)}$. Consequently, by (56) and (59),

$$\mathrm{E}_P\big[\psi(W,\theta,\beta,\alpha)\big] = \bar{m}(\theta) + b^Y - \theta\,b^D \qquad \text{for every } \theta \in \Theta \tag{81}$$

where $b^Y = (b_1^Y, \ldots, b_{d_z}^Y)' \in \mathbb{R}^{d_z}$ and $b^D = (b_1^D, \ldots, b_{d_z}^D)' \in \mathbb{R}^{d_z}$.

Assumption 1(a) holds by Assumption 5(d), and $\bar{m}(\theta)$ is linear in θ, and $\Omega \equiv \Omega_0$ is constant. Ω_0 and $J_0 = \tau^{D\prime}\Omega_0^{-1}\tau^D$ are positive definite by Assumption 5(c). Finally, $\partial_{\theta'}\psi = \psi^a = -\psi^D$

is free of θ, and each ψ_j^D is bounded, because D and Z_j are binary, $\nu_j(z,X) \in [0,1]$ and $p_j \in [\epsilon, 1-\epsilon]$ by Assumption 5(e). By (80) and (81), the maps $(\beta,\alpha) \mapsto \mathrm{E}_P[\psi(W,\theta,\beta,\alpha)]$, $(\beta,\alpha) \mapsto \mathrm{E}_P[\partial_{\theta'}\psi(W,\theta,\beta,\alpha)]$ differ from their values at (β_0,α_0) only through b^Y and b^D. Along the segment $(\beta_0,\alpha_0)+s\big(\beta-\beta_0,\alpha-\alpha_0\big)$, $s\in[0,1)$, the numerators in (80) scale with s^2, while the denominators p_j and $1-p_j$ are bounded away from zero by Assumption 5(a) and (e). Hence, b_j^Y, b_j^D and their first two derivatives are bounded by Assumption 5(e). By construction, $\mathrm{E}_P[\zeta]$ is constant. Thus, all three maps are twice continuously Gateaux differentiable on $\mathcal{B}_N \times \mathcal{A}_N$, and Assumption 1(b) holds.

For Assumption 1(c), b^Y and b^D vanish at (β_0,α_0), so that (81) gives $\bar{m}(\theta) = \mathrm{E}_P[m(W,\theta,\beta_0)]$ by (59). Furthermore, from the proof of part (b), derivatives of b^Y, b^D at $s=0$ vanish for every $\theta\in\Theta_0$ and every direction, which implies (28). Assumption 1(d) and Assumption 2 hold by construction.

Assumption 3(a) is imposed in Assumption 5(e). For part (b), ζ is constant and $\psi^a = -\psi^D$ is bounded, while each term of ψ_j^Y is a bounded function of (Z_j,X) multiplied by Y or $\ell_j(z,X)$, whose $L^q(P)$ norms are bounded by Assumption 5(a) and (e), together with $\|\ell_{0j}(z,\cdot)\|_{P,q} \le \epsilon^{-1/q}\|Y\|_{P,q}$, which follows from the conditional Jensen inequality and Assumption 5(a). Thus, ψ, ψ^a and ζ have bounded $L^q(P)$ norms.

For the $L^2(P)$ deviations, take the following terms of ψ_j^Y,

$$\frac{Z_j\big(Y-\ell_j(1,X)\big)}{p_j} - \frac{Z_j\big(Y-\ell_{0j}(1,X)\big)}{p_{0j}} = Z_j\big(Y-\ell_{0j}(1,X)\big)\frac{p_{0j}-p_j}{p_j p_{0j}} - \frac{Z_j\,\delta_{\ell j}^{(1)}}{p_j}.$$

