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. We also develop a multiway cluster robust standard error formula. Simulations indicate that the proposed procedure has favorable finite sample performance. Applying the proposed method to market share data for demand analysis, we obtain larger two-way cluster robust standard errors for the price coefficient than non-robust ones in the demand model.

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

JEL Codes: C10, C13, C14

*First arXiv date: September 8, 2019
†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
∥We benefited from useful comments by seminar participants at Southern Methodist University, Stony Brook University, University of Bristol, and University of Colorado - Boulder, and participants at CeMMAP UCL/Vanderbilt Joint Conference on Advances in Econometrics and CeMMAP Workshop on Causal Learning with Interactions. All remaining errors are ours.
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 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, Newey, and Robins (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 microeconometric 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. 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 repre-

1As of December 31, 2019, Cameron et al. (2011) has received over 2,500 citations. The majority of such citations came from applied economic papers.
sentation 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. Using 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 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 Kozbur (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 multi-way 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 \( \{ W_{ij} \mid i \in \{1, \ldots, N\}, j \in \{1, \ldots, M\} \} \) of double-indexed observations of size \( NM \). Let \( P \) denote the probability law of \( \{ W_{ij} \}_{ij} \), and let \( E_P \) denote the expectation with respect to \( P \). Let \( C = N \wedge M \) 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 at the expense of involved notations – see Appendix C for a general case.

\(^2\)See also Davezies, D’Haultfoeuille, and Guyonvarch (2019) for further generalization of the empirical process theory for dyadic data under joint exchangeability assumption.
The structural model is assumed to entail the moment restriction

\[ \mathbb{E}_P[\psi(W_{11}; \theta_0, \eta_0)] = 0 \]  

(2.1)

for some score \( \psi \) 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 subset \( T \) of a normed linear space. The nuisance parameter \( \eta \) may be finite-, high-, or infinite-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 \( D_r : \tilde{T} \rightarrow \mathbb{R}^{d_\eta} \) by

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

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 \((\theta_0, \eta_0)\) with respect to a nuisance realization set \( T_n \subset T \) if the score \( \psi \) satisfies (2.1), the pathwise derivative \( D_r[\eta - \eta_0] \) exists for all \( r \in [0, 1) \) and \( \eta \in T_n \), and the orthogonality equation

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

(2.2)

holds for all \( \eta \in T_n \). Furthermore, we also say that the \( \lambda_n \) Neyman near-orthogonality condition holds at \((\theta_0, \eta_0)\) with respect to a nuisance realization set \( T_n \subset T \) if the score \( \psi \) satisfies (2.1), the pathwise derivative \( D_r[\eta - \eta_0] \) exists for all \( r \in [0, 1) \) and \( \eta \in T_n \), and the orthogonality equation

\[ \sup_{\eta \in T_n} \left\| \partial_\eta \mathbb{E}_P \psi(W; \theta_0, \eta_0)[\eta - \eta_0] \right\| \leq \lambda_n \]  

(2.3)

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

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). Consider linear Neyman orthogonal scores \( \psi \) of the form

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

(2.4)

A generalization to nonlinear score follows from linearization with Gateaux differentiability as in Section 3.3 of CCDDHNR (2018). We focus on linear scores as they cover a wide range of applications.
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 \( \theta_0 \). For any \( r \in \mathbb{N} \), we use the notation \( [r] = \{1, ..., r\} \). With a fixed positive integer \( K \), randomly partition \( [N] \) into \( K \) parts \( \{I_1, ..., I_K\} \) and \( [M] \) into \( K \) parts \( \{J_1, ..., J_K\} \). For each \( (k, \ell) \in [K]^2 \), obtain an estimate

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

of the nuisance parameter \( \eta \) 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 \( \theta_0 \), as the solution to

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

where \( E_{n,k\ell}[f(W)] = \frac{1}{|I_k||J_\ell|} \sum_{(i,j)\in I_k \times J_\ell} f(W_{ij}) \) 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 \( E_{n,k\ell}[\psi(W; \cdot, \hat{\eta}_{k\ell})] \) 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 \( \alpha \)-way clustering for an integer \( \alpha > 2 \), the above algorithm can be easily generalized into a \( K^\alpha \)-fold multiway DML estimator. See Appendix [C] for a generalization.
We propose to estimate the asymptotic variance of $\sqrt{C}(\tilde{\theta} - \theta_0)$ by

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

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

$$\tilde{\Gamma} = \frac{1}{K^2} \sum_{(k,\ell) \in [K]^2} \left\{ \frac{|I| \wedge |J|}{(|I| |J|)^2} \sum_{i \in I_k} \sum_{j, j' \in J_\ell} \psi(W_{ij} ; \tilde{\theta}, \tilde{\eta}_{k\ell}) \psi(W_{ij'} ; \tilde{\theta}, \tilde{\eta}_{k\ell}) \right\} \quad \text{and}$$

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

accounting for multiway cluster dependence. For a $d_\theta$-dimensional vector $r$, the $(1 - a)$ confidence interval for the linear functional $r' \theta_0$ can be constructed by

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

### 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. Okui, Small, Tan and Robins, 2012; 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 \( \theta_0 \) is an object of interest.

A Neyman orthogonal score \( \psi \) for such model is given by

\[
\psi(w; \theta, \eta) = (y - g_1(x) - \theta(d - g_2(x)))(z - m(x))
\]

as in Okui et al. (2012) and CCDDHN (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 \( m_0(X) = E_P[Z|X] \).

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 \( \{I_1, \ldots, I_K\} \) and \([M] \) into \( K \) parts \( \{J_1, \ldots, J_K\} \).

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 \( \theta \) to obtain the multiway DML estimate \( \tilde{\theta} \).

4. Let \( \tilde{\epsilon}_{ij} = Y_{ij} - X_{ij}'\hat{\beta}_{k\ell} - \tilde{\theta}(D_{ij} - X_{ij}'\hat{\gamma}_{k\ell}), \) \( \tilde{u}_{ij} = D_{ij} - X_{ij}'\hat{\gamma}_{k\ell}, \) and \( \tilde{v}_{ij} = Z_{ij} - X_{ij}'\hat{\xi}_{k\ell} \) 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
\[
\hat{\sigma}^2 = \hat{J}^{-1} \frac{1}{K^2} \sum_{k=1}^{K} \sum_{\ell=1}^{K} \left\{ \frac{|I| \wedge |J|}{(|I| \wedge |J|)^2} \sum_{i \in I_k, j \in J_\ell} \hat{\epsilon}_{ij} \hat{v}_{ij} \hat{v}_{ij}' \hat{\epsilon}_{ij}' + \frac{|I| \wedge |J|}{(|I| \wedge |J|)^2} \sum_{i,i' \in I_k, j \in J_\ell} \hat{\epsilon}_{ij} \hat{v}_{ij} \hat{v}_{ij}' \hat{\epsilon}_{ij}' \right\} (\hat{J}^{-1})',
\]
where
\[
\hat{J} = - \frac{1}{K^2} \sum_{k=1}^{K} \sum_{\ell=1}^{K} E_{n,kl}[\hat{u}_{ij} \hat{v}_{ij}].
\]

5. Report the estimate \( \tilde{\theta} \), its standard error \( \sqrt{\hat{\sigma}^2 / C} \), and/or the \((1-a)\) confidence interval
\[
CL_a := \left[ \tilde{\theta} \pm \Phi^{-1}(1-a/2) \sqrt{\hat{\sigma}^2 / C} \right].
\]

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 \( \varepsilon_{cij} \) independently follows the Type I Extreme Value distribution, \( \alpha_c \) is a random coefficient, and the mean utility \( \delta_{ij} \) 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 \( \theta_0 \), 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 \( E_P[\epsilon_{ij}|X_{ij},Z_{ij}] = 0 \). This yields the reduced-form equation (2.7), together with the innocuous nonparametric projection equation (2.8). Since the random vector \( W_{ij} = (Y_{ij},D_{ij},X_{ij},Z_{ij}) \) is double-indexed by product \( i \) and market \( j \),
the sample naturally entails two-way dependence. Specifically, for each product \( i \), \( \{W_{ij}\}_{j=1}^{M} \) is likely dependent through a supply shock by the producer of product \( i \). Similarly, for each market \( j \), \( \{W_{ij}\}_{i=1}^{N} \) is likely dependent through a demand shock in market \( j \). 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. Let \( \{P_n\}_n \) 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 \( P = P_n \in P_n \) 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 \( C = N \wedge M, \mu_N = C/N, \) and \( \mu_M = C/M, \) and suppose that \( \mu_N \to \mu_N, \mu_M \to \mu_M. \) 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 \( \ell_2 \) or Euclidean norm of \( v \). For any matrix \( A, \|A\| \) denotes the induced \( \ell_2 \)-norm of the matrix. For any set \( B, |B| \) denotes the cardinality of the set.

We state the following assumption on multiway clustered sampling.

**Assumption 1 (Sampling).** Suppose \( C \to \infty \). 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 \( \pi_1 \) and \( \pi_2 \) of \( \mathbb{N} \), we have

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

(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 $\delta_n \geq C^{-1/2}$. Let $K \geq 2$ be a fixed integer. Let $W_{00}$ denote a copy of $W_{11}$ that is independent from the data and the random set $T_n$ of nuisance realization. With these notations, we consider the following assumptions.

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

(i) The true parameter value $\theta_0$ satisfies (2.1).

(ii) $\psi$ is linear in the sense that it satisfies (2.4).

(iii) The map $\eta \mapsto E_P[\psi(W_{00}; \theta, \eta)]$ is twice continuously Gateaux differentiable on $T$.

