

Multiway Cluster Robust Double/Debiased Machine Learning

Harold D. Chiang^{*} Kengo Kato[†] Yukun Ma[‡] Yuya Sasaki[§]

Abstract

This paper investigates double/debiased machine learning (DML) under multiway clustered sampling environments. We propose a novel multiway cross fitting algorithm and a multiway DML estimator based on this algorithm. Simulations indicate that the proposed procedure has favorable finite sample performance.

Keywords: double/debiased machine learning, multiway clustering, multiway cross fitting

JEL Codes: C10, C13, C14

1 Introduction

We propose a novel multiway cross fitting algorithm and a double/debiased machine learning (DML) estimator based on the proposed algorithm. This objective is motivated by recently growing interest in use of dependent cross sectional data and recently increasing demand for DML methods in empirical

^{*}Harold D. Chiang: harold.d.chiang@vanderbilt.edu. Department of Economics, Vanderbilt University, VU Station B #351819, 2301 Vanderbilt Place, Nashville, TN 37235-1819, USA

[†]Kengo Kato: kk976@cornell.edu. Department of Statistics and Data Science, Cornell University, 1194 Comstock Hall, Ithaca, NY 14853, USA

[‡]Yukun Ma: yukun.ma@vanderbilt.edu. Department of Economics, Vanderbilt University, VU Station B #351819, 2301 Vanderbilt Place, Nashville, TN 37235-1819, USA

[§]Yuya Sasaki: yuya.sasaki@vanderbilt.edu. Department of Economics, Vanderbilt University, VU Station B #351819, 2301 Vanderbilt Place, Nashville, TN 37235-1819, USA

research. On one hand, researchers frequently use multiway cluster sampled data in empirical studies, such as network data, matched employer-employee data, matched student-teacher data, scanner data where observations are double-indexed by stores and products, and market share data where observations are double-indexed by market and products. On the other hand, we have witnessed rapidly increasing popularity of machine learning methods in empirical studies, such as random forests, lasso, post-lasso, elastic nets, ridge, deep neural networks, and boosted trees among others. To date, available DML methods focus on i.i.d. sampled data. In light of the aforementioned research environments today, a new method of DML that is applicable to multiway cluster sampled data may well be of interest by empirical researchers.

The DML was proposed by the recent influential paper by Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Nevo (CCDDHNR, 2018). They provide a general DML toolbox for estimation and inference for structural parameters with high-dimensional and/or infinite-dimensional nuisance parameters. In that paper, the estimation method and properties of the estimator are presented under the typical microeconomic assumption of i.i.d. sampling. We advance this frontier literature of DML by proposing a modified DML estimation procedure with multiway cross fitting, which accommodates multiway cluster sampled data. Even for multiway cluster sampled data, we show that the proposed DML procedure works under nearly identical set of assumptions to that of CCDDHNR (2018). To our best knowledge, the present paper is the first to consider generic DML methods under multiway cluster sampling.

Another branch of the literature following the seminal work by Cameron, Gelbach, and Miller (2011) proposes multiway cluster robust inference methods. Menzel (2017) conducts formal analyses of bootstrap validity under multiway cluster sampling robustly accounting for non-degenerate and degenerate cases. Davezies, D'Haultfoeuille, and Guyonvarch (2018) develop empirical process theory under multiway cluster sampling which applies to a large class of models. We advance this practically important literature by developing a multiway cluster robust inference method based on DML. In deriving theoretical properties of the proposed estimator, we take advantage of the Aldous-Hoover representation employed by the preceding papers. To our knowledge, the present paper is the first in this literature on multiway clustering to develop generic DML methods.

1.1 Relations to the Literature

The past few years have seen a fast growing literature in machine learning based econometric methods. For general overviews of the field, see, e.g., Athey and Imbens (2019) or Mullainathan and Spiess (2017). For a review of estimation and inference methods for high-dimensional data, see Belloni, Chernozhukov, and Hansen (2014a). For an overview of data sketching methods tackling computationally impractically large number of observations, see Lee and Ng (2019). The DML of CCDDHNR (2018) is built upon Belloni, Chernozhukov, and Kato (2015), which proposes to use Neyman orthogonal moments for a general class of Z-estimation statistical problems in the presence of high-dimensional nuisance parameters. This framework is further generalized in different directions by Belloni, Chernozhukov, Fernández-Val, and Hansen (2017) and Belloni, Chernozhukov, Chetverikov, and Wei (2018). CCDDHNR (2018) combine the use of Neyman orthogonality condition with cross fitting to provide a simple yet widely applicable framework that covers a large class of models under i.i.d. settings. The DML is also compatible with various types of machine learning based methods for nuisance parameter estimation.

Driven by the need from empiricists, the literature on cluster robust inference has a long history in econometrics. For recent review of the literature, see, e.g., Cameron and Miller (2015) and MacKinnon (2019). On the other hand, coping with cross-sectional dependence using a multiway cluster robust variance estimator is a relatively recent phenomenon. Cameron et al. (2011) first provide a multiway cluster robust variance estimator for linear regression models without imposing additional parametric assumptions on the intra-cluster correlation structure. This variance estimator has significantly reshaped the landscape of econometric practices in applied microeconomics in the past decade.¹ In contrast to the popularity among empirical researchers, theoretical justification of the validity of this type of procedures was lagging behind. The first rigorous treatment of asymptotic properties of multiway cluster robust estimators are established by Menzel (2017) using the Aldous-Hoover representation under the assumptions of separable exchangeability and dissociation. The asymptotic theory of Menzel (2017) covers both non-degenerate and degenerate cases. Focusing on non-degenerate situations, Davezies et al. (2018) further extend this approach to a general empirical process theory.² Using

¹As of July 2019, Cameron et al. (2011) has received over 2,300 citations. The majority of such citations came from applied economic papers.

²See also Davezies, D’Haultfoeuille, and Guyonvarch (2019) for further generalization of the empirical process theory

this asymptotic framework, MacKinnon, Nielsen, and Webb (2019) study linear regression models under the non-degenerate case and examine the validity of several types of wild bootstrap procedures and the robustness of multiway cluster robust variance estimators under different cluster sampling settings.

Despite of the popularity of both machine learning and cluster robust inference among empirical researchers, relatively limited cluster robust inference results exist for machine learning based methods. Inference for machine learning based methods with one-way clustering is studied by Belloni, Chernozhukov, Hansen, and Kolesár (2016), Kock (2016), Kock and Tang (2019), Semenova, Goldman, Chernozhukov, and Taddy (2018) and Hansen and Liao (2019) for different variations of regularized regression estimators and Athey and Wager (2019) for random forests. Chiang and Sasaki (2019) investigate the performance of lasso and post-lasso in the partially linear model setting of Belloni, Chernozhukov, and Hansen (2014b) under multiway cluster sampling. To our best knowledge, there is no general machine learning based procedures with known validity under multiway cluster sampling environments.

2 Overview

2.1 Setup

Suppose that the researcher observes a sample $\{W_{ij} | i \in \{1, \dots, N\}, j \in \{1, \dots, M\}\}$ of double-indexed observations of size $NM = n$. Let P denote the probability law of $\{W_{ij}\}_{ij}$, and let E_P denote the expectation with respect to P . Let $\underline{C} = N \wedge M$ denote the sample size in the smaller dimension. We consider two-way clustering where each cell contains one observation for simplicity of notations, but results for higher cluster dimensions and random cluster sizes can be obtained following the notation of Davezies et al. (2018).

The structural model is assumed to entail the moment restriction

$$E_P[\psi(W_{11}; \theta_0, \eta_0)] = 0 \tag{2.1}$$

for some score ψ that depends on a low-dimensional parameter vector $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ and a nuisance parameter $\eta \in T$ for a convex set T . The nuisance parameter η may be finite-, high-, or infinite-dimensional for dyadic data under joint exchangeability assumption.

dimensional, and its true value is denoted by $\eta_0 \in T$. In this setup, the true value of the low-dimensional target parameter, denoted by $\theta_0 \in \Theta$, is the object of interest.

Let $\tilde{T} = \{\eta - \eta_0 : \eta \in T\}$, and define the Gateaux derivative map $D_r : \tilde{T} \rightarrow \mathbb{R}^{d_\theta}$ by

$$D_r[\eta - \eta_0] := \partial_r \left\{ \mathbb{E}_P[\psi(W_{11}; \theta_0, \eta_0 + r(\eta - \eta_0))] \right\}$$

for all $r \in [0, 1)$. Also denote its limit by

$$\partial_\eta \mathbb{E}_P \psi(W_{11}; \theta_0, \eta_0)[\eta - \eta_0] := D_0[\eta - \eta_0].$$

We say that the Neyman orthogonality condition holds at (θ_0, η_0) with respect to a nuisance realization set $\mathcal{T}_n \subset T$ if the score ψ satisfies (2.1), the pathwise derivative $D_r[\eta - \eta_0]$ exists for all $r \in [0, 1)$ and $\eta \in \mathcal{T}_n$, and the orthogonality equation

$$\partial_\eta \mathbb{E}_P \psi(W_{11}; \theta_0, \eta_0)[\eta - \eta_0] = 0 \tag{2.2}$$

holds for all $\eta \in \mathcal{T}_n$. Furthermore, we also say that the λ_n Neyman near-orthogonality condition holds at (θ_0, η_0) with respect to a nuisance realization set $\mathcal{T}_n \subset T$ if the score ψ satisfies (2.1), the pathwise derivative $D_r[\eta - \eta_0]$ exists for all $r \in [0, 1)$ and $\eta \in \mathcal{T}_n$, and the orthogonality equation

$$\sup_{\eta \in \mathcal{T}_n} \left\| \partial_\eta \mathbb{E}_P \psi(W; \theta_0, \eta_0)[\eta - \eta_0] \right\| \leq \lambda_n \tag{2.3}$$

holds for all $\eta \in \mathcal{T}_n$ for some positive sequence $\{\lambda_n\}_n$ such that $\lambda_n = o(\underline{C}^{-1/2})$.

Throughout, we will consider structural models satisfying the moment restriction (2.1) and either form of the Neyman orthogonality conditions, (2.2) or (2.3). We focus on linear Neyman orthogonal scores ψ of the form

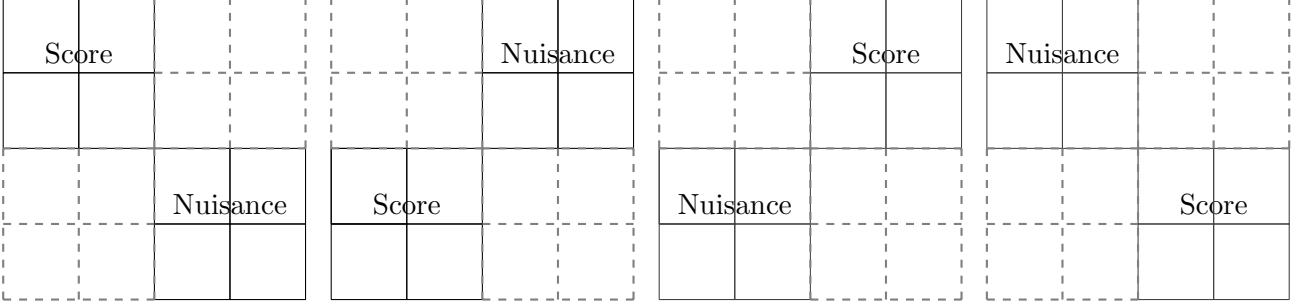
$$\psi(w; \theta, \eta) = \psi^a(w; \eta)\theta + \psi^b(w; \eta), \text{ for all } w \in \text{supp}(W), \theta \in \Theta, \eta \in T. \tag{2.4}$$

A generalization to nonlinear score follows from linearization with Gateaux differentiability as in Section 3.3 of CCDDHNR (2018).