The $L^2(P)$ norm of the second term is bounded by $\epsilon^{-1}\|\delta_{\ell j}^{(1)}\|_{P,2}$, and the first term has squared $L^2(P)$ norm bounded by $\epsilon^{-4}\mathrm{E}_P\big[\delta_{pj}^2\,\mathrm{E}_P[(Y-\ell_{0j}(1,X))^2 \mid X]\big] \le C\epsilon^{-4}\|\delta_{pj}\|_{P,2}^2 \le \epsilon^{-4}C\delta_N^2$, since $\mathrm{E}_P[(Y-\ell_{0j}(1,X))^2 \mid X] \le 2\mathrm{E}_P[Y^2 \mid X] + 2\ell_{0j}(1,X)^2$ by Assumption 5(b) and (e). The remaining terms in ψ_j^Y with $(1-Z_j)$ are handled identically, and so are those of ψ_j^D, with binary D in place of Y. Therefore, Assumption 3(b) holds.

For Assumption 3(c), derivation (80) and Assumption 5(e), together with the Cauchy–Schwarz inequality give

$$\big\|b^Y\big\| + \big\|b^D\big\| \le C\sum_{j=1}^{d_z}\|\delta_{pj}\|_{P,2}\Big(\sum_{z\in\{0,1\}}\|\delta_{\ell j}^{(z)}\|_{P,2} + \sum_{z\in\{0,1\}}\|\delta_{\nu j}^{(z)}\|_{P,2}\Big) \le C\delta_N N^{-1/2}$$

by (65). Then, Assumption 3(c) holds by the above bound, Assumption 5(d) and $\|\bar{m}_0\| = \|\mathrm{E}_P[\psi(W,\theta_0,\beta_0,\alpha_0)]\| \le C$. Finally, Assumption 3(d) holds by Assumption 5(c).

By Theorem 2, the DML estimator has the stated limit distribution, and (64) uses $\widetilde{\theta} = \check{\theta}$, which satisfies $\|\widetilde{\theta}-\theta_0\| = O_P(N^{-1/2})$, so that $\widehat{\sigma}^2$ is consistent and the intervals are uniformly valid. □

B Non-linear scores and general weighting

In this appendix, we consider the construction of orthogonal scores and the properties of the DML for non-linear scores and general weightings by relaxing Assumption 1(d) (one-step weight) and Assumption 2 (linear scores) in Section 3. Therefore, we can consider nonlinear moment conditions and the iterated GMM weights, where $\zeta(w,\theta,\beta,\alpha) = \psi(w,\theta,\beta,\alpha)\psi(w,\theta,\beta,\alpha)'$ depends on θ and on nuisance parameters (β,α).

The orthogonal score construction in Theorem 1 allows general nonlinear scores, and we consider the orthogonal score ψ^* of the following form with the general weight function correction term $\kappa(\cdot)$

$$\psi^*(w,\theta,\eta) := \mu\, h(w,\theta,\lambda,\beta,\alpha) + \kappa(w,\theta,\eta), \tag{82}$$

where the first part is the orthogonal score (30) of Section 3.1 based on the stacked moment function with the nuisance parameter that is enlarged with an additional nuisance parameter χ, with true value χ_0,

$$\eta := \big(\lambda', \mathrm{vec}(\mu)', \beta, \alpha, \chi\big), \qquad \eta_0 := \big(\lambda_0', \mathrm{vec}(\mu_0)', \beta_0, \alpha_0, \chi_0\big). \tag{83}$$

Evaluated at the true nuisance and with (A_h, Ω_h) chosen so that $\mu_0 J_\theta = J_0$, the first term has the closed form (32), so that for every $\theta \in \Theta$

$$\begin{aligned}\psi^*(w,\theta,\eta_0) = {} & G_0'\Omega_0^{-1}\psi(w,\theta,\beta_0,\alpha_0) + \big(\partial_{\theta'}\psi(w,\theta,\beta_0,\alpha_0) - G_0\big)'\Omega_0^{-1}\bar m_0 \\ & - G_0'\Omega_0^{-1}\big(\zeta(w,\theta,\beta_0,\alpha_0) - \Omega_0\big)\Omega_0^{-1}\bar m_0 + \kappa(w,\theta,\eta_0),\end{aligned} \tag{84}$$

and the Jacobian is

$$J_0 = G_0'\Omega_0^{-1}G_0 + \big(I_{d_\theta} \otimes \bar m_0'\Omega_0^{-1}\big)\,\partial_{\theta'}\,\mathrm{vec}\, G(\theta_0) - \big(\bar m_0'\Omega_0^{-1} \otimes G_0'\Omega_0^{-1}\big)\,\partial_{\theta'}\,\mathrm{vec}\,\Omega(\theta_0). \tag{85}$$