(iv) $\psi$ satisfies either the Neyman orthogonality condition (2.2) or more generally the Neyman $\lambda_n$ near orthogonality condition at $(\theta_0, \eta_0)$ with respect to a nuisance realization set $T_n \subset T$ as

$$
\lambda_n := \sup_{\eta \in T_n} \| \partial_\eta E_P[\psi(W_{00}; \theta_0, \eta_0)[\eta - \eta_0]] \| \leq \delta_n C^{-1/2}.
$$

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

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

(i) Given random subsets $I \subset [N]$ and $J \subset [M]$ such that $|I| \times |J| = [NM/K^2]$, the nuisance parameter estimator $\hat{\eta} = \hat{\eta}((W_{ij})_{(i,j) \in I \times J})$, where the complements are taken with respect to $[N]$ and $[M]$, respectively, belongs to the realization set $T_n$ with probability at least $1 - \Delta_n$, where $T_n$ contains $\eta_0$. 

11
The following moment conditions hold:

\[ m_n := \sup_{\eta \in \mathcal{T}_n} \mathbb{E}_P[\|\psi(W_{00}; \theta_0, \eta)\|^q]^{1/q} \leq c_1, \]
\[ m'_n := \sup_{\eta \in \mathcal{T}_n} \mathbb{E}_P[\|\psi^a(W_{00}; \eta)\|^q]^{1/q} \leq c_1. \]

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

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

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 3** (Discussion of the Assumptions). Assumption 1 is similar to those of the preceding work on multiway cluster robust inference (cf. Menzel, 2017; Davezies et al., 2018; Chiang and Sasaki, 2019). Menzel (2017) does not invoke the dissociation, and follows an alternative approach to inference. The other papers assume both the separate exchangeability and dissociation, and conduct unconditional inference as in this paper. See Kallenberg (2006, Corollary 7.23 and Lemma 7.35) for representations with and without the dissociation under the separate exchangeability. 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 4** (Partial Distributions). Assumptions 2 and 3 state conditions based on \( W_{00} \), differently from CCDDHNR (2018), because of our 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 \( \delta_n \geq C^{-1/2} \) for all \( C \geq 1 \), then

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

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

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

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

\[
\sigma^2 := J_0^{-1}\Gamma(J_0^{-1})'.
\]  

(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 \( \sigma^2 \).

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 \( d_\theta \)-dimensional vector. The \((1 - a)\) confidence interval of \( r'\theta_0 \) given by

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

satisfies

\[
\sup_{P \in \mathcal{P}_n} |P_P(\theta_0 \in CI_a) - (1 - a)| \to 0.
\]
As in Section 3.4 of CCDDHNREF (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

$$Y_{ij} = D_{ij} \theta_0 + X_{ij} \xi_0 + \epsilon_{ij}$$

$$D_{ij} = Z_{ij} \pi_{10} + X_{ij} \pi_{20} + v_{ij},$$

$$Z_{ij} = X_{ij} \xi_0 + V_{ij},$$

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

$$X_{ij} = (1 - \omega_1^X - \omega_2^X) \alpha_{ij}^X + \omega_1^X \alpha_i^X + \omega_2^X \alpha_j^X,$$

$$\epsilon_{ij} = (1 - \omega^\epsilon_1 - \omega^\epsilon_2) \alpha^\epsilon_{ij} + \omega^\epsilon_1 \alpha_i^\epsilon + \omega^\epsilon_2 \alpha_j^\epsilon,$$

$$v_{ij} = (1 - \omega^V_1 - \omega^V_2) \alpha^V_{ij} + \omega^V_1 \alpha_i^V + \omega^V_2 \alpha_j^V,$$

and

$$V_{ij} = (1 - \omega^V_1 - \omega^V_2) \alpha^V_{ij} + \omega^V_1 \alpha_i^V + \omega^V_2 \alpha_j^V$$

with two-way clustering weights $(\omega_1^X, \omega_2^X), (\omega_1^\epsilon, \omega_2^\epsilon), (\omega_1^V, \omega_2^V)$, and $(\omega_1^V, \omega_2^V)$, where $\alpha_{ij}^X$, $\alpha_i^X$, and $\alpha_j^X$ are independently generated according to

$$\alpha_{ij}^X, \alpha_i^X, \alpha_j^X \sim N(0, \begin{pmatrix}
    s_0^X & s_1^X & \cdots & s_{\dim(X)-2}^X & s_{\dim(X)-1}^X \\
    s_1^X & s_0^X & \cdots & s_{\dim(X)-3}^X & s_{\dim(X)-2}^X \\
    \vdots & \vdots & \ddots & \vdots & \vdots \\
    s_{\dim(X)-2}^X & s_{\dim(X)-3}^X & \cdots & s_0^X & s_1^X \\
    s_{\dim(X)-1}^X & s_{\dim(X)-2}^X & \cdots & s_1^X & s_0^X
  \end{pmatrix})$$
\((\alpha_{ij}^\epsilon, \alpha_{ij}^V)', (\alpha_i^\epsilon, \alpha_i^V)', \text{ and } (\alpha_j^\epsilon, \alpha_j^V)'\) 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 \(\alpha_{ij}^V, \alpha_i^V, \text{ and } \alpha_j^V\) are independently generated according to

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

The weights \((\omega_1^X, \omega_2^X), (\omega_1^\epsilon, \omega_2^\epsilon), (\omega_1^V, \omega_2^V), \text{ and } (\omega_1^Y, \omega_2^Y)\) specify the extent of dependence in two-way clustering in \(X_{ij}, \epsilon_{ij}, v_{ij}, \text{ and } V_{ij}\), 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 \((\omega_1^X, \omega_2^X) = (\omega_1^\epsilon, \omega_2^\epsilon) = (\omega_1^V, \omega_2^V) = (0.25, 0.25)\) 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 \((N, M, C, \text{ and } \operatorname{dim}(X))\). 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 \(\hat{\eta}_{hl}\). 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 \(\operatorname{dim}(X) \in \{100, 200\}\), 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 \(C = N \wedge M\) increases: 1) the bias tends to zero; 2) the standard deviation decreases approximately at the \(\sqrt{C}\) 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. This result is perhaps specific to the data generating process used for our simulations. 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 Empirical Illustration: Demand Analysis with Market Share Data

Let us revisit the demand model of Example 1 in Section 2.3. Recall that, for the consumer demand model of Berry (1994) introduced in Example 1, Lu et al. (2019, Equation (9)) derive the partial-linear equation

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

(5.1)

for estimation of $\theta_0$, 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 in market $j$, $D_{ij}$ denotes the log price of product $i$ in market $j$, and $X_{ij}$ denotes a vector of observed attributes of product $i$ in market $j$. To deal with the likely endogeneity of $D_{ij}$, researchers often use instruments $Z_{ij}$ such that $E_P[\epsilon_{ij} | X_{ij}, Z_{ij}] = 0$. Such instruments often consist of observed attributes of other products in the market.

The implied equation (5.1) together with this mean independence assumption yields the reduced-form model (2.7). Furthermore, we write the innocuous nonparametric projection equation (2.8). Therefore, we apply Algorithm 1 in Section 2.3 for the two-way cluster robust DML estimation of $\theta_0$ with a robust standard error.

We present an application of the proposed algorithm to the U.S. automobile data of Berry, Levinsohn, and Pakes (1995). The sample consists of unbalanced two-way clustered observations with $N = 557$ models of
automobiles and $M = 20$ markets. The observed attributes $X_{ij}$ consist of horsepower per weight, miles per dollar, miles per gallon, and size. The instrument $Z_{ij}$ is defined as the sum of the values of these attributes of other products.

For the purpose of highlighting the effect of clustering assumptions, we report estimates and standard errors under the zero-way cluster robust DML (based on the i.i.d. assumption) and the one-way cluster robust DML (based on clustering along each of the product and market dimensions), as well as the two-way cluster robust DML (along both of the product and market dimensions). The number $K = 4$ of folds of cross fitting is used for the zero- and one-way cluster robust DML, while the number $K^2 = 4$ of folds of two-way cross fitting is used for the two-way cluster robust DML following the recommendations from Section 4 and those by CCDDHNR (2018, Remark 3.1). To mitigate the uncertainty induced by sample splitting, we compute estimates based on the average of ten rerandomized DML following CCDDHNR (2018, Section 3.4) with variance estimation according to CCDDHNR (2018, Equation 3.13) adapted to our two-way cluster-robustness.

Table 2 summarizes the results. For each of the zero-, one-, and two-way cluster robust DML, both the point estimates and standard errors are similar across all the choices of instrument. Furthermore, the point estimates are also similar across all of the zero-, one-, and two-way cluster robust DML. On the other hand, the standard errors tend to increase as the assumed number of ways of clustering increases. In other words, the zero-way cluster robust DML reports the smallest standard error while the two-way cluster robust DML reports the largest standard error. To robustly account for possible cross-sectional dependence of observations in such two-way cluster sampled data as this market share data, we recommend that researchers use the two-way cluster robust DML although it may incur larger standard errors as is the case with this application.

6 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.

There are a couple of possible directions for future research. First, whereas we focused on linear orthogonal scores that cover a wide range of applications, it may be possible to develop a method and theories for non-linear orthogonal scores as in CCDDHNR (2018; Section 3.3). Second, whereas we focused on unconditional moment restrictions, it may be possible and important to develop a method and theories for conditional moment restrictions (Ai and Chen, 2003, 2007; Chen, Linton, and Van Keilegom, 2003; Chen and Pouzo, 2015). We leave these and other extensions for future research.
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 \(E_n\) for the event \(\hat{\eta}_{k\ell} \in T_n\) for all \(k, \ell \in [K]^2\). Assumption 3 (i) implies \(P(E_n) \geq 1 - K^2 \Delta_n\).