2.2 The Multiway Double/Debiased Machine Learning

For the class of models introduced in Section 2.1, we propose a novel K^2 -fold multiway cross fitting procedure for estimation of θ_0 . For any $r \in \mathbb{N}$, we use the notation $[r] = \{1, \dots, r\}$. With a fixed positive

Figure 1: An illustration of 2^2 -fold cross fitting.



integer K , randomly partition $[N]$ into K parts $\{I_1, \dots, I_K\}$ and $[M]$ into K parts $\{J_1, \dots, J_K\}$. For each $(k, \ell) \in [K]^2$, obtain an estimate

$$\hat{\eta}_{k\ell} = \hat{\eta}((W_{ij})_{(i,j) \in ([N] \setminus I_k) \times ([M] \setminus J_\ell)})$$

of the nuisance parameter η by some machine learning method (e.g., lasso, post-lasso, elastic nets, ridge, deep neural networks, and boosted trees) using only the subsample of those observations with multiway indices (i, j) in $([N] \setminus I_k) \times ([M] \setminus J_\ell)$. In turn, we define $\tilde{\theta}$, the multiway double/debiased machine learning (multiway DML) estimator for θ_0 , as the solution to

$$\frac{1}{K^2} \sum_{(k,\ell) \in [K]^2} \mathbb{E}_{n,k\ell}[\psi(W; \tilde{\theta}, \hat{\eta}_{k\ell})] = 0, \quad (2.5)$$

where $\mathbb{E}_{n,k\ell}[f(W)] = \frac{1}{|I_k||J_\ell|} \sum_{(i,j) \in I_k \times J_\ell} f(W_{ij})$ denotes the subsample empirical expectation using only the those observations with multiway indices (i, j) in $I_k \times J_\ell$.

We call this procedure the K^2 -fold multiway cross fitting. Note that, for each $(k, \ell) \in [K]^2$, the nuisance parameter estimate $\hat{\eta}_{k\ell}$ is computed using the subsample of those observations with multiway indices $(i, j) \in ([N] \setminus I_k) \times ([M] \setminus J_\ell)$, and in turn the score term $\mathbb{E}_{n,k\ell}[\psi(W; \cdot, \hat{\eta}_{k\ell})]$ is computed using the subsample of those observations with multiway indices $(i, j) \in I_k \times J_\ell$. This two-step computation is repeated K^2 times for every partitioning pair $(k, \ell) \in [K]^2$. Figure 1 illustrates this K^2 -fold cross fitting for the case of $K = 2$ and $N = M = 4$, where the cross fitting repeats for $K^2 (= 2^2 = 4)$ times.

Remark 1. *This estimator is a multiway-counterpart of DML2 in CCDDHNR (2018). It is also possible to consider the multiway-counterpart of their DML1. With this said, we focus on this current estimator following their simulation finding that DML2 outperforms their DML1 in most situation settings due to the stability of the score function.*

Remark 2 (Higher Cluster Dimensions). *When we have α -way clustering for an integer $\alpha > 2$, the above algorithm can be easily generalized into a K^α -fold multiway DML estimator following the general notation of Davezies et al. (2018).*

We propose to estimate the asymptotic variance of $\sqrt{\underline{C}}(\tilde{\theta} - \theta_0)$ by

$$\hat{\sigma}^2 = \hat{J}^{-1} \hat{\Gamma} (\hat{J}^{-1})', \quad (2.6)$$

where $\hat{\Gamma}$ and \hat{J} are given by

$$\begin{aligned} \hat{\Gamma} = & \frac{1}{K^2} \sum_{(k,\ell) \in [K]^2} \left\{ \frac{|I| \wedge |J|}{(|I||J|)^2} \sum_{i \in I_k} \sum_{j, j' \in J_\ell} \psi(W_{ij}; \tilde{\theta}, \hat{\eta}_{k\ell}) \psi(W_{ij'}; \tilde{\theta}, \hat{\eta}_{k\ell})' \right. \\ & \left. + \frac{|I| \wedge |J|}{(|I||J|)^2} \sum_{i, i' \in I_k} \sum_{j \in J_\ell} \psi(W_{ij}; \tilde{\theta}, \hat{\eta}_{k\ell}) \psi(W_{i'j}; \tilde{\theta}, \hat{\eta}_{k\ell})' \right\} \quad \text{and} \\ \hat{J} = & \frac{1}{K^2} \sum_{(k,\ell) \in [K]^2} \mathbb{E}_{n,k\ell}[\psi^a(W; \hat{\eta}_{k\ell})]. \end{aligned}$$

For a d_θ -dimensional vector r , the $(1 - a)$ confidence interval for the linear functional $r'\theta_0$ can be constructed by

$$\text{CI}_a := [r'\tilde{\theta} \pm \Phi^{-1}(1 - a/2) \sqrt{r'\hat{\sigma}^2 r / \underline{C}}].$$

2.3 Example: Partially Linear IV Model with Multiway Cluster Sample

For an illustration, consider as a concrete example the partially linear IV model (cf. CCDDHNR, 2018, Section 4.2) adapted to the multiway cluster sample data:

$$Y_{ij} = D_{ij}\theta_0 + g_0(X_{ij}) + \epsilon_{ij}, \quad \mathbb{E}_P[\epsilon_{ij}|X_{ij}, Z_{ij}] = 0, \quad (2.7)$$

$$Z_{ij} = m_0(X_{ij}) + v_{ij}, \quad \mathbb{E}_P[v_{ij}|X_{ij}] = 0. \quad (2.8)$$

A researcher observes the random variables Y_{ij} , D_{ij} , X_{ij} , and Z_{ij} , which are typically interpreted as the outcome, endogenous regressor, exogenous regressors, and instrumental variable, respectively. The low-dimensional parameter vector θ_0 is an object of interest.

A Neyman orthogonal score ψ is given by

$$\psi(w; \theta, \eta) = (y - g_1(x) - \theta(d - g_2(x)))(z - m(x)) \quad (2.9)$$

as in CCDDHNR (2018), where $w = (y, d, x, z)$, $\eta = (g_1, g_2, m)$ and $g_1, g_2, m \in L^2(P)$. It is straightforward to verify that this score satisfies both the moment restriction (2.1) $E_P[\psi(W_{11}; \theta_0, \eta_0)] = 0$ and the Neyman orthogonality condition (2.2) $\partial_\eta E_P \psi(W_{11}; \theta_0, \eta_0)[\eta - \eta_0] = 0$ for all $\eta \in \mathcal{T}_n$ at $\eta_0 = (g_{10}, g_{20}, m_0)$, where $g_{10}(X) = E_P[Y|X]$, $g_{20}(X) = E_P[D|X]$, and $m_0(X) = E_P[Z|X]$.

Remark 3. *This partially linear IV model nests several commonly used high-dimensional econometric models when nuisance parameter estimates are based on lasso or post-lasso. For example, it reduces to the many IV models studied in Belloni, Chen, Chernozhukov, and Hansen (2012) if we set $g_{10} = 0$. For another example, it reduces to the high-dimensional linear regression model studied in Belloni et al. (2014b) and others if we set $Z = D$.*

The following algorithm is our proposed multiway DML procedure introduced in Section 2.2, specifically applied to this partially linear IV model.

Algorithm 1 (K^2 -fold Multiway DML for partially linear IV Model with Lasso).

1. Randomly partition $[N]$ into K parts $\{I_1, \dots, I_K\}$ and $[M]$ into K parts $\{J_1, \dots, J_K\}$.

2. For each $(k, \ell) \in [K]^2$:

(a) Run a lasso of Y on X to obtain $\hat{g}_{1,k\ell}(x) = x' \hat{\beta}_{k\ell}$ using observations from $I_k^c \times J_\ell^c$.

(b) Run a lasso of D on X to obtain $\hat{g}_{2,k\ell}(x) = x' \hat{\gamma}_{k\ell}$ using observations from $I_k^c \times J_\ell^c$.

(c) Run a lasso of Z on X to obtain $\hat{m}_{k\ell}(x) = x' \hat{\xi}_{k\ell}$ using observations from $I_k^c \times J_\ell^c$.

3. Solve the equation

$$\frac{1}{K^2} \sum_{(k,\ell) \in [K]^2} \mathbb{E}_{n,k\ell}[(Y_{ij} - X'_{ij} \hat{\beta}_{k\ell} - \theta(D_{ij} - X'_{ij} \hat{\gamma}_{k\ell}))(Z_{ij} - X'_{ij} \hat{\xi}_{k\ell})] = 0$$

for θ to obtain the multiway DML estimate $\tilde{\theta}$.

4. Let $\widehat{\varepsilon}_{ij} = Y_{ij} - X'_{ij}\widehat{\beta}_{k\ell} - \widetilde{\theta}(D_{ij} - X'_{ij}\widehat{\gamma}_{k\ell})$, $\widehat{u}_{ij} = D_{ij} - X'_{ij}\widehat{\gamma}_{k\ell}$, and $\widehat{v}_{ij} = Z_{ij} - X'_{ij}\widehat{\xi}_{k\ell}$ for each $(i, j) \in I_k \times J_\ell$ for each $(k, \ell) \in [K]^2$, and let the multiway DML asymptotic variance estimator be given by

$$\widehat{\sigma}^2 = \widehat{J}^{-1} \frac{1}{K^2} \sum_{k=1}^K \sum_{\ell=1}^K \left\{ \frac{|I| \wedge |J|}{(|I||J|)^2} \sum_{i \in I_k} \sum_{j, j' \in J_\ell} \widehat{\varepsilon}_{ij} \widehat{v}_{ij} \widehat{v}_{ij'} \widehat{\varepsilon}_{ij'} + \frac{|I| \wedge |J|}{(|I||J|)^2} \sum_{i, i' \in I_k} \sum_{j \in J_\ell} \widehat{\varepsilon}_{ij} \widehat{v}_{ij} \widehat{v}_{i'j} \widehat{\varepsilon}_{i'j} \right\} (\widehat{J}^{-1})',$$

where

$$\widehat{J} = - \frac{1}{K^2} \sum_{k=1}^K \sum_{\ell=1}^K \mathbb{E}_{n, k\ell} [\widehat{u}_{ij} \widehat{v}_{ij}].$$

5. Report the estimate $\widetilde{\theta}$, its standard error $\sqrt{\widehat{\sigma}^2/\underline{C}}$, and/or the $(1 - a)$ confidence interval

$$CI_a := \left[\widetilde{\theta} \pm \Phi^{-1}(1 - a/2) \sqrt{\widehat{\sigma}^2/\underline{C}} \right].$$

For the sake of concreteness, we present this algorithm specifically based on lasso (in the three sub-steps under step 2), but another machine learning method (e.g., post-lasso, elastic nets, ridge, deep neural networks, and boosted trees) may be substituted for lasso.