As noted in Remark 3, the correction term $\kappa(w,\theta,\eta)$ is required to relax Assumption 1(d) in Theorem 1. In Assumption 6 below, we give the condition on κ, which exists under mild conditions and can be constructed using the Riesz representer or influence-function adjustment by Ichimura and Newey (2022).

Now we define the DML estimator $\check\theta$ for general non-linear scores and the efficient weight. We make the whole nuisance parameters fold-specific, as in the original definition of Chernozhukov et al. (2018), whereas Definition 2 constructs $(\widehat\Gamma, \widehat\Omega, \widehat{\bar m})$ as averages over all folds for linear scores and a one-step weight, and exploits the resulting closed form of $\check\theta$ in Theorem 2.

Definition 3 (DML estimator)**.** *(a) Take a K-fold random partition $(I_k)_{k=1}^K$ of $[N]$ with $|I_k| = n = N/K$.*

(b) For each $k \in [K]$, construct an ML estimator $\widehat\eta_k := (\widehat\lambda_k', \mathrm{vec}(\widehat\mu_k)', \widehat\beta_k, \widehat\alpha_k, \widehat\chi_k)$ of η_0 from $(W_i)_{i\in I_k^c}$ alone. To be specific, construct $\widehat\beta_k$, $\widehat\alpha_k$ and a preliminary estimator $\widetilde\theta_k$ of θ_0 from $(W_i)_{i\in I_k^c}$, and evaluate the cross-fitted Jacobian, weight matrix and moment $\widehat\Gamma_k$, $\widehat\Omega_k$, $\widehat{\bar m}_k$, and $\widehat\chi_k$ at $\widetilde\theta_k$, again using $(W_i)_{i\in I_k^c}$ only. Then set $\widehat\lambda_k := (\mathrm{vec}(\widehat\Gamma_k)', \mathrm{vec}(\widehat\Omega_k)')'$ and $\widehat\mu_k =$

$\Big[-\big(I_{d_\theta} \otimes \widehat{\bar{m}}_k' \widehat{\Omega}_k^{-1}\big) \quad \widehat{\bar{m}}_k' \widehat{\Omega}_k^{-1} \otimes \widehat{\Gamma}_k' \widehat{\Omega}_k^{-1} \quad I_{d_\theta} \Big]$ *as in Definition 2. Then, by* (82),

$$\begin{aligned}\psi^*(W_i, \theta, \widehat{\eta}_k) = \widehat{\Gamma}_k' \widehat{\Omega}_k^{-1} \psi(W_i, \theta, \widehat{\beta}_k, \widehat{\alpha}_k) + \big(\partial_{\theta'} \psi(W_i, \theta, \widehat{\beta}_k, \widehat{\alpha}_k) - \widehat{\Gamma}_k\big)' \widehat{\Omega}_k^{-1} \widehat{\bar{m}}_k \\ - \widehat{\Gamma}_k' \widehat{\Omega}_k^{-1} \big(\zeta(W_i, \theta, \widehat{\beta}_k, \widehat{\alpha}_k) - \widehat{\Omega}_k\big) \widehat{\Omega}_k^{-1} \widehat{\bar{m}}_k + \kappa(W_i, \theta, \widehat{\eta}_k).\end{aligned}$$

(c) Construct the DML estimator $\check{\theta}$ *in either of the two ways below:*