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

\[
\hat{J} := \frac{1}{K^2} \sum_{(k, \ell) \in [K]^2} E_n, k\ell[\psi^a(W; \hat{\eta}_{k\ell})], \quad R_{n,1} := \hat{J} - J_0,
\]

\[
R_{n,2} := \frac{1}{K^2} \sum_{(k, \ell) \in [K]^2} E_n, k\ell[\psi(W; \theta_0, \hat{\eta}_{k\ell})] - \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \psi(W_{ij}; \theta_0, \eta_0).
\]

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

\[
\|R_{n,1}\| = O_P_n(C^{-1/2} + r_n), \quad (A.1)
\]

\[
\|R_{n,2}\| = O_P_n(C^{-1/2} r_n' + \lambda_n + \lambda_n'), \quad (A.2)
\]

\[
\left\| \sqrt{C}(NM)^{-1} \sum_{i=1}^{N} \sum_{j=1}^{M} \psi(W_{ij}; \theta_0, \eta_0) \right\| = O_P_n(1), \quad (A.3)
\]

\[
\|\sigma^{-1}\| = O_P_n(1). \quad (A.4)
\]

Then, under Assumptions 2 and 3, \(C^{-1/2} + r_N \leq \rho_n = o(1)\) 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
Furthermore, from (A.4) and (A.5), we have

\[
\sqrt{C}(\tilde{\theta} - \theta_0) = - \sqrt{C} \tilde{J}^{-1} \frac{1}{K^2} \sum_{(k,\ell) \in [K]^2} \mathbb{E}_{n,k\ell}[\psi^b(W; \tilde{n}_{k\ell})];
\]

and

\[
\sqrt{C}(\tilde{\theta} - \theta_0) = - \sqrt{C} \tilde{J}^{-1} \frac{1}{K^2} \sum_{(k,\ell) \in [K]^2} \left( \mathbb{E}_{n,k\ell}[\psi^b(W; \tilde{n}_{k\ell})] + J\theta_0 \right)
\]

\[
= - \sqrt{C} \tilde{J}^{-1} \frac{1}{K^2} \sum_{(k,\ell) \in [K]^2} \mathbb{E}_{n,k\ell}[\psi(W; \theta_0, \tilde{n}_{k\ell})]
\]

\[
= - \left( J_0 + R_{n,1} \right)^{-1} \times \left( \frac{\sqrt{C}}{N M} \sum_{i=1}^{N} \sum_{j=1}^{M} \psi(W_{ij}; \theta_0, \eta_0) + \sqrt{C} R_{n,2} \right). \tag{A.5}
\]

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

\[
\| (J_0 + R_{n,1})^{-1} - J_0^{-1} \| \leq \| (J_0 + R_{n,1})^{-1} R_{n,1} J_0^{-1} \| \leq \| (J_0 + R_{n,1})^{-1} \| \| R_{n,1} \| \| J_0^{-1} \|
\]

\[
= O_{P_n}(1) O_{P_n}(C^{-1/2} + r_n) O_{P_n}(1) = O_{P_n}(C^{-1/2} + r_n).
\]

Furthermore, \( r_n' + \sqrt{C} (\lambda_n + \lambda_n') \leq \rho_n = o(1) \), it holds that

\[
\left\| \frac{\sqrt{C}}{N M} \sum_{i=1}^{N} \sum_{j=1}^{M} \psi(W_{ij}; \theta_0, \eta_0) + \sqrt{C} R_{n,2} \right\| \leq \left\| \frac{\sqrt{C}}{N M} \sum_{i=1}^{N} \sum_{j=1}^{M} \psi(W_{ij}; \theta_0, \eta_0) \right\| + \left\| \sqrt{C} R_{n,2} \right\|
\]

\[
= O_{P_n}(1) + o_{P_n}(1) = O_{P_n}(1),
\]

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

\[
\left\| \left( J_0 + R_{n,1} \right)^{-1} - J_0^{-1} \right\| \times \left\| \frac{\sqrt{C}}{N M} \sum_{i=1}^{N} \sum_{j=1}^{M} \psi(W_{ij}; \theta_0, \eta_0) + \sqrt{C} R_{n,2} \right\| = O_{P_n}(C^{-1/2} + r_n) O_{P_n}(1)
\]

\[
= O_{P_n}(C^{-1/2} + r_n). \tag{A.6}
\]

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

\[
\sqrt{C} \sigma^{-1}(\tilde{\theta} - \theta_0) = \frac{\sqrt{C}}{N M} \sum_{i=1}^{N} \sum_{j=1}^{M} \tilde{\psi}(W_{ij}) + O_{P_n}(\rho_n).
\]
The first term on the RHS above can be written as $G_n \bar{\psi}$. Applying Lemma 1, we obtain the independent linear representation

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

and it holds $P_n$-a.s. that

$$V(G_n \bar{\psi}) = V(H_n \bar{\psi}) + O(C^{-1}) = J_0^{-1}(J_0^{-1})' + O(C^{-1})$$

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

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

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

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

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

$$\mathcal{I}_{1, k\ell} := \left| \mathbb{E}_{n, k\ell}[\psi^a(W; \hat{\eta}_{k\ell})] - \mathbb{E}_{P_n}[\psi^a(W_{ij}; \hat{\eta}_{k\ell})|\mathbb{I}_k \times J^c] \right|$$

$$\mathcal{I}_{2, k\ell} := \left| \mathbb{E}_{P_n}[\psi^a(W_{ij}; \hat{\eta}_{k\ell})|\mathbb{I}_k \times J^c] - \mathbb{E}_{P_n}[\psi^a(W_{11}; \eta_0)] \right|.$$ 

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 $\tilde{\psi}^a_{ij,m} = \psi^a_m(W_{ij}; \hat{\eta}_{k\ell}) - \mathbb{E}_{P_n}[\psi^a_m(W_{ij}; \hat{\eta}_{k\ell})|\mathbb{I}_k \times J^c]$ and $\tilde{\psi}^a_{ij} = (\tilde{\psi}^a_{ij,m})_{m \in [d_\theta]}$. To bound $\mathcal{I}_{1, k\ell}$,
note that conditional on $I^c_k \times J^c_\ell$, it holds that

$$
\mathbb{E}_P_n[\mathcal{I}_1^{2} \mid I^c_k \times J^c_\ell] = \mathbb{E}_P_n\left[\left\| \mathbb{E}_{n,k}[\psi^a(W; \tilde{\eta}_{k\ell})] - \mathbb{E}_P_n[\psi^a(W_{ij}; \tilde{\eta}_{k\ell})] \right\|^2 \mid I^c_k \times J^c_\ell \right]
$$

$$
= \frac{1}{(|I| \wedge |J|)^2} \mathbb{E}_P_n \left[ \sum_{m=1}^{d_\theta} \left( \sum_{(i,j) \in I_k \times J_\ell} \tilde{\psi}^a_{ij,m} \right)^2 \mid I^c_k \times J^c_\ell \right]
$$

$$
= \frac{1}{(|I| \wedge |J|)^2} \sum_{(i,j) \in I_k \times J_\ell} \sum_{j' \in J_k, j' \neq j} \mathbb{E}_P_n \left[ \sum_{m=1}^{d_\theta} \tilde{\psi}^a_{ij,m} \tilde{\psi}^a_{ij',m} \mid I^c_k \times J^c_\ell \right]
$$

$$
+ \frac{1}{(|I| \wedge |J|)^2} \sum_{(i,j) \in I_k \times J_\ell} \sum_{i' \in I_k, i' \neq i} \mathbb{E}_P_n \left[ \sum_{m=1}^{d_\theta} \tilde{\psi}^a_{ij',m} \tilde{\psi}^a_{ij',m} \mid I^c_k \times J^c_\ell \right]
$$

$$
= \frac{1}{(|I| \wedge |J|)^2} \sum_{(i,j) \in I_k \times J_\ell} \sum_{j' \in J_k, j' \neq j} \mathbb{E}_P_n \left[ \sum_{m=1}^{d_\theta} \tilde{\psi}^a_{ij,j'} \mid I^c_k \times J^c_\ell \right]
$$

$$
+ \frac{1}{(|I| \wedge |J|)^2} \sum_{(i,j) \in I_k \times J_\ell} \sum_{i' \in I_k, i' \neq i} \mathbb{E}_P_n \left[ \sum_{m=1}^{d_\theta} \tilde{\psi}^a_{ij,i'} \mid I^c_k \times J^c_\ell \right]
$$

$$
\leq \frac{1}{|I| \wedge |J|} \mathbb{E}_P_n \left[ \left\| \psi^a(W_{ij}; \tilde{\theta}_{k\ell}) - \mathbb{E}_P_n[\psi^a(W_{ij}; \tilde{\eta}_{k\ell})] \right\|^2 \mid I^c_k \times J^c_\ell \right]
$$

$$
\leq \frac{1}{|I| \wedge |J|} \mathbb{E}_P_n \left[ \left\| \psi^a(W_{ij}; \tilde{\theta}_{k\ell}) \right\|^2 \mid I^c_k \times J^c_\ell \right]
$$

$$
\leq \frac{c_1^2}{|I| \wedge |J|}
$$

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

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

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

Denote

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

22
where $\phi$ is $P_n$ an integrable function on $\text{supp}(W)$. Then

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

where

$$
\mathcal{I}_{3,k\ell} := \left\| \mathbb{G}_{n,k\ell}[\psi(W; \theta_0, \tilde{\eta}_{k\ell})] - \mathbb{G}_{n,k\ell}[\psi(W; \theta_0, \eta_0)] \right\|
$$

$$
\mathcal{I}_{4,k\ell} := \sqrt{C} \left[ \mathbb{E}_{P_n}[\psi(W_{ij}; \theta_0, \tilde{\eta}_{k\ell})|I_k \times J_\ell] - \mathbb{E}_{P_n}[\psi(W_{11}; \theta_0, \eta_0)] \right].
$$