Example 1 (Demand Analysis). *Consider the model of Berry (1994) in which consumer c derives the utility*

$$\delta_{ij} + X_{ij}\alpha_c + \varepsilon_{cij}$$

from choosing product i in market j , where ε_{cij} independently follows the Type I Extreme Value distribution, α_c is a random coefficient, and the mean utility δ_{ij} takes the linear-index form

$$\delta_{ij} = D_{ij}\theta_0 + \epsilon_{ij}.$$

In this framework, Lu, Shi, and Tao (2019, Equation (9)) derive the partial-linear equation

$$Y_{ij} = D_{ij}\theta_0 + g_0(X_{ij}) + \epsilon_{ij}$$

for estimation of θ_0 , where $Y_{ij} = \log(S_{ij}) - \log(S_{0j})$ denotes the observed log share of product i relative to the log of the outside share. Since D_{ij} usually consists of the endogenous price of product i in market j , researchers often use instruments Z_{ij} such that $\mathbb{E}_P[\epsilon_{ij}|X_{ij}, Z_{ij}] = 0$. This yields the reduced-form

equation (2.7), together with the innocuous (i.e., nonparametric) projection equation (2.8). Since the variables Y_{ij} , D_{ij} , X_{ij} , and Z_{ij} are double-indexed by product i and market j , the sample naturally entails two-way dependence. As such, instead of using standard errors based on i.i.d. sampling, we recommend that a researcher uses the two-way cluster-robust standard error based on Algorithm 1. \triangle

3 Theory of the Multiway DML

In this section, we present formal theories to guarantee that the multiway DML method proposed in Section 2 works. We first fix some notations for convenience. The two-way sample sizes $(N, M) \in \mathbb{N}^2$ will be indexed by a single index $n \in \mathbb{N}$ as $(N, M) = (N(n), M(n))$ where $M(n)$ and $N(n)$ are non-decreasing in n and $M(n)N(n)$ is increasing in n . With this said, we will suppress the index notation and write (N, M) for simplicity of notations. Let $\{\mathcal{P}_n\}_n$ be a sequence of sets of probability laws of $\{W_{ij}\}_{ij}$ – note that we allow for increasing dimensionality of W_{ij} in the sample size n . Let $P = P_n \in \mathcal{P}_n$ denote the law with respect to sample size (N, M) . Throughout, we assume that this random vector W_{ij} is Borel measurable. Recall the notations $\underline{C} = N \wedge M$, $\mu_N = \underline{C}/N$, and $\mu_M = \underline{C}/M$, and suppose that $\mu_N \rightarrow \bar{\mu}_N$, $\mu_M \rightarrow \bar{\mu}_M$. We write $a \lesssim b$ to mean $a \leq cb$ for some $c > 0$ that does not depend on n . We also write $a \lesssim_P b$ to mean $a = O_P(b)$. For any finite dimensional vector v , $\|v\|$ denotes the ℓ_2 or Euclidean norm of v . For any matrix A , $\|A\|$ denotes the induced ℓ_2 -norm of the matrix.

We state the following assumption on multiway clustered sampling.

Assumption 1 (Sampling). *Suppose $\underline{C} \rightarrow \infty$. The following conditions hold for each n .*

(i) $(W_{ij})_{(i,j) \in \mathbb{N}^2}$ is an infinite sequence of separately exchangeable p -dimensional random vectors.

That is, for any permutations π_1 and π_2 of \mathbb{N} , we have

$$(W_{ij})_{(i,j) \in \mathbb{N}^2} \stackrel{d}{=} (W_{\pi_1(i)\pi_2(j)})_{(i,j) \in \mathbb{N}^2}.$$

(ii) $(W_{ij})_{(i,j) \in \mathbb{N}^2}$ is dissociated. That is, for any $(c_1, c_2) \in \mathbb{N}^2$, $(W_{ij})_{i \in [c_1], j \in [c_2]}$ is independent of

$$(W_{ij})_{i \in [c_1]^c, j \in [c_2]^c}.$$

(iii) For each n , an econometrician observes $(W_{ij})_{i \in [N], j \in [M]}$.

Recall that we focus on the linear Neyman orthogonal score of the form

$$\psi(w; \theta, \eta) = \psi^a(w; \eta)\theta + \psi^b(w; \eta), \text{ for all } w \in \text{supp}(W), \theta \in \Theta, \eta \in T.$$

Let $c_0 > 0$, $c_1 > 0$, $s > 0$, $q \geq 4$ be some finite constants with $c_0 \leq c_1$. Let $\{\delta_n\}_{n \geq 1}$ (estimation errors) and $\{\Delta_n\}_{n \geq 1}$ (probability bounds) be sequences of positive constants that converge to zero such that $\delta_n \geq \underline{C}^{-1/2}$. Let $K \geq 2$ be a fixed integer. Let W_{00} denote an independent copy of W_{11} and therefore is independent from the data and the random set \mathcal{T}_n of nuisance realization. With these notations, we consider the following assumptions.

Assumption 2 (Linear Neyman Orthogonal Score). *For $\underline{C} \geq 3$ and $P \in \mathcal{P}_n$, the following conditions hold.*

- (i) *The true parameter value θ_0 satisfies (2.1).*
- (ii) *ψ is linear in the sense that it satisfies (2.4).*
- (iii) *The map $\eta \mapsto \mathbb{E}_P[\psi(W_{00}; \theta, \eta)]$ is twice continuously Gateaux differentiable on T .*
- (iv) *ψ satisfies either the Neyman orthogonality condition (2.2) or more generally the Neyman λ_n near orthogonality condition at (θ_0, η_0) with respect to a nuisance realization set $\mathcal{T}_n \subset T$ as*

$$\lambda_n := \sup_{\eta \in \mathcal{T}_n} \left\| \partial_\eta \mathbb{E}_P \psi(W_{00}; \theta_0, \eta_0) [\eta - \eta_0] \right\| \leq \delta_n \underline{C}^{-1/2}.$$

- (v) *The identification condition holds as the singular values of the matrix $J_0 := \mathbb{E}_P[\psi^a(W_{11}; \eta_0)]$ are between c_0 and c_1 .*

Assumption 3 (Score Regularity and Nuisance Parameter Estimators). *For all $\underline{C} \geq 3$ and $P \in \mathcal{P}_n$, the following conditions hold.*

- (i) *Given random subsets $I \subset [N]$ and $J \subset [M]$ such that $|I| \times |J| = \lfloor NM/K^2 \rfloor$, the nuisance parameter estimator $\hat{\eta} = \hat{\eta}((W_{ij})_{(i,j) \in I^c \times J^c})$, where the complements are taken with respect to $[N]$ and $[M]$, respectively, belongs to the realization set \mathcal{T}_n with probability at least $1 - \Delta_n$, where \mathcal{T}_n contains η_0 .*

(ii) The following moment conditions hold:

$$m_n := \sup_{\eta \in \mathcal{T}_n} (\mathbb{E}_P[\|\psi(W_{00}; \theta_0, \eta)\|^q])^{1/q} \leq c_1,$$

$$m'_n := \sup_{\eta \in \mathcal{T}_n} (\mathbb{E}_P[\|\psi^a(W_{00}; \eta)\|^q])^{1/q} \leq c_1.$$

(iii) The following conditions on the rates r_n , r'_n and λ'_n hold:

$$r_n := \sup_{\eta \in \mathcal{T}_n} \|\mathbb{E}_P[\psi^a(W_{00}; \eta)] - \mathbb{E}_P[\psi^a(W_{00}; \eta_0)]\| \leq \delta_n,$$

$$r'_n := \sup_{\eta \in \mathcal{T}_n} (\|\mathbb{E}_P[\psi(W_{00}; \theta_0, \eta)] - \mathbb{E}_P[\psi(W_{00}; \theta_0, \eta_0)]\|^2)^{1/2} \leq \delta_n,$$

$$\lambda'_n = \sup_{r \in (0,1), \eta \in \mathcal{T}_n} \|\partial_r^2 \mathbb{E}_P[\psi(W_{00}; \theta_0, \eta_0 + r(\eta - \eta_0))]\| \leq \delta_n / \sqrt{\underline{C}}.$$

(iv) All eigenvalues of the matrix

$$\Gamma := \bar{\mu}_N \Gamma_N + \bar{\mu}_M \Gamma_M = \bar{\mu}_N \mathbb{E}_P[\psi(W_{11}; \theta_0, \eta_0) \psi(W_{12}; \theta_0, \eta_0)'] + \bar{\mu}_M \mathbb{E}_P[\psi(W_{11}; \theta_0, \eta_0) \psi(W_{21}; \theta_0, \eta_0)'].$$

are bounded from below by c_0 .

Remark 4 (Discussion of the Assumptions). *Assumption 1 is exactly the same as Assumption 1 of Chiang and Sasaki (2019), which is closely related to Assumption 1 of Davezies et al. (2018). Assumption 2 is closely related to Assumptions 3.1 of CCDDHNR (2018). It requires the score to be Neyman near orthogonal – see their Section 2.2.1. for the procedure of orthogonalizing a non-orthogonal score. It also imposes some mild smoothness and identification conditions. Assumption 3 corresponds to Assumption 3.2 of CCDDHNR (2018). It imposes some high level conditions on the quality of the nuisance parameter estimator as well as the non-degeneracy of the asymptotic variance. This rules out the degenerate cases such as Example 1.6 of Menzel (2017).*

Remark 5 (Partial Distributions). *Assumptions 2 and 3 state conditions based on W_{00} , differently from CCDDHNR (2018), because of the need to deal with dependent observations in cross fitting in our multiway DML framework.*

The following result presents the main theorem of this paper, establishing the linear representation and asymptotic normality of the multiway DML estimator. It corresponds to Theorem 3.1 of CCDDHNR (2018), and is an extension of it to the case of multiway cluster sampling.

Theorem 1 (Main Result). *Suppose that Assumptions 1, 2 and 3 are satisfied. If $\delta_n \geq \underline{C}^{-1/2}$ for all $\underline{C} \geq 1$, then*

$$\sqrt{\underline{C}}\sigma^{-1}(\tilde{\theta} - \theta_0) = \frac{\sqrt{\underline{C}}}{NM} \sum_{i=1}^N \sum_{j=1}^M \bar{\psi}(W_{ij}) + O_P(\rho_n) \rightsquigarrow N(0, I_{d_\theta})$$

holds uniformly over $P \in \mathcal{P}_n$, where the size of the remainder terms follows

$$\rho_n := \underline{C}^{-1/2} + r_n + r'_n + \underline{C}^{1/2}\lambda_n + \underline{C}^{1/2}\lambda'_n \lesssim \delta_n,$$

the influence function takes the form $\bar{\psi}(\cdot) := -\sigma^{-1}J_0^{-1}\psi(\cdot; \theta_0, \eta_0)$, and the approximate variance is given by

$$\sigma^2 := J_0^{-1}\Gamma(J_0^{-1})'. \quad (3.1)$$

As is commonly the case in practice, we need to estimate the unknown asymptotic variance. The following theorem shows the validity of our proposed multiway DML variance estimator.

Theorem 2 (Variance Estimator). *Under the assumptions required by Theorem 1, we have*

$$\hat{\sigma}^2 = \sigma^2 + O_P(\rho_n).$$

Furthermore, the statement of Theorem 1 holds true with $\hat{\sigma}^2$ in place of σ^2 .

Theorems 1 and 2 can be used for constructing confidence intervals.

Corollary 1. *Suppose that all the Assumptions required by Theorem 1 are satisfied. Let r be a d_θ -dimensional vector. The $(1 - a)$ confidence interval of $r'\theta_0$ given by*

$$CI_a := [r'\tilde{\theta} \pm \Phi^{-1}(1 - a/2)\sqrt{r'\hat{\sigma}^2 r / \underline{C}}]$$

satisfies

$$\sup_{P \in \mathcal{P}_n} |P_P(\theta_0 \in CI_a) - (1 - a)| \rightarrow 0.$$

As in Section 3.4 of CCDDHNR (2018), we can also repeatedly compute multiway DML estimates and variance estimates S -times for some fixed $S \in \mathbb{N}$ and consider the average or median of the estimates as the new estimate. This does not have an asymptotic impact, yet it can reduce the impact of a random sample splitting on the estimate.