(DML1) For each $k \in [K]$*, let* $\check{\theta}_k$ *be an approximate solution of*

$$\Big\| \frac{1}{n} \sum_{i \in I_k} \psi^*\big(W_i, \check{\theta}_k, \widehat{\eta}_k\big) \Big\| \leq \inf_{\theta \in \Theta} \Big\| \frac{1}{n} \sum_{i \in I_k} \psi^*\big(W_i, \theta, \widehat{\eta}_k\big) \Big\| + \varepsilon_N, \tag{86}$$

where $\varepsilon_N = o(\delta_N N^{-1/2})$*, and* δ_N *is some sequence of positive constants converging to zero. Then, aggregate over the folds,* $\check{\theta} := K^{-1} \sum_{k=1}^K \check{\theta}_k$*.*

(DML2) Construct $\check{\theta}$ *as an approximate solution of*

$$\Big\| \frac{1}{N} \sum_{k=1}^K \sum_{i \in I_k} \psi^*\big(W_i, \check{\theta}, \widehat{\eta}_k\big) \Big\| \leq \inf_{\theta \in \Theta} \Big\| \frac{1}{N} \sum_{k=1}^K \sum_{i \in I_k} \psi^*\big(W_i, \theta, \widehat{\eta}_k\big) \Big\| + \varepsilon_N. \tag{87}$$

B.1 Asymptotic distribution of the DML for non-linear scores and general weighting

Assumption 6 (Weight correction)**.** *For all* $P \in \mathcal{P}_N$*, the correction term* $\kappa(w, \theta, \eta)$ *in* (82) *satisfies the following conditions:* $\mathrm{E}_P[\kappa(W, \theta_0, \eta_0)] = 0$ *and* $\partial_{\theta'} \mathrm{E}_P[\kappa(W, \theta, \eta_0)]|_{\theta=\theta_0} = 0$*. The Gateaux derivatives of* $\mathrm{E}_P[\kappa(W, \theta_0, \eta_0)]$ *in the directions of* λ*,* $\mathrm{vec}(\mu)$ *and* χ *vanish. For all* $(\beta, \alpha) \in \mathcal{B}_N \times \mathcal{A}_N$*,*

$$\begin{aligned}&\partial_{(\beta,\alpha)} \mathrm{E}_P\big[\kappa(W, \theta_0, \eta_0)\big][\beta - \beta_0, \alpha - \alpha_0] \\ &\quad = G_0' \Omega_0^{-1} \, \partial_{(\beta,\alpha)} \mathrm{E}_P\big[\zeta(W, \theta_0, \beta_0, \alpha_0)\big][\beta - \beta_0, \alpha - \alpha_0] \, \Omega_0^{-1} \bar{m}_0.\end{aligned} \tag{88}$$

Assumption 6 replaces Assumption 1(d) and provides the conditions on κ. The correction term κ takes into account the derivative of the weight function with respect to the first step, and the assumption is used only for the orthogonality of ψ^* in Theorem 5 below. Assumption 6 keeps the moment condition and leaves the Jacobian unchanged.

Assumption 7 (Regularity conditions for non-linear scores)**.** *Let* $0 < c_0 \leq c_1$*,* $c_2 > 1$*,* $q > 2$*, and* $v > 0$ *be constants and* $\{\delta_N\}$*,* $\{\Delta_N\}$*,* $\{\tau_N\}$ *sequences of positive constants converging to zero, with* $\delta_N \geq N^{-1/2+1/q} \log N$ *and* $N^{-1/2} \log N \leq \tau_N \leq \delta_N$*. For all* $N \geq 3$ *and* $P \in \mathcal{P}_N$*, the following conditions hold.*

(a) Θ *is bounded and contains a ball of radius* $c_1 N^{-1/2} \log N$ *centred at* θ_0 *of* (21)*.*

(b) The map $(\theta, \eta) \mapsto \mathrm{E}_P[\psi^*(W, \theta, \eta)]$ *is twice continuously Gateaux-differentiable on* $\Theta \times T$*, where* $T := \Lambda \times \mathbb{R}^{d_\theta d_h} \times \mathcal{B} \times \mathcal{A} \times \mathcal{X}$ *is convex. For every* $k \in [K]$*,* $\widehat{\eta}_k$ *belongs to the nuisance realization*

set T_N with probability at least $1-\Delta_N$, where $T_N := \Lambda_N \times \mathcal{M}_N \times \mathcal{B}_N \times \mathcal{A}_N \times \mathcal{X}_N \subset T$ contains η_0 and is constrained by conditions given below, and Λ_N, $\mathcal{M}_N$ and $\mathcal{X}_N$ are neighbourhoods of λ_0, μ_0 and χ_0.