Denote $\tilde{\psi}_{ij,m} := \psi_m(W_{ij}; \theta_0, \tilde{\eta}_{k\ell}) - \psi_m(W_{ij}; \theta_0, \eta_0)$ 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

$$
\mathbb{E}_{P_n}[||\mathcal{I}_{3,k\ell}||^2 | I^c_k \times J_k^c] = \mathbb{E}_{P_n}[||\mathbb{G}_{n,k\ell}[\psi(W_{ij}; \theta_0, \tilde{\eta}_{k\ell})] - \psi(W_{ij}; \theta_0, \eta_0)||^2 | I^c_k \times J_k^c]
$$

$$
= \mathbb{E}_{P_n} \left[ \frac{C}{(|I||J|)^2} \sum_{m=1}^{d_\theta} \left( \sum_{(i,j) \in I_k \times J_\ell} \left( \tilde{\psi}_{ij,m} - \mathbb{E}_{P_n} \tilde{\psi}_{ij,m} \right) \right)^2 | I^c_k \times J_k^c \right]
$$

$$
= \mathbb{E}_{P_n} \left[ \frac{C}{(|I||J|)^2} \sum_{(i,j) \in I_k \times J_\ell} \sum_{j' \in J_\ell, j' \neq j} \mathbb{E}_{P_n} \left[ \left( \tilde{\psi}_{ij,j',m} - \mathbb{E}_{P_n} \tilde{\psi}_{ij,j',m} \right) \left( \tilde{\psi}_{ij',m} - \mathbb{E}_{P_n} \tilde{\psi}_{ij',m} \right) \right] | I^c_k \times J_k^c \right]
$$

where the first inequality follows from Cauchy-Schwartz’s inequality, the second-to-last equality is due to Assumption $\mathbb{H}$ and the last equality is due to Assumption $\mathbb{M}$ (iii).
Hence, $I_{3,k\ell} = O_{P_n}(r_n'')$. To bound $I_{4,k\ell}$, let

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

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

$$\|f'_{k\ell}(0)\| = \left\| \partial_{\eta_0} E_{P_n}[\psi(W; \theta_0, \eta_0)|I_k^c \times J_\ell^c] \right\| \leq \lambda_n$$

under Assumption \[2\] (iv). Moreover, under Assumption \[3\] (iii), on the event $E_n$, 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

$$E_{P_n}\left[ \left\| \frac{\sqrt{C}}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \psi(W_{ij}; \theta_0, \eta_0) \right\|^2 \right] = \frac{C}{(NM)^2} E_{P_n}\left[ \sum_{m=1}^{d_\phi} \left( \sum_{i=1}^{N} \sum_{j=1}^{M} \psi_m(W_{ij}; \theta_0, \eta_0) \right)^2 \right]$$

$$= \frac{C}{(NM)^2} \sum_{i=1}^{N} \sum_{1 \leq j < j' \leq M} E_{P_n}\left[ \sum_{m=1}^{d_\phi} \psi_m(W_{ij}; \theta_0, \eta_0) \psi_m(W_{ij'; \theta_0, \eta_0}) \right]$$

$$+ \frac{C}{(NM)^2} \sum_{1 \leq i < i' \leq N} \sum_{j=1}^{M} E_{P_n}\left[ \sum_{m=1}^{d_\phi} \psi_m(W_{ij'; \theta_0, \eta_0}) \right]$$

$$+ \frac{C}{(NM)^2} \sum_{i=1}^{N} \sum_{j=1}^{M} E_{P_n}\left[ \sum_{m=1}^{d_\phi} \psi_m^2(W_{ij'; \theta_0, \eta_0}) \right] + 0$$

$$\lesssim E_{P_n} \left[ \left\| \psi(W_{ij}; \theta_0, \eta_0) \right\|^2 \right] \leq c_1^2$$

under Assumptions \[4\] 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 $\Gamma$ 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).
A.2 Proof of Theorem 2

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

\[
\left\| \frac{1}{K^2} \sum_{(k, \ell) \in [K]^2} \left\{ \frac{|I| \wedge |J|}{(|I| |J|)^2} \sum_{i \in I_k, j' \in J_k} \psi(W_{ij}; \tilde{\theta}, \tilde{\eta}_{k\ell}) \psi(W_{ij'}; \tilde{\theta}, \tilde{\eta}_{k\ell})' + \frac{|I| \wedge |J|}{(|I| |J|)^2} \sum_{i,j \in I_k} \sum_{j' \in J_k} \psi(W_{ij}; \tilde{\theta}, \tilde{\eta}_{k\ell}) \psi(W_{ij'}; \tilde{\theta}, \tilde{\eta}_{k\ell})' \right\} \right. \\
- \bar{\mu}_N E_P[\psi(W_{11}; \theta_0, \eta_0) \psi(W_{12}; \theta_0, \eta_0)'] - \bar{\mu}_M E_P[\psi(W_{21}; \theta_0, \eta_0) \psi(W_{21}; \theta_0, \eta_0)'] = O_P(\rho_n).
\]

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

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

and

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

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

\[
\mathcal{I}_{kl,lm} \leq \mathcal{I}_{kl,lm,1} + \mathcal{I}_{kl,lm,2},
\]

where

\[
\mathcal{I}_{kl,lm,1} := \left\| \frac{1}{|I|^2 |J|} \sum_{i,j' \in I_k, j \in J_k} \sum_{j' \in J_k} \left\{ \psi_l(W_{ij}; \tilde{\theta}, \tilde{\eta}_{k\ell}) \psi_m(W_{ij'}; \tilde{\theta}, \tilde{\eta}_{k\ell}) - \psi_l(W_{ij}; \theta_0, \eta_0) \psi_m(W_{ij'}; \theta_0, \eta_0) \right\} \right\|
\]

\[
\mathcal{I}_{kl,lm,2} := \left\| \frac{1}{|I|^2 |J|} \sum_{i,j' \in I_k, j \in J_k} \sum_{j' \in J_k} \psi_l(W_{ij}; \theta_0, \eta_0) \psi_m(W_{ij'}; \theta_0, \eta_0) - E_P[\psi_l(W_{11}; \theta_0, \eta_0) \psi_m(W_{21}; \theta_0, \eta_0)] \right\|.
\]
We first find a bound for $I_{kt,lm,2}$. Since $q > 4$, it holds that

$$E_P[I_{kt,lm,2}^2] = \frac{1}{|I|^4} \frac{1}{|J|^2} E_P \left[ \sum_{i,j} \sum_{i',j'} \psi_i(W_{ij}; \theta_0, \eta_0) \psi_j(W_{ij}; \theta_0, \eta_0) - E_P[\psi(W_{ij}; \theta_0, \eta_0) \psi_j(W_{ij}; \theta_0, \eta_0)]^2 \right]$$

$$\leq \frac{1}{|I|^4} \frac{1}{|J|^2} E_P \left[ \sum_{i,i',j,j'} \psi_i(W_{ij}; \theta_0, \eta_0) \psi_j(W_{ij}; \theta_0, \eta_0) \psi_i(W_{ij}; \theta_0, \eta_0) \psi_j(W_{ij}; \theta_0, \eta_0) \right]$$

$$+ \frac{1}{|I|^4} \frac{1}{|J|^2} E_P \left[ \sum_{i,i',j,j'} \psi_i(W_{ij}; \theta_0, \eta_0) \psi_j(W_{ij}; \theta_0, \eta_0) \psi_i(W_{ij}; \theta_0, \eta_0) \psi_j(W_{ij}; \theta_0, \eta_0) \right]$$

$$+ \frac{1}{|I|^4} \frac{1}{|J|^2} E_P \left[ \sum_{i,i',j,j'} \psi_i(W_{ij}; \theta_0, \eta_0) \psi_j(W_{ij}; \theta_0, \eta_0) \psi_i(W_{ij}; \theta_0, \eta_0) \psi_j(W_{ij}; \theta_0, \eta_0) \right]$$

$$+ o((|I| \land |J|)^{-2}) = O(C^{-1/2}) .$$

Now, to bound $I_{kt,lm,1}$, 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 $|a| \lor |b| \leq c$ and $|\delta a| \lor |\delta b| \leq r$, it holds that $|(a + \delta a)(b + \delta b) - ab| \leq 2r(c + r)$. Denote $\psi_{ij,h} := \psi_i(W_{ij}; \theta_0, \eta_0)$ and $\hat{\psi}_{ij,h} := \psi_i(W_{ij}; \theta, \eta)$ for $h \in \{l, m\}$ and apply the above identity with $a = \psi_{ij,l}, b = \psi_{ij,m}, a + \delta a = \hat{\psi}_{ij,l}, b + \delta b = \hat{\psi}_{ij,m}, r = |\hat{\psi}_{ij,l} - \psi_{ij,l}| \lor |\hat{\psi}_{ij,m} - \psi_{ij,m}|$. Then

$$I_{kt,lm,1} = \frac{1}{|I|^2} \frac{1}{|J|} \sum_{i,i' \in I_k \land J_\ell} \sum_{j,j' \in J_\ell} \left( |\hat{\psi}_{ij,l} \hat{\psi}_{ij,m} - \psi_{ij,l} \psi_{ij,m}| \right)$$