4 Simulation Studies

4.1 Simulation Setup

Consider the partially linear IV model introduced in Section 2.3. We specifically focus on the following high-dimensional linear representations

$$\begin{aligned} Y_{ij} &= D_{ij}\theta_0 + X'_{ij}\zeta_0 + \epsilon_{ij} \\ D_{ij} &= Z_{ij}\pi_{10} + X'_{ij}\pi_{20} + v_{ij}, \\ Z_{ij} &= X'_{ij}\xi_0 + V_{ij}, \end{aligned}$$

where the parameter values are set to $\theta_0 = \pi_{10} = 1.0$ and $\zeta_0 = \pi_{20} = \xi_0 = (0.5, .0.5^2, \dots, 0.5^{\dim(X)})'$ for some large $\dim(X)$. The primitive random vector $(X'_{ij}, \epsilon_{ij}, v_{ij}, V_{ij})'$ is constructed by

$$\begin{aligned} X_{ij} &= (1 - \omega_1^X - \omega_2^X)\alpha_{ij}^X + \omega_1^X\alpha_i^X + \omega_2^X\alpha_j^X, \\ \epsilon_{ij} &= (1 - \omega_1^\epsilon - \omega_2^\epsilon)\alpha_{ij}^\epsilon + \omega_1^\epsilon\alpha_i^\epsilon + \omega_2^\epsilon\alpha_j^\epsilon, \\ v_{ij} &= (1 - \omega_1^v - \omega_2^v)\alpha_{ij}^v + \omega_1^v\alpha_i^v + \omega_2^v\alpha_j^v, \quad \text{and} \\ V_{ij} &= (1 - \omega_1^V - \omega_2^V)\alpha_{ij}^V + \omega_1^V\alpha_i^V + \omega_2^V\alpha_j^V \end{aligned}$$

with two-way clustering weights (ω_1^X, ω_2^X) , $(\omega_1^\epsilon, \omega_2^\epsilon)$, (ω_1^v, ω_2^v) , and (ω_1^V, ω_2^V) , where α_{ij}^X , α_i^X , and α_j^X are independently generated according to

$$\alpha_{ij}^X, \alpha_i^X, \alpha_j^X \sim N \left(0, \begin{pmatrix} \begin{pmatrix} s_X^0 & s_X^1 & \cdots & s_X^{\dim(X)-2} & s_X^{\dim(X)-1} \\ s_X^1 & s_X^0 & \cdots & s_X^{\dim(X)-3} & s_X^{\dim(X)-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ s_X^{\dim(X)-2} & s_X^{\dim(X)-3} & \cdots & s_X^0 & s_X^1 \\ s_X^{\dim(X)-1} & s_X^{\dim(X)-2} & \cdots & s_X^1 & s_X^0 \end{pmatrix} \end{pmatrix} \right),$$

$(\alpha_{ij}^\epsilon, \alpha_{ij}^v)'$, $(\alpha_i^\epsilon, \alpha_i^v)'$, and $(\alpha_j^\epsilon, \alpha_j^v)'$ are independently generated according to

$$\begin{pmatrix} \alpha_{ij}^\epsilon \\ \alpha_{ij}^v \end{pmatrix}, \begin{pmatrix} \alpha_i^\epsilon \\ \alpha_i^v \end{pmatrix}, \begin{pmatrix} \alpha_j^\epsilon \\ \alpha_j^v \end{pmatrix} \sim N \left(0, \begin{pmatrix} 1 & s_{\epsilon v} \\ s_{\epsilon v} & 1 \end{pmatrix} \right),$$

and α_{ij}^V , α_i^V , and α_j^V are independently generated according to

$$\alpha_{ij}^V, \alpha_i^V, \alpha_j^V \sim N(0, 1).$$

The weights (ω_1^X, ω_2^X) , $(\omega_1^\epsilon, \omega_2^\epsilon)$, (ω_1^v, ω_2^v) , and (ω_1^V, ω_2^V) specify the extent of dependence in two-way clustering in X_{ij} , ϵ_{ij} , v_{ij} , and V_{ij} , respectively. The parameter s_X specifies the extent of collinearity among the high-dimensional regressors X_{ij} . The parameter $s_{\epsilon v}$ specifies the extent of endogeneity. We set the values of these parameters to $(\omega_1^X, \omega_2^X) = (\omega_1^\epsilon, \omega_2^\epsilon) = (\omega_1^v, \omega_2^v) = (\omega_1^V, \omega_2^V) = (0.25, 0.25)$ and $s_X = s_{\epsilon v} = 0.25$.

4.2 Results

Monte Carlo simulations are conducted with 2,500 iterations for each set. Table 1 reports simulation results. The first four columns in the table indicate the data generating process (N , M , \underline{C} , and $\dim(X)$). The next column indicates the integer K for our K^2 -fold cross fitting method. We use $K = 2$ and 3 in the simulations for the displayed results, since $2^2 (\approx 5)$ and $3^2 (\approx 10)$ are close to the common numbers of folds used in cross fitting in practice. The next column indicates the machine learning method for estimation of $\hat{\eta}_{k\ell}$. We use the ridge, elastic net, and lasso. The last four columns of the table report Monte Carlo simulation statistics, including the bias (Bias), standard deviation (SD), root mean square error (RMSE), and coverage frequency for the nominal probability of 95% (Cover).

For each covariate dimension $\dim(X) \in \{100, 200\}$, for each choice $K \in \{2, 3\}$ for the number K^2 of multiway cross fitting, and for each of the three machine learning methods, we observe the following patterns as the effective sample size $\underline{C} = N \wedge M$ increases: 1) the bias tends to zero; 2) the standard deviation decreases approximately at the $\sqrt{\underline{C}}$ rate; and 3) the coverage frequency converges to the nominal probability. These results confirm the theoretical properties of the proposed method. We ran several other sets of simulations besides those displayed in the table, and this pattern remains the same across different sets.

Comparing the results across the three machine learning methods, we observe that the ridge entails larger bias and smaller variance relative to the elastic net and lasso in finite sample. This makes the coverage frequency of the ridge less accurate compared with the elastic net and lasso. On one hand, the choice $K = 3$ (i.e., 9-fold) of the multiway cross fitting contributes to mitigating the large bias of the ridge relative to the choice $K = 2$, and hence $K = 3$ produces more preferred results for the ridge. On the other hand, the choice $K = 2$ tends to yield preferred results in terms of coverage accuracy for

the elastic net and lasso. In light of these results, we recommend the elastic net or lasso along with the use of 2^2 - fold (i.e., 4-fold) cross fitting. This number of folds in cross fitting is in fact similar to that recommended by CCDDHNR (2018) for i.i.d. sampling – see their Remark 3.1 where they recommend 4- or 5-fold cross fitting.

5 Conclusion

In this paper, we propose a multiway DML procedure based on a new multiway cross fitting algorithm. This multiway DML procedure is valid in the presence of multiway cluster sampled data, which is frequently used in empirical research. We present an asymptotic theory showing that multiway DML is valid under nearly identical regularity conditions to those of CCDDHNR (2018). The proposed method covers a large class of econometric models as is the case with CCDDHNR (2018), and is compatible with various machine learning based estimation methods. Simulation studies indicate that the proposed procedure has attractive finite sample performance under various multiway cluster sampling environments for various machine learning methods. To accompany the theoretical findings, we provide easy-to-implement algorithms for multiway DML. Such algorithms are readily implementable using existing statistical packages.

Appendix

A Proofs of the Main Results

For any $(i, j) \in I_k \times J_\ell$, we use the shorthand notation $E_P[f(W_{ij})|I_k^c \times J_\ell^c]$ to denote the conditional expectation $E_P[f(W_{ij})|(W_{i'j'})_{(i',j') \in ([N] \setminus I_k) \times ([M] \setminus J_\ell)}]$ whenever one exists.

A.1 Proof of Theorem 1

Proof. In this proof we try to follow as parallelly as possible the five steps of the proof of Theorem 3.1 of CCDDHNR (2018) although all the asymptotic arguments are properly modified to account for multiway cluster sampling.

Denote \mathcal{E}_n for the event $\hat{\eta}_{k\ell} \in \mathcal{T}_n$ for all $k, \ell \in [K]^2$. Assumption 3 (i) implies $P(\mathcal{E}_n) \geq 1 - K^2 \Delta_n$.

Step 1. This is the main step showing linear representation and asymptotic normality for the proposed estimator. Denote

$$\begin{aligned}\hat{J} &:= \frac{1}{K^2} \sum_{(k,\ell) \in [K]^2} \mathbb{E}_{n,k\ell}[\psi^a(W; \hat{\eta}_{k\ell})], & R_{n,1} &:= \hat{J} - J_0, \\ R_{n,2} &:= \frac{1}{K^2} \sum_{(k,\ell) \in [K]^2} \mathbb{E}_{n,k\ell}[\psi(W; \theta_0, \hat{\eta}_{k\ell})] - \frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M \psi(W_{ij}; \theta_0, \eta_0).\end{aligned}$$

We will later show in Steps 2, 3, 4 and 5, respectively, that

$$\|R_{n,1}\| = O_{P_n}(\underline{C}^{-1/2} + r_n), \quad (\text{A.1})$$

$$\|R_{n,2}\| = O_{P_n}(\underline{C}^{-1/2} r'_n + \lambda_n + \lambda'_n), \quad (\text{A.2})$$

$$\left\| \sqrt{\underline{C}} (NM)^{-1} \sum_{i=1}^N \sum_{j=1}^M \psi(W_{ij}; \theta_0, \eta_0) \right\| = O_{P_n}(1), \quad (\text{A.3})$$

$$\|\sigma^{-1}\| = O_{P_n}(1). \quad (\text{A.4})$$

Then, under Assumptions 2 and 3, $\underline{C}^{-1/2} + r_n \leq \rho_n = o(1)$ and all singular values of J_0 are bounded away from zero. Therefore, with P_n -probability at least $1 - o(1)$, all singular values of \hat{J} are bounded away from zero. Thus with the same P_n probability, the multiway DML solution is uniquely written as

$$\tilde{\theta} = -\hat{J}^{-1} \frac{1}{K^2} \sum_{(k,\ell) \in [K]^2} \mathbb{E}_{n,k\ell}[\psi^b(W; \hat{\eta}_{k\ell})],$$

and

$$\begin{aligned}\sqrt{\underline{C}}(\tilde{\theta} - \theta_0) &= -\sqrt{\underline{C}}\hat{J}^{-1} \frac{1}{K^2} \sum_{(k,\ell) \in [K]^2} \left(\mathbb{E}_{n,k\ell}[\psi^b(W; \hat{\eta}_{k\ell})] + \hat{J}\theta_0 \right) \\ &= -\sqrt{\underline{C}}\hat{J}^{-1} \frac{1}{K^2} \sum_{(k,\ell) \in [K]^2} \mathbb{E}_{n,k\ell}[\psi(W; \theta_0, \hat{\eta}_{k\ell})] \\ &= -\left(J_0 + R_{n,1}\right)^{-1} \times \left(\frac{\sqrt{\underline{C}}}{NM} \sum_{i=1}^N \sum_{j=1}^M \psi(W_{ij}; \theta_0, \eta_0) + \sqrt{\underline{C}}R_{n,2} \right).\end{aligned} \quad (\text{A.5})$$