(c) For all $\theta \in \Theta$, $2\|\mathrm{E}_P[\psi^*(W,\theta,\eta_0)]\| \geq \|J_0(\theta-\theta_0)\| \wedge c_0$, where J_0 of (85) has singular values between c_0 and c_1. The variance of the score is non-degenerate. All eigenvalues of $\mathrm{E}_P[\psi^*(W,\theta_0,\eta_0)\psi^*(W,\theta_0,\eta_0)']$ are bounded below by c_0.

(d) For each $\eta \in T_N$, the function class

$$\mathcal{F}_\eta := \big\{\psi_j^*(\cdot,\theta,\eta) : j = 1,\dots,d_\theta,\ \theta \in \Theta\big\}$$

is suitably measurable with a measurable envelope F_η satisfying $\|F_\eta\|_{P,q} \leq c_1$, and its uniform covering entropy obeys

$$\sup_Q \ \log N\big(t\,\|F_\eta\|_{Q,2},\ \mathcal{F}_\eta,\ \|\cdot\|_{Q,2}\big) \ \leq\ v\log(c_2/t) \qquad \text{for all } 0<t\leq 1,$$

where $N(s,\mathcal{F},\|\cdot\|)$ denotes the s-covering number of $\mathcal{F}$ with respect to $\|\cdot\|$ and the supremum runs over all finitely discrete probability measures Q on the sample space.

(e) The following conditions on the statistical rates r_N, r'_N, and λ'_N hold:

$$r_N := \sup_{\eta\in T_N,\ \theta\in\Theta} \big\|\mathrm{E}_P[\psi^*(W,\theta,\eta)] - \mathrm{E}_P[\psi^*(W,\theta,\eta_0)]\big\| \leq \delta_N\tau_N,$$
$$r'_N := \sup_{\eta\in T_N,\ \|\theta-\theta_0\|\leq\tau_N} \big(\mathrm{E}_P\big\|\psi^*(W,\theta,\eta)-\psi^*(W,\theta_0,\eta_0)\big\|^2\big)^{1/2},\ \text{and } r'_N\log^{1/2}(1/r'_N)\leq\delta_N,$$
$$\lambda'_N := \sup_{r\in(0,1),\ \eta\in T_N,\ \|\theta-\theta_0\|\leq\tau_N} \big\|\partial_r^2\mathrm{E}_P\big[\psi^*\big(W,\theta_0+r(\theta-\theta_0),\eta_0+r(\eta-\eta_0)\big)\big]\big\| \leq \delta_N N^{-1/2}.$$

Assumption 7 is similar to Chernozhukov et al. (2018, Assumptions 3.3 and 3.4) imposed on the misspecification-robust score $\psi^*(W,\theta,\eta)$ and the enlarged nuisance parameter $\eta = \big(\lambda', \mathrm{vec}(\mu)', \beta, \alpha, \chi\big)$. As discussed in Chernozhukov et al. (2018), these are mild regularity conditions and standard in moment condition models. Under correct specification ($\bar{m}_0 = 0$), $\psi^* = G_0'\Omega_0^{-1}\psi$ and the conditions in Assumption 7 reduce to those of Chernozhukov et al. (2018); under misspecification ($\bar{m}_0 \neq 0$), we use conditions on a different Jacobian J_0 with additional smoothness conditions (Assumption 1(a)), and the identification condition at the GMM estimand (21), instead of the true value.