$$\leq \frac{1}{|I|^2} \frac{1}{|J|} \sum_{i,i' \in I_k \land J_\ell} \sum_{j,j' \in J_\ell} \left( |\hat{\psi}_{ij,l} - \psi_{ij,l}| \lor |\hat{\psi}_{ij,m} - \psi_{ij,m}| \right)$$

$$\leq \frac{2}{|I|^2} \frac{1}{|J|} \sum_{i,i' \in I_k \land J_\ell} \sum_{j,j' \in J_\ell} \left( |\hat{\psi}_{ij,l} - \psi_{ij,l}| \lor |\hat{\psi}_{ij,m} - \psi_{ij,m}| \right) \left( |\hat{\psi}_{ij,l} \lor |\psi_{ij,m}| + |\hat{\psi}_{ij,m} \lor |\psi_{ij,l}| \right)$$

$$\leq \left( \frac{2}{|I|^2} \frac{1}{|J|} \sum_{i,i' \in I_k \land J_\ell} \sum_{j,j' \in J_\ell} \left( |\hat{\psi}_{ij,l} - \psi_{ij,l}| \lor |\hat{\psi}_{ij,m} - \psi_{ij,m}| \right) \right)^{1/2}$$

$$\leq \left( \frac{2}{|I|^2} \frac{1}{|J|} \sum_{i,i' \in I_k \land J_\ell} \sum_{j,j' \in J_\ell} \left( |\hat{\psi}_{ij,l} - \psi_{ij,l}| \lor |\hat{\psi}_{ij,m} - \psi_{ij,m}| \right) \right)^{1/2}$$

$$\leq \left( \frac{2}{|I|^2} \frac{1}{|J|} \sum_{i,i' \in I_k \land J_\ell} \sum_{j,j' \in J_\ell} \left( |\hat{\psi}_{ij,l} - \psi_{ij,l}| \lor |\hat{\psi}_{ij,m} - \psi_{ij,m}| \right) \right)^{1/2}$$

$$\leq \left( \frac{2}{|I|^2} \frac{1}{|J|} \sum_{i,i' \in I_k \land J_\ell} \sum_{j,j' \in J_\ell} \left( |\hat{\psi}_{ij,l} - \psi_{ij,l}| \lor |\hat{\psi}_{ij,m} - \psi_{ij,m}| \right) \right)^{1/2}$$

$$\leq \frac{1}{|I|^2} \frac{1}{|J|} \sum_{i,i' \in I_k \land J_\ell} \sum_{j,j' \in J_\ell} \left( |\hat{\psi}_{ij,l} - \psi_{ij,l}| \lor |\hat{\psi}_{ij,m} - \psi_{ij,m}| \right)$$

$$+ \left( \frac{2}{|I|^2} \frac{1}{|J|} \sum_{i,i' \in I_k \land J_\ell} \sum_{j,j' \in J_\ell} \left( |\hat{\psi}_{ij,l} - \psi_{ij,l}| \lor |\hat{\psi}_{ij,m} - \psi_{ij,m}| \right) \right)^{1/2} .$$
Thus, the above bound for $I_{R}$ under Assumptions 1 and 3 (ii). Finally, to bound the inequality. Notice that

$$\sum_{i,i' \in I_k \times J_\ell} |\psi_{ij}|^2 \vee |\psi_{ij,m}|^2 \leq |I| \sum_{i,j=1}^{N,M} \|\psi(W_{ij}; \theta_0, \eta_0)\|^2,$$

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

Thus, the above bound for $I_{k\ell,m,1}$ implies that

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

where

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

which is implied by Markov’s inequality and the calculations

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

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

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

The first term on RHS is bounded by

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

due to Assumption 3 (ii), Markov’s inequality, and Theorem 1. Furthermore, given that $(W_{ij})_{(i,j) \in I_k \times J_\ell}$ satisfies $\tilde{\eta}_{k\ell} \in \mathcal{T}_n$,

$E_P[\|\psi(W_{ij}; \theta_0, \tilde{\eta}_{k\ell}) - \psi(W_{ij}; \theta_0, \eta_0)\|^2 |I_k \times J_\ell] \leq \sup_{\eta \in \mathcal{T}_n} E_P[\|\psi(W_{ij}; \theta_0, \eta) - \psi(W_{ij}; \theta_0, \eta_0)\|^2 |I_k \times J_\ell] \leq (r_n)^2$
due to Assumptions \(1\) and \(3\) (iii). Also, the event \(\bar{\eta}_{k\ell} \in \mathcal{T}_n\) happens with probability \(1 - o(1)\), we have \(R_n = O_P(\mathcal{C}^{-1} + \langle r_n'\rangle^2)\). Thus we conclude that

\[
\mathcal{I}_{k\ell,l,m,1} = O_P(\mathcal{C}^{-1/2} + r_n').
\]

This completes the proof. \(\square\)

B Useful Lemmas

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

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

\[
\mathcal{G}_n f := \sqrt{\mathcal{C}} \left\{ \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} f(W_{ij}) - E_P[f(W_{11})] \right\}
\]

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 \(2\) holds and \(f : \text{supp}(W) \rightarrow \mathbb{R}^d\) for some fixed \(d \in \mathbb{N}\) and suppose \(E_P\|f(W_{11})\|^2 < K\) 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 \(\mathcal{G}_n f\) 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

\[
H_n f = \frac{\sqrt{\mathcal{C}}}{N} \sum_{i=1}^{N} E_P \left[ f(W_{i1}) - E_P f(W_{11}) \mid U_{i0} \right] + \frac{\sqrt{\mathcal{C}}}{M} \sum_{j=1}^{M} E_P \left[ f(W_{1j}) - E_P f(W_{11}) \mid U_{0j} \right]
\]

for each \(n\). Furthermore,

\[
V(\mathcal{G}_n f) = V(H_n f) + O(\mathcal{C}^{-1}) = \tilde{\mu}_N \text{Cov}(f(W_{11}), f(W_{12})) + \tilde{\mu}_M \text{Cov}(f(W_{11}), f(W_{21})) + O(\mathcal{C}^{-1})
\]

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 CCDDHN (2018):

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

(i) If for \(\epsilon_n \to 0\), \(P(\|X_n\| > \epsilon_n|Y_n) = o_P(1)\) 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 \(l_n \to \infty\), \(P(\|X_n\| > l_n A_n) = o(1)\).

## C Extension to General Multiway Clustering

In this section, we extend the main results to general multiway cluster sampling framework. Notations in the current section are independent of those in the remaining parts of the paper – we introduce different notations in order to enhance the readability of the main results of the paper while economizing complicated notations in the current extension section. Consider the \(\ell\)-way clustered data for a fixed dimension \(\ell \in \mathbb{N}\). With \(C_i \in \mathbb{N}\) denoting the number of clusters in the \(i\)-th cluster dimension for each \(i \in \{1, ..., \ell\}\), each cell of the \(\ell\)-way clustered sample is indexed by the \(\ell\)-dimensional multiway cluster indices \(j = (j_1, ..., j_\ell) \in \times_{i=1}^\ell [C_i]\). The \(\ell\)-dimensional size \((C_1, ..., C_\ell) \in \mathbb{N}^\ell\) of the \(\ell\)-way clustered sample will be index by \(n \in \mathbb{N}\) as \((C_1, ..., C_\ell) = (C_1(n), ..., C_\ell(n))\), where \(C_i(n)\) is non-decreasing in \(n\) for each \(i \in \{1, ..., \ell\}\) and \(\prod_{i=1}^\ell C_i(n)\) is increasing in \(n\). With this said, we will suppress the index notation and write \((C_1, ..., C_\ell)\) without \(n\) for simplicity. Also define the notations \(C = (C_1, ..., C_\ell)\), \(\prod C = \prod_{i=1}^\ell C_i\), \(\underline{C} = \min_{1 \leq i \leq \ell} C_i\), \(\overline{C} = \max_{1 \leq i \leq \ell} C_i\), and \(\mu_i = \underline{C}/C_i\) for each \(i \in \{1, ..., \ell\}\). Suppose that \(\mu_i \to \bar{\mu}_i\) for some constant \(\bar{\mu}_i\) for each \(i \in \{1, ..., \ell\}\). The number of observations in the \(j\)-th cell is denoted by \(N_j\), which is treated as an \([0, 1, ..., N]\)-valued random variable for some \(N \in \mathbb{N}\) not depending on \(n\). When \([\cdot]\) takes the random variable \(N_j\) as an argument, we extend the definition of \([\cdot]\) to \([N_j] := \{1, ..., N_j\}\) if \(N_j \geq 1\) and \([\cdot] := \emptyset\) if \(N_j = 0\). The observed vector for unit \(i \in [N_j]\) in the \(j\)-th cell is denoted by \(W_{i,j}\). Let \(\{P_n\}_n\) be a sequence of sets of probability laws of \((N_j, (W_{i,j})_{1 \leq i \leq N_j})_{\geq 1}\), where \(1 := (1, ..., 1)\) for a short-hand notation and we write \(j \geq j'\) to mean \(j_i \geq j'_i\) for all \(i \in \{1, ..., \ell\}\).
Example 2. The sampling setting in Section 2.1 fits in the current general framework with $\ell = 2$, $C_1 = N$, $C_2 = M$, and $(N_j, W_{i,j}) = (1, W_{j1,2})$ for all $j \in [N] \times [M]$ with probability one. \(\triangle\)

The econometric model has the true parameters $(\theta_0, \eta_0) \in \Theta \times T$ satisfying the score moment restriction

$$E_P \left[ \sum_{i=1}^{N_1} \psi(W_{i,1}; \theta_0, \eta_0) \right] = 0,$$

where we focus on the linear Neyman orthogonal score of the form

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

for $\text{supp}(W) := \bigcup_{i=1}^{N_1} \text{supp}(W_{i,1})$, $\Theta \subset \mathbb{R}^{d_\theta}$ and a convex set $T$.