Using the fact that

$$\left(J_0 + R_{n,1}\right)^{-1} - J_0^{-1} = -(J_0 + R_{n,1})^{-1} R_{n,1} J_0^{-1},$$

we have

$$\begin{aligned}\|(J_0 + R_{n,1})^{-1} - J_0^{-1}\| &= \|(J_0 + R_{n,1})^{-1} R_{n,1} J_0^{-1}\| \leq \|(J_0 + R_{n,1})^{-1}\| \|R_{n,1}\| \|J_0^{-1}\| \\ &= O_{P_n}(1) O_{P_n}(\underline{C}^{-1} + r_n) O_{P_n}(1) = O_{P_n}(\underline{C}^{-1} + r_n).\end{aligned}$$

Furthermore, $r'_n + \sqrt{\underline{C}}(\lambda_n + \lambda'_n) \leq \rho_n = o(1)$, it holds that

$$\begin{aligned}\left\| \frac{\sqrt{\underline{C}}}{NM} \sum_{i=1}^N \sum_{j=1}^M \psi(W_{ij}; \theta_0, \eta_0) + \sqrt{\underline{C}} R_{n,2} \right\| &\leq \left\| \frac{\sqrt{\underline{C}}}{NM} \sum_{i=1}^N \sum_{j=1}^M \psi(W_{ij}; \theta_0, \eta_0) \right\| + \left\| \sqrt{\underline{C}} R_{n,2} \right\| \\ &= O_{P_n}(1) + o_{P_n}(1) = O_{P_n}(1),\end{aligned}$$

where the first equality is due to (A.3) and (A.4). Combining above two bounds gives

$$\begin{aligned}\left\| (J_0 + R_{n,1})^{-1} - J_0^{-1} \right\| \times \left\| \frac{\sqrt{\underline{C}}}{NM} \sum_{i=1}^N \sum_{j=1}^M \psi(W_{ij}; \theta_0, \eta_0) + \sqrt{\underline{C}} R_{n,2} \right\| &= O_{P_n}(\underline{C}^{-1} + r_n) O_{P_n}(1) \\ &= O_{P_n}(\underline{C}^{-1} + r_n).\end{aligned}\tag{A.6}$$

Therefore, from (A.4), (A.5) and (A.6), we have

$$\sqrt{\underline{C}} \sigma^{-1} (\tilde{\theta} - \theta_0) = \frac{\sqrt{\underline{C}}}{NM} \sum_{i=1}^N \sum_{j=1}^M \bar{\psi}(W_{ij}) + O_{P_n}(\rho_n).$$

The first term on the RHS above can be written as $\mathbb{G}_{\underline{C}} \bar{\psi}$. Applying Lemma 1, we obtain the independent linear representation

$$H_n \bar{\psi} := \sum_{i=1}^N \frac{\sqrt{\underline{C}}}{N} \mathbb{E}_{P_n}[\bar{\psi}(W_{ij}) | U_{i0}] + \sum_{j=1}^M \frac{\sqrt{\underline{C}}}{M} \mathbb{E}_{P_n}[\bar{\psi}(W_{ij}) | U_{0j}]$$

and it holds P_n -a.s. that

$$\begin{aligned}V(\mathbb{G}_{\underline{C}} \bar{\psi}) &= V(H_n \bar{\psi}) + O(\underline{C}^{-1}) = J_0^{-1} \Gamma(J_0^{-1})' + O(\underline{C}^{-1}) \quad \text{and} \\ \mathbb{G}_{\underline{C}} \bar{\psi} &= H_n \bar{\psi} + O_P(\underline{C}^{-1/2})\end{aligned}$$

under Assumption 3 (iv). Recall that $q \geq 4$, the third moments of both summands of $H_n \bar{\psi}$ are bounded over n under Assumptions 2(v) and 3 (ii) (iv). We have verified all the conditions for Lyapunov's CLT. An application of Lyapunov's CLT and Cramer-Wold device gives

$$H_n \bar{\psi} \rightsquigarrow N(0, I_{d_\theta})$$

and an application of Theorem 2.7 of van der Vaart (1998) concludes the proof.

Step 2. Since K is fixed, it suffices to show for any $(k, \ell) \in [K]^2$,

$$\left\| \mathbb{E}_{n, k\ell}[\psi^a(W; \hat{\eta}_{k\ell})] - \mathbb{E}_P[\psi^a(W_{11}; \eta_0)] \right\| = O_{P_n}(\underline{C}^{-1/2} + r_n).$$

Fix $(k, \ell) \in [K]^2$,

$$\left\| \mathbb{E}_{n, k\ell}[\psi^a(W; \hat{\eta}_{k\ell})] - \mathbb{E}_{P_n}[\psi^a(W_{ij}; \eta_0)] \right\| \leq \mathcal{I}_{1, k\ell} + \mathcal{I}_{2, k\ell}.$$

where

$$\begin{aligned} \mathcal{I}_{1, k\ell} &= \left\| \mathbb{E}_{n, k\ell}[\psi^a(W; \hat{\eta}_{k\ell})] - \mathbb{E}_{P_n}[\psi^a(W_{ij}; \hat{\eta}_{k\ell}) | I_k^c \times J_\ell^c] \right\| \\ \mathcal{I}_{2, k\ell} &= - \left\| \mathbb{E}_{P_n}[\psi^a(W_{ij}; \hat{\eta}_{k\ell}) | I_k^c \times J_\ell^c] - \mathbb{E}_{P_n}[\psi^a(W_{11}; \eta_0)] \right\|. \end{aligned}$$

Notice that $\mathcal{I}_{2, k\ell} \leq r_n$ with P_n -probability $1 - o(1)$ follows directly from Assumptions 1 (ii) and 3 (iii).

Now denote $\tilde{\psi}_{ij, m}^a = \psi_m^a(W_{ij}; \hat{\eta}_{k\ell}) - \mathbb{E}_{P_n}[\psi_m^a(W_{ij}; \hat{\eta}_{k\ell}) | I_k^c \times J_\ell^c]$ and $\tilde{\psi}_{ij}^a = (\tilde{\psi}_{ij, m}^a)_{m \in [d_\theta]}$. To bound $\mathcal{I}_{1, k\ell}$,

note that conditional on $I_k^c \times J_\ell^c$, it holds that

$$\begin{aligned}
\mathbb{E}_{P_n}[\mathcal{I}_{1,k\ell}^2 | I_k^c \times J_\ell^c] &= \mathbb{E}_{P_n} \left[\left\| \mathbb{E}_{n,k\ell}[\psi^a(W; \hat{\eta}_{k\ell})] - \mathbb{E}_{P_n}[\psi^a(W_{ij}; \hat{\eta}_{k\ell}) | I_k^c \times J_\ell^c] \right\|^2 \middle| I_k^c \times J_\ell^c \right] \\
&= \frac{1}{(|I||J|)^2} \mathbb{E}_{P_n} \left[\sum_{m=1}^{d_\theta} \left(\sum_{(i,j) \in I_k \times J_\ell} \tilde{\psi}_{ij,m}^a \right)^2 \middle| I_k^c \times J_\ell^c \right] \\
&= \frac{1}{(|I||J|)^2} \sum_{(i,j) \in I_k \times J_\ell} \sum_{j' \in J_\ell, j' \neq j} \mathbb{E}_{P_n} \left[\sum_{m=1}^{d_\theta} \tilde{\psi}_{ij,m}^a \tilde{\psi}_{ij',m}^a \middle| I_k^c \times J_\ell^c \right] \\
&\quad + \frac{1}{(|I||J|)^2} \sum_{(i,j) \in I_k \times J_\ell} \sum_{i' \in I_k, i' \neq i} \mathbb{E}_{P_n} \left[\sum_{m=1}^{d_\theta} \tilde{\psi}_{ij,m}^a \tilde{\psi}_{i'j,m}^a \middle| I_k^c \times J_\ell^c \right] \\
&\quad + \frac{1}{(|I||J|)^2} \sum_{(i,j) \in I_k \times J_\ell} \mathbb{E}_{P_n} \left[\sum_{m=1}^{d_\theta} (\tilde{\psi}_{ij,m}^a)^2 \middle| I_k^c \times J_\ell^c \right] + 0 \\
&= \frac{1}{(|I||J|)^2} \sum_{(i,j) \in I_k \times J_\ell} \sum_{j' \in J_\ell, j' \neq j} \mathbb{E}_{P_n} [\langle \tilde{\psi}_{ij}^a, \tilde{\psi}_{ij'}^a \rangle | I_k^c \times J_\ell^c] \\
&\quad + \frac{1}{(|I||J|)^2} \sum_{(i,j) \in I_k \times J_\ell} \sum_{i' \in I_k, i' \neq i} \mathbb{E}_{P_n} [\langle \tilde{\psi}_{ij}^a, \tilde{\psi}_{i'j}^a \rangle | I_k^c \times J_\ell^c] \\
&\quad + \frac{1}{(|I||J|)^2} \sum_{(i,j) \in I_k \times J_\ell} \mathbb{E}_{P_n} [\|\tilde{\psi}_{ij}^a\|^2 | I_k^c \times J_\ell^c] \\
&\lesssim \frac{1}{|I| \wedge |J|} \mathbb{E}_{P_n} \left[\left\| \psi^a(W_{ij}; \hat{\theta}_{k\ell}) - \mathbb{E}_{P_n}[\psi^a(W_{ij}; \hat{\theta}_{k\ell}) | I_k^c \times J_\ell^c] \right\|^2 \middle| I_k^c \times J_\ell^c \right] \\
&\leq \frac{1}{|I| \wedge |J|} \mathbb{E}_{P_n} [\|\psi^a(W_{ij}; \hat{\theta}_{k\ell})\|^2 | I_k^c \times J_\ell^c] \\
&\leq c_1^2 / |I| \wedge |J|
\end{aligned}$$

under an application of Cauchy-Schwartz's inequality and Assumptions 1 and 3 (ii). Note that $\underline{C} \lesssim |I| \wedge |J| \lesssim \underline{C}$. Hence an application of Lemma 2 (i) implies $\mathcal{I}_{1,k\ell} = O_{P_n}(\underline{C}^{-1/2})$. This completes a proof of (A.1).