Remark 4 (Partially linear IV model (cont.))**.** *In the partially linear IV model with the efficient weights $\zeta(w,\theta,\beta,\alpha) = \psi(w,\theta,\beta,\alpha)\psi(w,\theta,\beta,\alpha)'$, the target becomes the iterated GMM estimand, i.e., the fixed point of (21) with $\Omega(\theta) = \mathrm{E}_P[\tilde{Z}\tilde{Z}'U(\theta)^2]$, where $U(\theta) = \tilde{Y} - \tilde{D}\theta$, so that $\Omega_0 = \Omega(\theta_0) = \mathrm{E}_P[\tilde{Z}\tilde{Z}'\tilde{U}^2]$ with $\tilde{U} = U(\theta_0)$. The weight function now depends on the first step and violates (29): its derivative in the direction $(\ell-\ell_0, r-r_0, z-z_0)$ does not vanish in general. The correction term is*

$$\kappa(w,\theta,\eta) = -\chi_1(X)\big(Z - z(X)\big) - 2\,\chi_2(X)\big(Y - \ell(X) - (D-r(X))\theta\big), \tag{89}$$

with the additional nuisance parameter $\chi = (\chi_1, \chi_2)$*, whose true value* $\chi_0 = (\chi_{10}, \chi_{20})$ *is*

$$\chi_{10}(X) := \mathrm{E}_P\big[\tilde{U}^2\big((\bar{m}_0'\Omega_0^{-1}\tilde{Z})\Pi'\Omega_0^{-1} + (\Pi'\Omega_0^{-1}\tilde{Z})\bar{m}_0'\Omega_0^{-1}\big) \,\big|\, X\big] \in \mathbb{R}^{1\times d_z},$$
$$\chi_{20}(X) := \mathrm{E}_P\big[(\Pi'\Omega_0^{-1}\tilde{Z})(\bar{m}_0'\Omega_0^{-1}\tilde{Z})\tilde{U} \,\big|\, X\big] \in \mathbb{R}.$$

Both are conditional expectations, which can be estimated by machine learning, and both are proportional to $\bar{m}_0$*, so that the correction vanishes under correct specification. It is straightforward to verify that* κ *in* (89) *satisfies Assumption 6. Then, the orthogonal score is the analogue of* (48)*,*

$$\begin{aligned}\psi^*(w, \theta_0, \eta_0) = &- \Pi'\Omega_0^{-1}\tilde{Z}\tilde{U} - \big(\tilde{Z}\tilde{D} - \Pi\big)'\Omega_0^{-1}\bar{m}_0 \\ &+ \Pi'\Omega_0^{-1}\big(\tilde{Z}\tilde{Z}'\tilde{U}^2 - \Omega_0\big)\Omega_0^{-1}\bar{m}_0 - \chi_{10}(X)\tilde{Z} - 2\,\chi_{20}(X)\tilde{U},\end{aligned}$$

with the Jacobian

$$J_0 = \Pi'\Omega_0^{-1}\Pi - 2\,\Pi'\Omega_0^{-1}\mathrm{E}_P\big[\tilde{Z}\tilde{Z}'\tilde{U}\tilde{D}\big]\Omega_0^{-1}\bar{m}_0,$$

in place of $J_0 = \Pi' S^{-1}\Pi$ *in Section 4.*

Theorem 5 (Orthogonal scores and DML inference for non-linear scores and general weights)**.** *Let* ψ^* *be given by* (82) *with* (A_h, Ω_h) *chosen so that* $\mu_0 J_\theta = J_0$*.*

(i) Under Assumption 1(a)–(c) and Assumption 6, $\mathrm{E}_P[\psi^*(W, \theta_0, \eta_0)] = 0$*, and the score* ψ^* *is Neyman orthogonal at* (θ_0, η_0)*.*