For a fixed integer $K > 1$, we randomly split the data into $K$ folds in each of the $\ell$ cluster dimensions, resulting in $K^\ell$ folds in total. Specifically, randomly partition $[C_i]$ into $K$ parts $\{I_{i}^1, ..., I_{i}^{K}\}$ for each $i \in \{1, ..., \ell\}$. We use the $\ell$-dimensional indices $k := (k_1, ..., k_\ell)$ to index the $\ell$-way fold $I_k := I_{k_1} \times \cdots \times I_{k_\ell}$ and its complementary product $I_k^c := I_{k_1}^c \times \cdots \times I_{k_\ell}^c$ for each $k \in [K]^\ell$. Let $
_k = \hat{\eta}((W_{i,j})_{i \in [N_j]} j \in I_k^c)$ be a machine learning estimate of $\eta$ using the subsample $((W_{i,j})_{i \in [N_j]} j \in I_k^c)$ for each $k \in [K]^\ell$. Let

$$\tilde{J} := \frac{1}{K^\ell} \sum_{k \in [K]^\ell} \mathbb{E}_{n,k} \left[ \sum_{i \in [N_j]} \psi^a(W_{i,j}; \n_k) \right]$$

where

$$\mathbb{E}_{n,k} \left[ \sum_{i \in [N_j]} f(W_{i,j}) \right] := \frac{1}{|I_k|} \sum_{j \in I_k} \sum_{i \in [N_j]} f(W_{i,j}) \text{ for each } k \in [K]^\ell$$

for any Borel measurable function $f$, the sum $\sum_{i \in [N_j]}$ is treated as zero when $N_j = 0$, and $|I_k| := \lfloor \prod_{i=1}^{\ell} C_i/K^\ell \rfloor$.

With these setup and notations, the multiway DML estimator is defined by

$$\tilde{\theta} = -\tilde{J}^{-1} \frac{1}{K^\ell} \sum_{k \in [K]^\ell} \mathbb{E}_{n,k} \left[ \sum_{i \in [N_j]} \psi^b(W_{i,j}; \n_k) \right].$$

(C.3)

Let $|I_k| = \min\{|I_{k_1}|, ..., |I_{k_\ell}|\}$ for a short-hand notation. Also let $I(\hat{j})$ denote the multiway fold containing the $\hat{j}$-th multiway cluster, i.e., $I(\hat{j}) \subset \times_{i=1}^{\ell} [C_i]$ satisfies $I_k = I(\hat{j})$ for some $k \in [K]^\ell$ and $\hat{j} \in I(\hat{j})$. With these additional notations, we propose to estimate the asymptotic variance of $\sqrt{C}(\tilde{\theta} - \theta_0)$
by
\[
\hat{\sigma}^2 = \tilde{J}^{-1} \left[ \frac{1}{K^\ell} \sum_{k \in [K]^\ell} \frac{|I_k|}{|I_k|^2} \sum_{j,j' \in I_k, j \neq j'} \sum_{w \in [N] \setminus I_k} \sum_{v' \in [N]} \psi(W_{i,j}; \tilde{\theta}, \tilde{\eta}) \psi(W_{i,j'}; \tilde{\theta}, \tilde{\eta}) ' \right] (\tilde{J}^{-1})'. \tag{C.4}
\]

**Example 2 Continued.** The two-way DML in Section 2.2 is a special case of the current general methodological framework with \( \{ I_1, ..., I_K \} = \{ I_1, ..., I_1 \}, \{ I_2, ..., I_K \} = \{ J_1, ..., J_K \}, \tilde{\eta}_{(k_1, k_2)} = \hat{\eta}(W_{j_1, j_2})_{(j_1, j_2) \in ([N] \setminus I_k) \times ([M] \setminus J_k)}, \tilde{J} = \frac{1}{K^2} \sum_{(j_1, j_2) \in [K]^2} \mathbb{E}_{n,(k_1, k_2)} \left[ \psi^a(W_{j_1, j_2}; \tilde{\eta}(k_1, k_2)) \right] \right] where \( \mathbb{E}_{n,(k_1, k_2)} \left[ \psi^a(W_{j_1, j_2}; \tilde{\eta}(k_1, k_2)) \right], \) and \( \tilde{\sigma}^2 = \tilde{J}^{-1} \hat{\Gamma}(\tilde{J}^{-1})' \) where \( \hat{\Gamma} = \frac{1}{K^2} \sum_{(k_1, k_2) \in [K]^2} \left( \frac{1}{|I_k| |J_k|^2} \right)^2 \sum_{(j_1, j_2) \in [K]^2} \sum_{(j_1, j_2) \in [K]^2} \psi(W_{j_1, j_2}; \tilde{\theta}, \tilde{\eta}(k_1, k_2)) \psi(W_{j_1, j_2}; \tilde{\theta}, \tilde{\eta}(k_1, k_2))' \right) \Delta \)

We now state assumptions under which (C.4) is an asymptotically valid variance estimator for \( \sqrt{\mathbb{C}(\hat{\theta} - \theta_0)} \) with the multiway DML estimator (C.3). 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 \( \ell_2 \) or Euclidean norm of \( v \). For any matrix \( A \), \( \| A \| \) denotes the induced \( \ell_2 \)-norm of the matrix. The following assumption concerns the multiway clustered sampling.

**Assumption 4 (Sampling).** The following conditions hold for each \( n \).

(i) The array \( (N_j, (W_{i,j})_{1 \leq i \leq N})_{j \geq 1} \) is an infinite sequence of separately exchangeable random vector.

That is, for any \( \ell \)-tuple of permutations \( (\pi_1, ..., \pi_\ell) \) of \( \mathbb{N} \), we have
\[
(N_j, (W_{i,j})_{1 \leq i \leq N})_{j \geq 1} \overset{d}{=} (N_{\pi_1(j_1)}, ..., \pi_\ell(j_\ell)), (W_{i,\pi_1(j_1)}, ..., \pi_\ell(j_\ell))_{1 \leq i \leq N})_{j \geq 1}.
\]

(ii) \( (N_j, (W_{i,j})_{1 \leq i \leq N})_{j \geq 1} \) is dissociated. That is, for any \( c \geq 1 \), \( (N_j, (W_{i,j})_{1 \leq i \leq N})_{1 \leq j \leq c} \) is independent of \( (N_{j'}, (W_{i',j'})_{1 \leq i' \leq N})_{j' \geq c+1} \)

(iii) \( E(N_1) > 0 \) and \( N_j \leq \overline{N} \) for each \( 1 \leq j \leq C \), where \( \overline{N} \in \mathbb{N} \) does not depend on \( n \).

(iv) The econometrician observes \( (N_j, (W_{i,j})_{1 \leq i \leq N_j})_{1 \leq j \leq C} \).

**Remark 5.** The dependence among \( (W_{i,j})_{1 \leq i \leq N} \) in each cell \( j \) is left unrestricted in this assumption. Assumption 4 is similar to Assumption 1 of Davezies et al. (2018), except for \( \overline{N} \). We introduce \( \overline{N} \) to simplify some concentration arguments.
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 $\delta_n \geq C^{-1/2}$. Let $K \geq 2$ be a fixed integer. Let $(N_0, (W_i, 0)_{0 \leq i \leq N})$ denote an independent copy of $(N_1, (W_i, 1)_{1 \leq i \leq N})$ and therefore is independent from the data and the random set $\mathcal{T}_n$ of nuisance realization. With these notations, we state the following assumptions for the model.

**Assumption 5** (Linear Neyman Orthogonal Score). For all $C \geq 3$ and $P \in \mathcal{P}_n$, the following conditions hold.

(i) The true parameter value $\theta_0$ satisfies (C.1).

(ii) $\psi$ is linear in the sense that it satisfies (C.2).

(iii) The map $\eta \mapsto E_P \left[ \sum_{i \in [N_0]} \psi(W_i, 0; \theta, \eta) \right]$ is twice continuously Gateaux differentiable on $T$.

(iv) $\psi$ satisfies the Neyman near orthogonality condition at $(\theta_0, \eta_0)$ as

$$\lambda_n := \sup_{\eta \in \mathcal{T}_n} \left\| \partial_\eta E_P \left[ \sum_{i \in [N_0]} \psi(W_i, 0; \theta_0, \eta_0) \right] \right\| \leq \delta_n C^{-1/2}.$$  

(v) The identification condition holds as the singular values of the matrix $J_0 := E_P \left[ \sum_{i \in [N_0]} \psi^a(W_i, 0; \eta_0) \right]$ are between $c_0$ and $c_1$.

**Assumption 6** (Score Regularity and Nuisance Parameter Estimators). For all $C \geq 3$ and $P \in \mathcal{P}_n$, the following conditions hold.

(i) The realization set $\mathcal{T}_n$ contains $\eta_0$, and the nuisance parameter estimator $\hat{\eta}_k = \hat{n}(W_i, j)_{i \in [N_0], j \in I_k}$ belongs to the realization set $\mathcal{T}_n$ for each $k \in [K]^\ell$ with probability at least $1 - \Delta_n$.