Step 3. It again suffices to show that for any $(k, \ell) \in [K]^2$, one has

$$\left\| \mathbb{E}_{n,k\ell}[\psi(W; \theta_0, \hat{\eta}_{k\ell})] - \frac{1}{|I||J|} \sum_{(i,j) \in I_k \times J_\ell} \psi(W_{ij}; \theta_0, \eta_0) \right\| = O_{P_n}(\underline{C}^{-1/2} r'_n + \lambda_n + \lambda'_n)$$

Denote

$$\mathbb{G}_{n,k\ell}[\phi(W)] = \frac{\sqrt{\underline{C}}}{|I||J|} \sum_{(i,j) \in I_k \times J_\ell} \left(\phi(W_{ij}) - \int \phi(w) dP_n \right),$$

where ϕ is P_n an integrable function on $\text{supp}(W)$. Then

$$\left\| \mathbb{E}_{n,k\ell}[\psi(W; \theta_0, \hat{\eta}_{k\ell})] - \frac{1}{|I||J|} \sum_{(i,j) \in I_k \times J_\ell} \psi(W_{ij}; \theta_0, \eta_0) \right\| \leq \frac{\mathcal{I}_{3,k\ell} + \mathcal{I}_{4,k\ell}}{\sqrt{\underline{C}}}$$

where

$$\begin{aligned} \mathcal{I}_{3,k\ell} &:= \left\| \mathbb{G}_{n,k\ell}[\psi(W; \theta_0, \hat{\eta}_{k,\ell})] - \mathbb{G}_{n,k\ell}[\psi(W; \theta_0, \eta_0)] \right\|, \\ \mathcal{I}_{4,k\ell} &:= \sqrt{\underline{C}} \left\| \mathbb{E}_{P_n}[\psi(W_{ij}; \theta_0, \hat{\eta}_{k,\ell}) | I_k \times J_\ell] - \mathbb{E}_{P_n}[\psi(W_{11}; \theta_0, \eta_0)] \right\|. \end{aligned}$$

Denote $\tilde{\psi}_{ij,m} := \psi_m(W_{ij}; \theta_0, \hat{\eta}_{k,\ell}) - \psi_m(W_{ij}; \theta_0, \eta_0)$ and $\tilde{\psi}_{ij} = (\tilde{\psi}_{ij,m})_{m \in [d_\theta]}$. To bound $\mathcal{I}_{3,k\ell}$, notice that using a similar argument as for the bound of $\mathcal{I}_{1,k\ell}$, one has

$$\begin{aligned} \mathbb{E}_{P_n}[\|\mathcal{I}_{3,k\ell}\|^2 | I_k^c \times J_\ell^c] &= \mathbb{E}_{P_n}[\|\mathbb{G}_{n,k\ell}[\psi(W_{ij}; \theta_0, \hat{\eta}_{k,\ell})] - \mathbb{G}_{n,k\ell}[\psi(W_{ij}; \theta_0, \eta_0)]\|^2 | I_k^c \times J_\ell^c] \\ &= \mathbb{E}_{P_n} \left[\frac{\underline{C}}{(|I||J|)^2} \sum_{m=1}^{d_\theta} \left\{ \sum_{(i,j) \in I_k \times J_\ell} (\tilde{\psi}_{ij,m} - \mathbb{E}_{P_n} \tilde{\psi}_{ij,m}) \right\}^2 \middle| I_k^c \times J_\ell^c \right] \\ &= \frac{\underline{C}}{(|I||J|)^2} \sum_{(i,j) \in I_k \times J_\ell} \sum_{j' \in J_\ell, j' \neq j} \mathbb{E}_{P_n} \left[\sum_{m=1}^{d_\theta} (\tilde{\psi}_{ij,m} - \mathbb{E}_{P_n} \tilde{\psi}_{ij,m}) (\tilde{\psi}_{ij',m} - \mathbb{E}_{P_n} \tilde{\psi}_{ij',m}) \middle| I_k^c \times J_\ell^c \right] \\ &\quad + \frac{\underline{C}}{(|I||J|)^2} \sum_{(i,j) \in I_k \times J_\ell} \sum_{i' \in I_k, i' \neq i} \mathbb{E}_{P_n} \left[\sum_{m=1}^{d_\theta} (\tilde{\psi}_{ij,m} - \mathbb{E}_{P_n} \tilde{\psi}_{ij,m}) (\tilde{\psi}_{i'j,m} - \mathbb{E}_{P_n} \tilde{\psi}_{i'j,m}) \middle| I_k^c \times J_\ell^c \right] \\ &\quad + \frac{\underline{C}}{(|I||J|)^2} \sum_{(i,j) \in I_k \times J_\ell} \mathbb{E}_{P_n} \left[\sum_{m=1}^{d_\theta} (\tilde{\psi}_{ij,m} - \mathbb{E}_{P_n} \tilde{\psi}_{ij,m})^2 \middle| I_k^c \times J_\ell^c \right] + 0 \\ &= \frac{\underline{C}}{(|I||J|)^2} \sum_{(i,j) \in I_k \times J_\ell} \sum_{j' \in J_\ell, j' \neq j} \mathbb{E}_{P_n} \left[\langle \tilde{\psi}_{ij} - \mathbb{E}_{P_n} \tilde{\psi}_{ij}, \tilde{\psi}_{ij'} - \mathbb{E}_{P_n} \tilde{\psi}_{ij'} \rangle \middle| I_k^c \times J_\ell^c \right] \\ &\quad + \frac{\underline{C}}{(|I||J|)^2} \sum_{(i,j) \in I_k \times J_\ell} \sum_{i' \in I_k, i' \neq i} \mathbb{E}_{P_n} \left[\langle \tilde{\psi}_{ij} - \mathbb{E}_{P_n} \tilde{\psi}_{ij}, \tilde{\psi}_{i'j} - \mathbb{E}_{P_n} \tilde{\psi}_{i'j} \rangle \middle| I_k^c \times J_\ell^c \right] \\ &\quad + \frac{\underline{C}}{(|I||J|)^2} \sum_{(i,j) \in I_k \times J_\ell} \mathbb{E}_{P_n} \left[\|\tilde{\psi}_{ij} - \mathbb{E}_{P_n} \tilde{\psi}_{ij}\|^2 \middle| I_k^c \times J_\ell^c \right] \\ &\lesssim \mathbb{E}_{P_n} \left[\left\| \psi(W_{ij}; \theta_0, \hat{\eta}) - \psi(W_{ij}; \theta_0, \eta_0) - \mathbb{E}_{P_n}[\psi(W_{ij}; \theta_0, \hat{\eta}) - \psi(W_{ij}; \theta_0, \eta_0)] \right\|^2 \middle| I_k^c \times J_\ell^c \right] \\ &\leq \mathbb{E}_{P_n}[\|\psi(W_{ij}; \theta_0, \hat{\eta}) - \psi(W_{ij}; \theta_0, \eta_0)\|^2 | I_k^c \times J_\ell^c] \\ &\leq \sup_{\eta \in \mathcal{T}_n} \mathbb{E}_{P_n}[\|\psi(W_{00}; \theta_0, \eta) - \psi(W_{00}; \theta_0, \eta_0)\|^2 | I_k^c \times J_\ell^c] \\ &= \sup_{\eta \in \mathcal{T}_n} \mathbb{E}_{P_n}[\|\psi(W_{00}; \theta_0, \eta) - \psi(W_{00}; \theta_0, \eta_0)\|^2] = (r'_n)^2, \end{aligned}$$

where the first inequality follows from Cauchy-Schwartz's inequality, the second-to-last equality is due to Assumption 1, and the last equality is due to Assumption 3 (iii).

Hence, $\mathcal{I}_{3,k\ell} = O_{P_n}(r'_n)$. To bound $\mathcal{I}_{4,k\ell}$, let

$$f_{k\ell}(r) := \mathbb{E}_{P_n}[\psi(W_{ij}; \theta_0, \eta_0 + r(\widehat{\eta}_{k\ell} - \eta_0)) | I_k^c \times J_\ell^c] - \mathbb{E}_{P_n}[\psi(W_{11}; \theta_0, \eta_0)], \quad r \in [0, 1].$$

An application of the mean value expansion coordinate-wise gives

$$f_{k\ell}(1) = f_{k\ell}(0) + f'_{k\ell}(0) + f''_{k\ell}(\tilde{r})/2,$$

where $\tilde{r} \in (0, 1)$. Note that $f_{k\ell}(0) = 0$ under Assumption 2 (i), and

$$\|f'_{k\ell}(0)\| = \left\| \partial_\eta \mathbb{E}_{P_n} \psi(W; \theta_0, \eta_0) [\widehat{\eta}_{k\ell} - \eta_0] \right\| \leq \lambda_n$$

under Assumption 2 (iv). Moreover, under Assumption 3 (iii), on the event \mathcal{E}_n , we have

$$\|f''_{k\ell}(\tilde{r})\| \leq \sup_{r \in (0,1)} \|f''_{k\ell}(r)\| \leq \lambda'_n.$$

This completes a proof of (A.2).

Step 4. Note that

$$\begin{aligned} \mathbb{E}_{P_n} \left[\left\| \frac{\sqrt{C}}{NM} \sum_{i=1}^N \sum_{j=1}^M \psi(W_{ij}; \theta_0, \eta_0) \right\|^2 \right] &= \frac{C}{(NM)^2} \mathbb{E}_{P_n} \left[\sum_{m=1}^{d_\theta} \left(\sum_{i=1}^N \sum_{j=1}^M \psi_m(W_{ij}; \theta_0, \eta_0) \right)^2 \right] \\ &= \frac{C}{(NM)^2} \sum_{i=1}^N \sum_{1 \leq j, j' \leq M} \mathbb{E}_{P_n} \left[\sum_{m=1}^{d_\theta} \psi_m(W_{ij}; \theta_0, \eta_0) \psi_m(W_{ij'}; \theta_0, \eta_0) \right] \\ &\quad + \frac{C}{(NM)^2} \sum_{1 \leq i, i' \leq N} \sum_{j=1}^M \mathbb{E}_{P_n} \left[\sum_{m=1}^{d_\theta} \psi_m(W_{ij}; \theta_0, \eta_0) \psi_m(W_{i'j}; \theta_0, \eta_0) \right] \\ &\quad + \frac{C}{(NM)^2} \sum_{i=1}^N \sum_{j=1}^M \mathbb{E}_{P_n} \left[\sum_{m=1}^{d_\theta} \psi_m^2(W_{ij}; \theta_0, \eta_0) \right] + 0 \\ &\lesssim \mathbb{E}_{P_n} [\|\psi(W_{ij}; \theta_0, \eta_0)\|^2] \leq c_1^2 \end{aligned}$$

under Assumptions 1 and 3 (ii). Therefore, an application of Markov's inequality implies

$$\left\| \frac{\sqrt{C}}{NM} \sum_{i=1}^N \sum_{j=1}^M \psi(W_{ij}; \theta_0, \eta_0) \right\| = O_{P_n}(1).$$

This completes a proof of (A.3).

Step 5. Note that all singular values of J_0 are bounded from above by c_1 under Assumption 2 (v) and all eigenvalues of Γ are bounded from below by c_0 under Assumption 3 (iv). Therefore, we have $\|\sigma^{-1}\| \leq c_1/\sqrt{c_0}$ and thus $\|\sigma^{-1}\| = O_{P_n}(1)$. This completes a proof of (A.4). \square

A.2 Proof of Theorem 2

Proof. Step 2 of the proof of Theorem 1 proves $\|\hat{J} - J_0\| = O_p(\underline{C}^{-1/2} + r_n)$ and Assumption 2 (v) implies $\|J_0^{-1}\| \leq c_0^{-1}$. Therefore, to prove the claim of the theorem, it suffices to show

$$\begin{aligned} & \left\| \frac{1}{K^2} \sum_{(k,\ell) \in [K]^2} \left\{ \frac{|I| \wedge |J|}{(|I||J|)^2} \sum_{i \in I_k} \sum_{j, j' \in J_\ell} \psi(W_{ij}; \tilde{\theta}, \hat{\eta}_{k\ell}) \psi(W_{ij'}; \tilde{\theta}, \hat{\eta}_{k\ell})' \right. \right. \\ & \quad \left. \left. + \frac{|I| \wedge |J|}{(|I||J|)^2} \sum_{i, i' \in I_k} \sum_{j \in J_\ell} \psi(W_{ij}; \tilde{\theta}, \hat{\eta}_{k\ell}) \psi(W_{i'j}; \tilde{\theta}, \hat{\eta}_{k\ell})' \right\} \right. \\ & \quad \left. - \bar{\mu}_N \mathbb{E}_P[\psi(W_{11}; \theta_0, \eta_0) \psi(W_{12}; \theta_0, \eta_0)'] - \bar{\mu}_M \mathbb{E}_P[\psi(W_{11}; \theta_0, \eta_0) \psi(W_{21}; \theta_0, \eta_0)'] \right\| = O_P(\rho_n). \end{aligned}$$