(ii) If, in addition, Assumption 7 holds, the estimator $\check{\theta}$ *of Definition 3, in either its DML1 or its DML2 form, satisfies*

$$\sqrt{N}\,\sigma^{-1}\big(\check{\theta} - \theta_0\big) \xrightarrow{d} \mathcal{N}\big(0, I_{d_\theta}\big), \qquad \sigma^2 = J_0^{-1}\,\mathrm{E}_P\big[\psi^*(W, \theta_0, \eta_0)\psi^*(W, \theta_0, \eta_0)'\big]\,J_0^{-1\prime},$$

uniformly over $P \in \mathcal{P}_N$*. Moreover,* σ^2 *can be replaced by a consistent estimator* $\widehat{\sigma}^2$ *with* $\widehat{\sigma}^2 = \sigma^2 + o_P(1)$ *uniformly over* $P \in \mathcal{P}_N$*, and the confidence intervals* (40) *built from* $\check{\theta}$ *and such a* $\widehat{\sigma}^2$ *are uniformly asymptotically valid.*

Theorem 5 covers the general non-linear scores and weightings, which the main results in Section 3 do not cover. Above all, we allow the efficient weight $\zeta = \psi\psi'$ of Remarks 3 and 4, where ζ depends on both θ and the first step (β, α), and thus $\kappa \not\equiv 0$. If ψ and ζ do not depend on the nonparametric first step (β, α), then $\kappa \equiv 0$, and Theorem 5 covers standard non-linear GMM with efficient weighting: θ_0 is the iterated GMM fixed point, J_0 is (85), and σ^2 is the misspecification-robust variance (Proposition 2) as in Hall and Inoue (2003) and Hansen and Lee (2021). Furthermore, Theorem 2 is not a special case of Theorem 5, because the two use different estimators: Definition 3 makes $(\widehat{\Gamma}, \widehat{\Omega}, \widehat{\bar{m}})$ fold-specific as in Chernozhukov et al. (2018), whereas Definition 2 pools them over folds and exploits the closed form (73) for linear scores, which requires neither the entropy condition (Assumption 7(d)) nor the global identification condition (Assumption 7(c)).

B.2 Proof of Theorem 5

Proof. (i) Note that Assumption 2 is not used in the proof of Theorem 1, and Assumption 1(d) implies $D_\zeta = \partial_{(\beta,\alpha)} \mathrm{E}_P[\zeta(W, \theta_0, \beta_0, \alpha_0)][\beta - \beta_0, \alpha - \alpha_0] = 0$ for $(\beta, \alpha) \in \mathcal{B}_N \times \mathcal{A}_N$. Therefore, similar arguments carry over from the proof of Theorem 1, and it remains to account for κ and for $D_\zeta \neq 0$.

The moment condition $\mathrm{E}_P[\mu_0 h(W, \theta_0, \lambda_0, \beta_0, \alpha_0)] + \mathrm{E}_P[\kappa(W, \theta_0, \eta_0)] = 0$ holds by (31), as in the proof of Theorem 1, with the enlarged nuisance parameter $\eta = \big(\lambda', \mathrm{vec}(\mu)', \beta, \alpha, \chi\big)$, and by Assumption 6. The derivatives of $\mathrm{E}_P[\mu_0 h]$ in λ and $\mathrm{vec}(\mu)$ vanish as in the proof of Theorem 1, and those of $\mathrm{E}_P[\kappa]$ in λ, $\mathrm{vec}(\mu)$ and χ vanish by Assumption 6.

For $(\beta, \alpha) \in \mathcal{B}_N \times \mathcal{A}_N$, similar calculations in the proof of Theorem 1 give

$$\partial_{(\beta,\alpha)} \mathrm{E}_P\big[\psi^*(W, \theta_0, \eta_0)\big][\beta - \beta_0, \alpha - \alpha_0] = G_0' \Omega_0^{-1} D_\psi + D'_{\partial_{\theta'}\psi}\, \Omega_0^{-1} \bar{m}_0 - G_0' \Omega_0^{-1} D_\zeta\, \Omega_0^{-1} \bar{m}_0 + D_\kappa = 0,$$

where the last two terms cancel because $D_\kappa = G_0' \Omega_0^{-1} D_\zeta \Omega_0^{-1} \bar{m}_0$ by (88). As in the proof of Theorem 1, $D_\psi = 0$ and $D_{\partial_{\theta'}\psi} = 0$ by Assumption 1(c), and $\partial_{\theta'} \mathrm{E}_P[\kappa(W, \theta, \eta_0)]|_{\theta=\theta_0} = 0$ by Assumption 6. The closed form (84) can be similarly constructed as (71) at (β_0, α_0) plus κ, and the Jacobian remains unchanged $\mu_0 J_\theta = J_0$ of (85) by Assumption 6.