(ii) The following moment conditions hold:

$$m_n := \sup_{\eta \in \mathcal{T}_n} (E_P \left[ \left\| \sum_{i \in [N_0]} \psi(W_i, 0; \theta, \eta) \right\|^q \right]^{1/q}) \leq c_1,$$

$$m'_n := \sup_{\eta \in \mathcal{T}_n} (E_P \left[ \left\| \sum_{i \in [N_0]} \psi^a(W_i, 0; \eta) \right\|^q \right]^{1/q}) \leq c_1.$$
(iii) The following conditions on the rates $r_n, r'_n$ and $\lambda'_n$ hold:

\[
  r_n := \sup_{\eta \in T_n} \left\| E_P \left[ \sum_{i \in [N_0]} \psi'(W_i,0;\eta) \right] - E_P \left[ \sum_{i \in [N_0]} \psi(W_i,0;\eta_0) \right] \right\| \leq \delta_n,
\]

\[
  r'_n := \sup_{\eta \in T_n} \left( \left\| E_P \left[ \sum_{i \in [N_0]} \psi(W_i,0;\theta_0,\eta) \right] - E_P \left[ \sum_{i \in [N_0]} \psi(W_i,0;\theta_0,\eta_0) \right] \right\|^2 \right)^{1/2} \leq \delta_n,
\]

\[
  \lambda'_n = \sup_{r \in (0,1), \eta \in T_n} \left\| \partial^2 E_P \left[ \sum_{i \in [N_0]} \psi(W_i,0;\theta_0,\eta_0 + r(\eta - \eta_0)) \right] \right\| \leq \delta_n / \sqrt{C}.
\]

(iv) All eigenvalues of the matrix

\[
  \Gamma := \sum_{i=1}^{\ell} \bar{\mu}_{i} \Gamma_i = \sum_{i=1}^{\ell} \bar{\mu}_{i} E_P \left[ \sum_{i=1}^{N_1} \sum_{i'=1}^{N_2} \psi(W_i,0;\theta_0,\eta_0) \psi(W_{i'},2;\theta_0,\eta_0) \right]
\]

are bounded from below by $c_0$, where $2_i$ denotes the $\ell$-tuple vector with 2 in each entry but for 1 in the $i$-th entry.

The following theorems generalize Theorems 1 and 2 to cover general $\ell$-way cluster sampling. Their proofs are contained in Section D.

**Theorem 3** (Main Result). Suppose that Assumptions 4, 5, and 6 are satisfied. If $\delta_n \geq C^{-1/2}$ for all $C \geq 1$, then

\[
  \sqrt{C} \sigma^{-1}(\bar{\theta} - \theta_0) = \sqrt{C} \sum_{j=1}^{C} \sum_{j=1}^{C} \sum_{i=1}^{N_j} \tilde{\psi}(W_i,j) + O_P(\rho_n) \sim N(0, I_{d_\theta})
\]

holds uniformly for all $P \in \mathcal{P}_n$, where $\prod_{C} = \prod_{i=1}^{\ell} C_i$, the influence function takes the form $\tilde{\psi}(\cdot) := -\sigma^{-1} J_0^{-1} \psi(\cdot;\theta_0,\eta_0)$, the size of the remainder terms follows

\[
  \rho_n := C^{-1/2} + r_n + r'_n + C_1^{1/2} \lambda_n + C_1^{1/2} \lambda'_n \lesssim \delta_n,
\]

and the asymptotic variance is given by

\[
  \sigma^2 := J_0^{-1} \Gamma(J_0^{-1})'.
\]

(C.5)

**Theorem 4** (Variance Estimator). Under the assumptions required by Theorem 3, we have

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

Furthermore, the statement of Theorem 3 holds true with $\hat{\sigma}^2$ in place of $\sigma^2$. 33
D Proofs of the Extended Results

D.1 Proof of Theorem \[\text{3}\]

Proof. Let \( \mathcal{E}_n \) denote the event \( \tilde{\eta}_{(k_1, \ldots, k_\ell)} \in \mathcal{T}_n \) for all \((k_1, \ldots, k_\ell) \in [K]^\ell \) and define \( k := (k_1, \ldots, k_\ell) \). Assumption \(6\) (i) implies \( P_n(\mathcal{E}_n) \geq 1 - K^\ell \Delta_n \). Let \( e \in \{0, 1\}^\ell \), and define \( A_e := \{(j, j') : 1 \leq j, j' \leq C : \forall i = 1, \ldots, \ell, e_i = 1 \Leftrightarrow j_i = j'_i \} \), and \( \epsilon_m := \{e \in \{0, 1\}^\ell : \sum_{i=1}^\ell e_i = m\} \).

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

\[
\tilde{J} := \frac{1}{K^\ell} \sum_{k \in [K]^\ell} \mathbb{E}_{n,k} \left[ \sum_{i \in [N_j]} \psi^a(W_{i,j}; \tilde{\eta}_k) \right], \quad R_{n,1} := \tilde{J} - J_0,
\]

\[
R_{n,2} := \frac{1}{K^\ell} \sum_{k \in [K]^\ell} \mathbb{E}_{n,k} \left[ \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \hat{\eta}_0) \right] - \frac{1}{\prod_{j=1}^C} \sum_{j_1=1}^{C_1} \cdots \sum_{j_\ell=1}^{C_\ell} \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta_0).
\]

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

\[
\|R_{n,1}\| = O_{P_n}(C^{-1/2} + r_n), \quad \text{(D.1)}
\]

\[
\|R_{n,2}\| = O_{P_n}(C^{-1/2} r_n + \lambda_n + \lambda'_n), \quad \text{(D.2)}
\]

\[
\left\| \sqrt{C} \frac{1}{\prod_{j=1}^C} \sum_{j_1=1}^{C_1} \cdots \sum_{j_\ell=1}^{C_\ell} \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta_0) \right\| = O_{P_n}(1), \quad \text{(D.3)}
\]

\[
\|\sigma^{-1}\| = O_{P_n}(1). \quad \text{(D.4)}
\]

Then, under Assumptions \(5\) and \(\mathbb{E}_n \leq \rho_n = o(1)\) 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 \(\tilde{J}\) are bounded away from zero. Thus with the same \(P_n\) probability, the multiway DML solution is uniquely written as

\[
\tilde{\theta} = -\tilde{J}^{-1} \frac{1}{K^\ell} \sum_{k \in [K]^\ell} \mathbb{E}_{n,k} \left[ \sum_{i \in [N_j]} \psi^b(W_{i,j}; \tilde{\eta}_k) \right],
\]

34
Therefore, from (D.4), (D.5) and (D.6), we have

\[
\sqrt{\bar{C}}(\tilde{\theta} - \theta_0) = - \frac{1}{\sqrt{C} \bar{J}} \sum_{k \in [K]} \mathbb{E}_{n,k} \left[ \sum_{i \in [N_j]} \psi^b(W_{i,j}; \tilde{\eta}_k) + \tilde{J}_0 \right]
\]

and

\[
- \frac{1}{\sqrt{C} \bar{J}} \sum_{k \in [K]} \mathbb{E}_{n,k} \left[ \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \tilde{\eta}_k) \right]
\]

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

\[
|| (J_0 + R_{n,1})^{-1} - J_0^{-1} || = || (J_0 + R_{n,1})^{-1} R_{n,1} J_0^{-1} || \leq ||(J_0 + R_{n,1})^{-1}|| ||R_{n,1}|| ||J_0^{-1}||
\]

\[
= O_{P_n}(1) O_{P_n}(C^{-1/2} + r_n) = O_{P_n}(C^{-1/2} + r_n).
\]

Furthermore, \( r_n' + \sqrt{C}(\lambda_n + \lambda_n') \leq \rho_n = o(1), \) it holds that

\[
\left\| \frac{\sqrt{C}}{\prod C} \sum_{j_1 = 1}^{C_1} \ldots \sum_{j_\ell = 1}^{C_\ell} \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta_0) + \sqrt{C} R_{n,2} \right\| \leq \left\| \frac{\sqrt{C}}{\prod C} \sum_{j_1 = 1}^{C_1} \ldots \sum_{j_\ell = 1}^{C_\ell} \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta_0) \right\| + \left\| \sqrt{C} R_{n,2} \right\|
\]

\[
= O_{P_n}(1) + o_{P_n}(1) = O_{P_n}(1),
\]

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

\[
\left\| \left( J_0 + R_{n,1} \right)^{-1} - J_0^{-1} \right\| \times \left\| \frac{\sqrt{C}}{\prod C} \sum_{j_1 = 1}^{C_1} \ldots \sum_{j_\ell = 1}^{C_\ell} \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta_0) + \sqrt{C} R_{n,2} \right\| = O_{P_n}(C^{-1/2} + r_n) O_{P_n}(1)
\]

\[
= O_{P_n}(C^{-1/2} + r_n).
\]

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

\[
\sqrt{C} \sigma^{-1}(\tilde{\theta} - \theta_0) = \frac{\sqrt{C}}{\prod C} \sum_{j_1 = 1}^{C_1} \ldots \sum_{j_\ell = 1}^{C_\ell} \sum_{i \in [N_j]} \tilde{\psi}(W_{i,j}) + O_{P_n}(\rho_n).
\]

The first term on the RHS above can be written as \( \mathbb{G}_n \tilde{\psi} \). Applying Lemma 3, we obtain the independent linear representation

\[
H_n \tilde{\psi} := \sum_{j_1 = 1}^{C_1} \frac{\sqrt{C}}{C_1} \mathbb{E}_{P_n} \left[ \sum_{i \in [N_j]} \tilde{\psi}(W_{i,j}) | U_{j_1,0 \ldots 0} \right] + \ldots + \sum_{j_\ell = 1}^{C_\ell} \frac{\sqrt{C}}{C_\ell} \mathbb{E}_{P_n} \left[ \sum_{i \in [N_j]} \tilde{\psi}(W_{i,j}) | U_{0 \ldots 0,j_\ell} \right]
\]

35
and it holds \( P_n \)-a.s. that
\[
V_n(G_n \bar{\psi}) = V_n(H_n \bar{\psi}) + O(C^{-1}) = J_0^{-1} \Gamma(J_0^{-1})' + O(C^{-1}) \quad \text{and}
\]
\[
G_n \bar{\psi} = H_n \bar{\psi} + O_P(C^{-1/2}),
\]
where \( V_n(\cdot) = E_{P_n}[(\cdot - E_{P_n}[\cdot])^2] \). Under Assumption 6 (iv). Recall that \( q \geq 4 \), the third moments of both summands of \( H_n \bar{\psi} \) are bounded over \( n \) under Assumptions 5(v) and 6 (ii) (iv). We have verified all the conditions for Lyapunov’s CLT. An application of Lyapunov’s CLT and Cramer-Wold device gives
\[
H_n \bar{\psi} \rightsquigarrow N(0, I_{d_0})
\]
and an application of Theorem 2.7 of van der Vaart (1998) concludes the proof.