Moreover, since K and d_θ are constants and $\mu_N \rightarrow \bar{\mu}_N \leq 1$ and $\mu_M \rightarrow \bar{\mu}_M \leq 1$, it suffices to show that for each $(k, \ell) \in [K]^2$ and $l, m \in [d_\theta]$, it holds that

$$\left| \frac{|I| \wedge |J|}{(|I||J|)^2} \sum_{i \in I_k} \sum_{j, j' \in J_\ell} \psi_l(W_{ij}; \tilde{\theta}, \hat{\eta}_{k\ell}) \psi_m(W_{ij'}; \tilde{\theta}, \hat{\eta}_{k\ell}) - \mu_N \mathbb{E}_P[\psi_l(W_{11}; \theta_0, \eta_0) \psi_m(W_{12}; \theta_0, \eta_0)] \right| = O_P(\rho_n)$$

and

$$\left| \frac{|I| \wedge |J|}{(|I||J|)^2} \sum_{i, i' \in I_k} \sum_{j \in J_\ell} \psi_l(W_{ij}; \tilde{\theta}, \hat{\eta}_{k\ell}) \psi_m(W_{i'j}; \tilde{\theta}, \hat{\eta}_{k\ell}) - \mu_M \mathbb{E}_P[\psi_l(W_{11}; \theta_0, \eta_0) \psi_m(W_{21}; \theta_0, \eta_0)] \right| = O_P(\rho_n).$$

We will show the second statement since the first one follows analogously. Denote the left-hand side of the equation as $\mathcal{I}_{k\ell, lm}$. First, note that $(|I| \wedge |J|)/|J| = \mu_M$, and apply the triangle inequality to get

$$\mathcal{I}_{k\ell, lm} \leq \mathcal{I}_{k\ell, lm, 1} + \mathcal{I}_{k\ell, lm, 2},$$

where

$$\begin{aligned} \mathcal{I}_{k\ell, lm, 1} &:= \left| \frac{1}{|I|^2 |J|} \sum_{i, i' \in I_k} \sum_{j \in J_\ell} \left\{ \psi_l(W_{ij}; \tilde{\theta}, \hat{\eta}_{k\ell}) \psi_m(W_{i'j}; \tilde{\theta}, \hat{\eta}_{k\ell}) - \psi_l(W_{ij}; \theta_0, \eta_0) \psi_m(W_{i'j}; \theta_0, \eta_0) \right\} \right| \\ \mathcal{I}_{k\ell, lm, 2} &:= \left| \frac{1}{|I|^2 |J|} \sum_{i, i' \in I_k} \sum_{j \in J_\ell} \psi_l(W_{ij}; \theta_0, \eta_0) \psi_m(W_{i'j}; \theta_0, \eta_0) - \mathbb{E}_P[\psi_l(W_{11}; \theta_0, \eta_0) \psi_m(W_{21}; \theta_0, \eta_0)] \right|. \end{aligned}$$

We first find a bound for $\mathcal{I}_{k\ell,lm,2}$. Since $q > 4$, it holds that

$$\begin{aligned}
\mathbb{E}_P[\mathcal{I}_{k\ell,lm,2}^2] &= \frac{1}{|I|^4|J|^2} \mathbb{E}_P \left[\left| \sum_{i,i' \in I_k} \sum_{j \in J_\ell} \psi_l(W_{ij}; \theta_0, \eta_0) \psi_m(W_{i'j}; \theta_0, \eta_0) - \mathbb{E}_P[\psi_l(W_{11}; \theta_0, \eta_0) \psi_m(W_{21}; \theta_0, \eta_0)] \right|^2 \right] \\
&\leq \frac{1}{|I|^4|J|^2} \mathbb{E}_P \left[\sum_{i,i',i'' \in I_k} \sum_{j,j' \in J_\ell} \psi_l(W_{ij}; \theta_0, \eta_0) \psi_m(W_{i'j}; \theta_0, \eta_0) \psi_l(W_{ij'}; \theta_0, \eta_0) \psi_m(W_{i''j'}; \theta_0, \eta_0) \right] \\
&\quad + \frac{1}{|I|^4|J|^2} \mathbb{E}_P \left[\sum_{i,i',i'',i''' \in I_k} \sum_{j \in J_\ell} \psi_l(W_{ij}; \theta_0, \eta_0) \psi_m(W_{i'j}; \theta_0, \eta_0) \psi_l(W_{i''j}; \theta_0, \eta_0) \psi_m(W_{i'''j}; \theta_0, \eta_0) \right] \\
&\quad + o((|I| \wedge |J|)^{-1}) + 0 \\
&\lesssim \frac{1}{|I| \wedge |J|} \mathbb{E}_P[\|\psi(W; \theta_0, \eta_0)\|^4] \lesssim c_1^4/\underline{C} = O(\underline{C}^{-1/2}).
\end{aligned}$$

Now, to bound $\mathcal{I}_{k\ell,lm,1}$, we make use of the following identity coming from the proof of Theorem 3.2 in CCDDHNR (2018): for any numbers $a, b, \delta a, \delta b$ such that $|a| \vee |b| \leq c$ and $|\delta a| \vee |\delta b| \leq r$, it holds that $|(a + \delta a)(b + \delta b) - ab| \leq 2r(c + r)$. Denote $\psi_{ij,h} := \psi_l(W_{ij}; \theta_0, \eta_0)$ and $\widehat{\psi}_{ij,h} := \psi_l(W_{ij}; \widehat{\theta}, \widehat{\eta}_{k\ell})$ for $h \in \{l, m\}$ and apply the above identity with $a = \psi_{ij,l}$, $b = \psi_{i'j,m}$, $a + \delta a = \widehat{\psi}_{ij,l}$, $b + \delta b = \widehat{\psi}_{i'j,m}$, $r = |\widehat{\psi}_{ij,l} - \psi_{ij,l}| \vee |\widehat{\psi}_{i'j,m} - \psi_{i'j,m}|$ and $c = |\psi_{ij,l}| \vee |\psi_{i'j,m}|$. Then

$$\begin{aligned}
\mathcal{I}_{k\ell,lm,1} &= \left| \frac{1}{|I|^2|J|} \sum_{i,i' \in I_k} \sum_{j \in J_\ell} \left\{ \widehat{\psi}_{ij,l} \widehat{\psi}_{i'j,m} - \psi_{ij,l} \psi_{i'j,m} \right\} \right| \\
&\leq \frac{1}{|I|^2|J|} \sum_{i,i' \in I_k} \sum_{j \in J_\ell} |\widehat{\psi}_{ij,l} \widehat{\psi}_{i'j,m} - \psi_{ij,l} \psi_{i'j,m}| \\
&\leq \frac{2}{|I|^2|J|} \sum_{i,i' \in I_k} \sum_{j \in J_\ell} (|\widehat{\psi}_{ij,l} - \psi_{ij,l}| \vee |\widehat{\psi}_{i'j,m} - \psi_{i'j,m}|) \\
&\quad \times (|\psi_{ij,l}| \vee |\psi_{i'j,m}| + |\widehat{\psi}_{ij,l} - \psi_{ij,l}| \vee |\widehat{\psi}_{i'j,m} - \psi_{i'j,m}|) \\
&\leq \left(\frac{2}{|I|^2|J|} \sum_{i,i' \in I_k} \sum_{j \in J_\ell} |\widehat{\psi}_{ij,l} - \psi_{ij,l}|^2 \vee |\widehat{\psi}_{i'j,m} - \psi_{i'j,m}|^2 \right)^{1/2} \\
&\quad \times \left(\frac{2}{|I|^2|J|} \sum_{i,i' \in I_k} \sum_{j \in J_\ell} \left\{ |\psi_{ij,l}| \vee |\psi_{i'j,m}| + |\widehat{\psi}_{ij,l} - \psi_{ij,l}| \vee |\widehat{\psi}_{i'j,m} - \psi_{i'j,m}| \right\}^2 \right)^{1/2} \\
&\leq \left(\frac{2}{|I|^2|J|} \sum_{i,i' \in I_k} \sum_{j \in J_\ell} |\widehat{\psi}_{ij,l} - \psi_{ij,l}|^2 \vee |\widehat{\psi}_{i'j,m} - \psi_{i'j,m}|^2 \right)^{1/2} \\
&\quad \times \left\{ \left(\frac{2}{|I|^2|J|} \sum_{i,i' \in I_k} \sum_{j \in J_\ell} |\psi_{ij,l}|^2 \vee |\psi_{i'j,m}|^2 \right)^{1/2} \right. \\
&\quad \left. + \left(\frac{2}{|I|^2|J|} \sum_{i,i' \in I_k} \sum_{j \in J_\ell} |\widehat{\psi}_{ij,l} - \psi_{ij,l}|^2 \vee |\widehat{\psi}_{i'j,m} - \psi_{i'j,m}|^2 \right)^{1/2} \right\}
\end{aligned}$$

where the second to the last inequality follows the Cauchy-Schwartz's inequality and Minkowski's inequality. Notice that

$$\begin{aligned} \sum_{i,i' \in I_k} \sum_{j \in J_\ell} |\psi_{ij,l}|^2 \vee |\psi_{i'j,m}|^2 &\leq |I| \sum_{i=1}^N \sum_{j=1}^M \|\psi(W_{ij}; \theta_0, \eta_0)\|^2, \\ \sum_{i,i' \in I_k} \sum_{j \in J_\ell} |\widehat{\psi}_{ij,l} - \psi_{ij,l}|^2 \vee |\widehat{\psi}_{i'j,m} - \psi_{i'j,m}|^2 &\leq |I| \sum_{i=1}^N \sum_{j=1}^M \|\psi(W_{ij}; \widetilde{\theta}, \widehat{\eta}_{k\ell}) - \psi(W_{ij}; \theta_0, \eta_0)\|^2. \end{aligned}$$

Thus, the above bound for $\mathcal{I}_{k\ell,lm,1}$ implies that

$$\mathcal{I}_{k\ell,lm,1}^2 \lesssim R_n \times \left(\frac{1}{|I||J|} \sum_{(i,j) \in I_k \times J_\ell} \|\psi(W_{ij}; \theta_0, \eta_0)\|^2 + R_n \right),$$

where

$$R_n := \frac{1}{|I||J|} \sum_{(i,j) \in I_k \times J_\ell} \|\psi(W_{ij}; \widetilde{\theta}, \widehat{\eta}_{k\ell}) - \psi(W_{ij}; \theta_0, \eta_0)\|^2.$$

Notice that

$$\frac{1}{|I||J|} \sum_{(i,j) \in I_k \times J_\ell} \|\psi(W_{ij}; \theta_0, \eta_0)\|^2 = O_P(1),$$

which is implied by Markov's inequality and the calculations

$$\mathbb{E}_P \left[\frac{1}{|I||J|} \sum_{(i,j) \in I_k \times J_\ell} \|\psi(W_{ij}; \theta_0, \eta_0)\|^2 \right] = \mathbb{E}_P [\|\psi(W_{11}; \theta_0, \eta_0)\|^2] \leq c_1^2$$

under Assumptions 1 and 3 (ii). Finally, to bound R_n , using Assumption 2 (ii),

$$R_n \lesssim \frac{1}{|I||J|} \sum_{(i,j) \in I_k \times J_\ell} \|\psi^a(W_{ij}; \widehat{\eta}_{k\ell})(\widetilde{\theta} - \theta_0)\|^2 + \frac{1}{|I||J|} \sum_{(i,j) \in I_k \times J_\ell} \|\psi(W_{ij}; \theta_0, \widehat{\eta}_{k\ell}) - \psi(W_{ij}; \theta_0, \eta_0)\|^2.$$