(ii) We give the argument for the DML1 form of Definition 3 and note that the DML2 form is treated in the same way as done in Chernozhukov et al. (2018). We closely follow the proof of Theorem 3.3 of Chernozhukov et al. (2018), and their Lemma 6.3, and show that the score ψ^* of (82), the nuisance parameter η of (83) and the GMM estimand θ_0 of (21) satisfy Assumptions 3.3 and 3.4 of Chernozhukov et al. (2018).

Fix a sequence $\{P_N\}_{N \geq 1}$ with $P_N \in \mathcal{P}_N$ for every N; since the sequence is arbitrary it suffices to establish the conclusion under P_N, and we write O_P for O_{P_N}. Definition 3 makes the whole of $\widehat{\eta}_k$ fold-specific, and its DML1 and DML2 forms are their Definitions 3.1 and 3.2 for the score ψ^* with $\varepsilon_N = o(\delta_N N^{-1/2})$. Note that $\mathrm{E}_P[\psi^*(W, \theta_0, \eta_0)] = 0$ and the Neyman orthogonality holds by part (i), and thus Assumption 3.3 of Chernozhukov et al. (2018) holds by Assumption 7(a)–(c), where $\partial_{\theta'} \mathrm{E}_P[\psi^*(W, \theta, \eta_0)]|_{\theta=\theta_0} = J_0$ of (85) by the proof of part (i). Similarly, Assumption 7 also implies their Assumption 3.4 with the score ψ^*. The growth conditions $\delta_N \geq N^{-1/2+1/q} \log N$ and $N^{-1/2} \log N \leq \tau_N \leq \delta_N$ of their Theorem 3.3 are imposed in Assumption 7.

Fix $k \in [K]$. Conditional on $(W_i)_{i \in I_k^c}$, the nuisance $\widehat{\eta}_k$ lies in T_N with probability at least $1 - \Delta_N$. We condition on the event in which all K first steps lie in T_N with probability at least $1 - K\Delta_N = 1 - o(1)$. Hence, Lemma 6.3 of Chernozhukov et al. (2018), which establishes approximate linearity of the subsample DML estimators, applies to the fold-k estimator (86) and gives, uniformly over $P \in \mathcal{P}_N$,

$$\sqrt{n}\, \sigma^{-1} \big(\check{\theta}_k - \theta_0\big) = -\frac{1}{\sqrt{n}} \sum_{i \in I_k} \sigma^{-1} J_0^{-1} \psi^*(W_i, \theta_0, \eta_0) + O_P(\delta_N),$$

where $n = N/K$. Hence their Theorem 3.3 applies to the DML1 form of $\check{\theta}$ (averaging over $k \in [K]$)

and gives, uniformly over $P \in \mathcal{P}_N$,

$$\sqrt{N}\,\sigma^{-1}\big(\check{\theta} - \theta_0\big) = -\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sigma^{-1} J_0^{-1} \psi^*(W_i, \theta_0, \eta_0) + O_P(\rho_N) \;\xrightarrow{d}\; \mathcal{N}\big(0, I_{d_\theta}\big),$$

where $\rho_N = N^{-1/2+1/q} \log N + r'_N \log^{1/2}(1/r'_N) + N^{1/2}\lambda_N + N^{1/2}\lambda'_N \leq C\delta_N$ by the condition on δ_N, and Assumption 7(e). The uniform validity of the confidence interval holds by their Theorem 3.3 together with their Corollary 3.1. □