The first term on RHS is bounded by

$$\left(\frac{1}{|I||J|} \sum_{(i,j) \in I_k \times J_\ell} \|\psi^a(W_{ij}; \widehat{\eta}_{k\ell})\|^2 \right) \times \|\widetilde{\theta} - \theta_0\|^2 = O_P(1) \times O_P(\underline{C}^{-1}) = O_P(\underline{C}^{-1})$$

due to Assumption 3 (ii), Markov's inequality, and Theorem 1. Furthermore, given that $(W_{ij})_{(i,j) \in I_k^c \times J_\ell^c}$

satisfies $\widehat{\eta}_{k\ell} \in \mathcal{T}_n$,

$$\mathbb{E}_P \left[\|\psi(W_{ij}; \theta_0, \widehat{\eta}_{k\ell}) - \psi(W_{ij}; \theta_0, \eta_0)\|^2 \Big| I_k^c \times J_\ell^c \right] \leq \sup_{\eta \in \mathcal{T}_n} \mathbb{E}_P \left[\|\psi(W_{ij}; \theta_0, \eta) - \psi(W_{ij}; \theta_0, \eta_0)\|^2 \Big| I_k^c \times J_\ell^c \right] \leq (r'_n)^2$$

due to Assumptions 1 and 3 (iii). Also, the event $\hat{\eta}_{kl} \in \mathcal{T}_n$ happens with probability $1 - o(1)$, we have $R_n = O_P(\underline{C}^{-1} + (r'_n)^2)$. Thus we conclude that

$$\mathcal{I}_{kl,lm,1} = O_P(\underline{C}^{-1/2} + r'_n).$$

This completes the proof. \square

B Useful Lemmas

We collect some of the useful auxiliary results in this section.

First, for any $f : \text{supp}(\mathbf{W}) \rightarrow \mathbb{R}^d$ for a fixed $d \in \mathbb{N}$, we use

$$\mathbb{G}_{\underline{C}}f := \sqrt{\underline{C}} \left\{ \frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M f(W_{ij}) - \mathbb{E}_P[f(W_{11})] \right\}$$

to denote its multiway empirical process. The following is a multivariate version of Chiang and Sasaki (2019), Lemma 1; see also Lemma D.2 in Davezies et al. (2018).

Lemma 1 (Independentization via Hájek Projections). *If Assumption 1 holds and $f : \text{supp}(\mathbf{W}) \rightarrow \mathbb{R}^d$ for some fixed $d \in \mathbb{N}$ and suppose $\mathbb{E}_P\|f(W_{11})\|^2 < K$ for a finite constant K that is independent of n , then there exist i.i.d. uniform random variables U_{i0} and U_{0j} such that the Hájek projection $H_n f$ of $\mathbb{G}_{\underline{C}}f$ on*

$$\mathcal{G}_n = \left\{ \sum_{i=1}^N g_{i0}(U_{i0}) + \sum_{j=1}^M g_{0j}(U_{0j}) : g_{i0}, g_{0j} \in L^2(P_n) \right\}$$

is equal to

$$H_n f = \frac{\sqrt{\underline{C}}}{N} \sum_{i=1}^N \mathbb{E}_P \left[f(W_{i1}) - \mathbb{E}_P f(W_{11}) \middle| U_{i0} \right] + \frac{\sqrt{\underline{C}}}{M} \sum_{j=1}^M \mathbb{E}_P \left[f(W_{1j}) - \mathbb{E}_P f(W_{11}) \middle| U_{0j} \right]$$

for each n . Furthermore,

$$V(\mathbb{G}_{\underline{C}}f) = V(H_n f) + O(\underline{C}^{-1}) = \bar{\mu}_N \text{Cov}(f(W_{11}), f(W_{12})) + \bar{\mu}_M \text{Cov}(f(W_{11}), f(W_{21})) + O(\underline{C}^{-1})$$

holds a.s.

Proof. The proof is essentially the same as the proof for Lemma 1 of Chiang and Sasaki (2019) and is therefore omitted. \square

The following re-states Lemma 6.1. of CCDDHNR (2018):

Lemma 2 (Conditional Convergence Implies Unconditional). *Let (X_n) and (Y_n) be sequences of random vectors.*

- (i) *If for $\epsilon_n \rightarrow 0$, $P(\|X_n\| > \epsilon_n | Y_n) = o_P(1)$ in probability, then $P(\|X_n\| > \epsilon_n) = o(1)$. In particular, this occurs if $E_P[\|X_n\|^q / \epsilon_n^q | Y_n] = o_P(1)$ for some $q \geq 1$.*
- (ii) *Let (A_n) be a sequence of positive constants. If $\|X_n\| = O_P(A_n)$ conditional on Y_n , then $\|X_n\| = O_P(A_n)$ unconditional, namely, for any $l_n \rightarrow \infty$, $P(\|X_n\| > l_n A_n) = o(1)$.*

References

- ATHEY, S. AND G. W. IMBENS (2019): “Machine Learning Methods That Economists Should Know About,” *Annual Review of Economics*, 11.
- ATHEY, S. AND S. WAGER (2019): “Estimating Treatment Effects with Causal Forests: An Application,” *arXiv preprint arXiv:1902.07409*.
- 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, 2369–2429.
- BELLONI, A., V. CHERNOZHUKOV, D. CHETVERIKOV, AND Y. WEI (2018): “Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework,” *The Annals of Statistics*, 46, 3643–3675.
- BELLONI, A., V. CHERNOZHUKOV, I. FERNÁNDEZ-VAL, AND C. HANSEN (2017): “Program evaluation and causal inference with high-dimensional data,” *Econometrica*, 85, 233–298.
- BELLONI, A., V. CHERNOZHUKOV, AND C. HANSEN (2014a): “High-dimensional methods and inference on structural and treatment effects,” *Journal of Economic Perspectives*, 28, 29–50.
- (2014b): “Inference on treatment effects after selection among high-dimensional controls,” *The Review of Economic Studies*, 81, 608–650.

- BELLONI, A., V. CHERNOZHUKOV, C. HANSEN, AND D. KOZBUR (2016): “Inference in high-dimensional panel models with an application to gun control,” *Journal of Business & Economic Statistics*, 34, 590–605.
- BELLONI, A., V. CHERNOZHUKOV, AND K. KATO (2015): “Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems,” *Biometrika*, 102, 77–94.
- BERRY, S. T. (1994): “Estimating discrete-choice models of product differentiation,” *The RAND Journal of Economics*, 242–262.
- CAMERON, A. C. AND D. L. MILLER (2015): “A practitioners guide to cluster-robust inference,” *Journal of Human Resources*, 50, 317–372.
- CAMERON, C. A., J. B. GELBACH, AND D. L. MILLER (2011): “Robust Inference With Multiway Clustering,” *Journal of Business and Economic Statistics*, 29, 238 – 249.
- 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, C1 – C68.
- CHIANG, H. AND Y. SASAKI (2019): “Lasso under Multi-way Clustering: Estimation and Post-selection Inference,” ArXiv:1905.02107.
- DAVEZIES, L., X. D’HAULTFOEUILLE, AND Y. GUYONVARCH (2018): “Asymptotic Results under Multiway Clustering,” ArXiv:1807.07925.
- (2019): “Empirical Process Results for Exchangeable Arrays,” *arXiv preprint arXiv:1906.11293*.
- HANSEN, C. AND Y. LIAO (2019): “The factor-lasso and k-step bootstrap approach for inference in high-dimensional economic applications,” *Econometric Theory*, 35, 465–509.
- KOCK, A. B. (2016): “Oracle Inequalities, Variable Selection and Uniform Inference in High-Dimensional Correlated Random Effects Panel Data Models,” *Journal of Econometrics*, 195, 71 – 85.

- KOCK, A. B. AND H. TANG (2019): “Uniform Inference in High-Dimensional Dynamic Panel Data Models with Approximately Sparse Fixed Effects,” *Econometric Theory*, 35, 295–359.
- LEE, S. AND S. NG (2019): “An Econometric View of Algorithmic Subsampling,” *arXiv preprint arXiv:1907.01954*.
- LU, Z., X. SHI, AND J. TAO (2019): “Semi-Nonparametric Estimation of Random Coefficient Logit Model for Aggregate Demand,” Working Paper.
- MACKINNON, J. G. (2019): “How cluster-robust inference is changing applied econometrics,” *Canadian Journal of Economics/Revue canadienne d’économique*.
- MACKINNON, J. G., M. O. NIELSEN, AND M. D. WEBB (2019): “Wild Bootstrap and Asymptotic Inference with Multiway Clustering,” Queen’s Economics Department Working Paper, No. 1415.
- MENZEL, K. (2017): “Bootstrap with Clustering in Two or More Dimensions,” ArXiv:1703.03043.
- MULLAINATHAN, S. AND J. SPIESS (2017): “Machine learning: an applied econometric approach,” *Journal of Economic Perspectives*, 31, 87–106.
- SEMENOVA, V., M. GOLDMAN, V. CHERNOZHUKOV, AND M. TADDY (2018): “Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels,” *arXiv preprint arXiv:1712.09988*.
- VAN DER VAART, A. (1998): *Asymptotic Statistics*, Cambridge University Press.

N	M	\underline{C}	$\dim(X)$	K (K^2)	Machine Learning	Bias	SD	RMSE	Cover
25	25	25	100	2 (4)	Ridge	0.069	0.074	0.102	0.835
					Elastic Net	0.010	0.079	0.080	0.963
					Lasso	0.005	0.080	0.080	0.965
50	50	50	100	2 (4)	Ridge	0.014	0.047	0.049	0.940
					Elastic Net	-0.002	0.048	0.048	0.956
					Lasso	-0.001	0.049	0.049	0.955
25	25	25	200	2 (4)	Ridge	0.190	0.053	0.197	0.118
					Elastic Net	0.016	0.077	0.079	0.969
					Lasso	0.006	0.080	0.080	0.968
50	50	50	200	2 (4)	Ridge	0.037	0.046	0.058	0.876
					Elastic Net	-0.000	0.048	0.048	0.960
					Lasso	-0.002	0.048	0.048	0.962
25	25	25	100	3 (9)	Ridge	0.042	0.074	0.085	0.962
					Elastic Net	0.004	0.074	0.074	0.993
					Lasso	0.002	0.075	0.075	0.992
50	50	50	100	3 (9)	Ridge	0.007	0.048	0.049	0.962
					Elastic Net	-0.001	0.047	0.047	0.972
					Lasso	-0.001	0.048	0.048	0.963
25	25	25	200	3 (9)	Ridge	0.081	0.067	0.105	0.896
					Elastic Net	0.005	0.073	0.073	0.994
					Lasso	0.003	0.076	0.077	0.992
50	50	50	200	3 (9)	Ridge	0.018	0.047	0.050	0.944
					Elastic Net	-0.002	0.048	0.048	0.968
					Lasso	-0.003	0.049	0.049	0.968

Table 1: Simulation results based on 5,000 Monte Carlo iterations. Results are displayed for each of the three machine learning methods, including the ridge, elastic net, and lasso. Reported statistics are the bias (Bias), standard deviation (SD), root mean square error (RMSE), and coverage frequency for the nominal probability of 95% (Cover).