\textbf{Step 2.} Since \( K \) is fixed, it suffices to show for any \( k \in [K]^\ell \),
\[
\left\| E_{n,k} \left[ \sum_{i \in [N]} \psi^a(W_{i,j}; \eta_k) \right] - E_P \left[ \sum_{i \in [N_0]} \psi^a(W_{i,0}; \eta_0) \right] \right\| = O_{P_n}(C^{-1/2} + r_n).
\]

Fix \( k \in [K]^\ell \),
\[
\left\| E_{n,k} \left[ \sum_{i \in [N]} \psi^a(W_{i,j}; \eta_k) \right] - E_{P_n} \left[ \sum_{i \in [N_0]} \psi^a(W_{i,0}; \eta_0) \right] \right\| \leq I_{1,k} + I_{2,k},
\]
where
\[
I_{1,k} := \left\| E_{n,k} \left[ \sum_{i \in [N]} \psi^a(W_{i,j}; \eta_k) \right] - E_{P_n} \left[ \sum_{i \in [N]} \psi^a(W_{i,j}; \eta_k) I_{k_1}^c \times \ldots \times I_{k_\ell}^c \right] \right\|,
\]
\[
I_{2,k} := \left\| E_{P_n} \left[ \sum_{i \in [N]} \psi^a(W_{i,j}; \eta_k) I_{k_1}^c \times \ldots \times I_{k_\ell}^c \right] - E_{P_n} \left[ \sum_{i \in [N_0]} \psi^a(W_{i,0}; \eta_0) \right] \right\|.
\]

Notice that \( I_{2,k} \leq r_n \) with \( P_n \)-probability \( 1 - o(1) \) follows directly from Assumptions 4 (ii) and 4 (iii).

Now denote
\[
\bar{\psi}_{j,m}^a = \sum_{i \in [N]} \psi^a(W_{i,j}; \eta_k) - E_{P_n} \left[ \sum_{i \in [N]} \psi^a(W_{i,j}; \eta_k) I_{k_1}^c \times \ldots \times I_{k_\ell}^c \right] \quad \text{and} \quad \bar{\psi}_j^a = (\bar{\psi}_{j,m}^a)_{m \in [d_0]},
\]
and \( |I_k| = \min\{|I_{k_1}|, \ldots, |I_{k_\ell}|\} \). Let us denote \( I_k : (I_{k_1} \times \ldots \times I_{k_\ell}) \) and \( I_k^c : (I_{k_1}^c \times \ldots \times I_{k_\ell}^c) \). Let \( j \mapsto I(j) \in I \), and define
\[
B_e := \{(j, j') : \forall i = 1, \ldots, \ell, e_i = 1 \iff I_i(j) = I_i(j') : j, j' \in I \}, \quad I := \{I_1^1, \ldots, I_{K_1}^1\} \times \ldots \times \{I_\ell^1, \ldots, I_{K_\ell}^\ell\}, \quad \text{and} \quad e_m := \{e \in (0, 1) : \sum_{i=1}^\ell e_i = m\}.
\]

To bound \( I_{1,k} \), note that
conditional on $I_k^c$, it holds that

$$
E_n \left[ I^2_{1,k} \mid I_k^c \right] = E_n \left[ \left\| \sum_{i \in [N]} \psi^a(W_{i,j}; \tilde{\eta}_k) \right\|^2 \mid I_k^c \right]
$$

$$
= \frac{1}{|I_k|^2} \sum_{m=1}^{d_q} \sum_{j \in I_k} \left( \sum_{m=1}^{d_q} \tilde{\psi}_{j,m}^a \right)^2 \left| I_k^c \right|
$$

$$
= \frac{1}{|I_k|^2} \sum_{e \in e_1} \sum_{(j', j) \in B_e} E_n \left[ \sum_{m=1}^{d_q} \tilde{\psi}_{j,m}^a \tilde{\psi}_{j',m}^a \mid I_k^c \right]
$$

$$
+ \frac{1}{|I_k|^2} \sum_{e \in e_r} \sum_{(j', j) \in B_e} E_n \left[ \sum_{m=1}^{d_q} \tilde{\psi}_{j,m}^a \tilde{\psi}_{j',m}^a \mid I_k^c \right]
$$

$$
+ \frac{1}{|I_k|^2} \sum_{e \in e_0} \sum_{(j', j) \in B_e} E_n \left[ \sum_{m=1}^{d_q} \tilde{\psi}_{j,m}^a \tilde{\psi}_{j',m}^a \mid I_k^c \right]
$$

$$
= \frac{1}{|I_k|^2} \sum_{e \in e_1} \sum_{(j', j) \in B_e} E_n \left[ \langle \tilde{\psi}_{j'}^a, \tilde{\psi}_{j}^a \rangle \mid I_k^c \right] + R + 0
$$

$$
\leq \frac{1}{|I_k|^2} \sum_{e \in e_1} \sum_{(j', j) \in B_e} E_n \left[ \left\| \sum_{i \in [N]} \psi^a(W_{i,j}; \tilde{\eta}_k) \right\|^2 \mid I_k^c \right]
$$

$$
\leq \frac{1}{|I_k|^2} \sum_{e \in e_1} \sum_{(j', j) \in B_e} E_n \left[ \left\| \sum_{i \in [N]} \psi^a(W_{i,j}; \tilde{\eta}_k) \right\|^2 \mid I_k^c \right] \leq c^2 \frac{1}{|I_k|^2}.
$$

In the third equality, the last term corresponds to the covariance between cells sharing no common cluster. By independence, the last term is zero. Let us denote the second term in the third equality by $R$. Under Cauchy-Schwarz inequality and Assumption 4 (ii),

$$
|R| \leq \frac{1}{|I_k|^2} \sum_{e \in e_1} |B_e| E_n \left[ \sum_{m=1}^{d_q} \left( \tilde{\psi}_{j,m}^a \right)^2 \mid I_k^c \right].
$$

(D.7)

For $r \geq 1$ and $e \in e_r$, we have

$$
|B_e| = |I_k| \times \prod_{i : e_i = 0} (|I_{k_i}| - 1).
$$

(D.8)

Therefore, $R = O(|I_k|^{-2})$. Note that $C \leq |I_k| \leq C$. Hence an application of Lemma 2 (i) implies $I_{1,k} = O_P_n(C^{-1/2})$. This completes a proof of (D.1).

**Step 3.** It again suffices to show that for any $k \in [K]^e$, one has

$$
\left\| \sum_{i \in [N]} \psi(W_{i,j}; \theta_0, \tilde{\eta}_k) \right\| \leq \frac{1}{|I_k^c|} \sum_{j \in I_k} \sum_{i \in [N]} \psi(W_{i,j}; \theta_0, \eta_0) = O_P(n(C^{-1/2} r_n' + \lambda_n' + \lambda_n)).
$$

37
Denote
\[ G_{n,k} \left[ \sum_{i \in [N_j]} \phi(W_{i,j}) \right] = \frac{\sqrt{C}}{|I_k|} \sum_{j \in I_k} \sum_{i \in [N_j]} \left( \phi(W_{i,j}) - \int \phi(w) dP_n \right). \]

Then
\[ \left\| E_{n,k} \left[ \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \tilde{\eta}_k) \right] - \frac{1}{|I_k|} \sum_{j \in I_k} \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta_0) \right\| \leq \frac{I_{3,k} + I_{4,k}}{\sqrt{C}}, \]

where
\[ I_{3,k} := \left\| G_{n,k} \left[ \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \tilde{\eta}_k) \right] - G_{n,k} \left[ \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta_0) \right] \right\|, \]
\[ I_{4,k} := \sqrt{C} \left\| E_{P_n} \left[ \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \tilde{\eta}_k) \right] I_k \right. - \left. E_{P_n} \left[ \sum_{i \in [N_0]} \psi(W_{i,0}; \theta_0, \eta_0) \right] \right\|. \]

Denote \( \tilde{\psi}_{j,m} := \sum_{i \in [N_j]} \psi_m(W_{i,j}; \theta_0, \tilde{\eta}_k) - \sum_{i \in [N_j]} \psi_m(W_{i,j}; \theta_0, \eta_0) \) and \( \tilde{\psi}_j = (\tilde{\psi}_{j,m})_{m \in [d_0]} \). To bound \( I_{3,k} \), notice that using a similar argument as for the bound of \( I_{1,k} \), one has
\[ E_{P_n} \left[ \| I_{3,k} \|^2 \right] = E_{P_n} \left[ \left\| G_{n,k} \left[ \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \tilde{\eta}_k) - \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta_0) \right] \right\|^2 \right] \]
\[ = E_{P_n} \left[ \frac{C}{|I_k|^2} \sum_{m=1}^{d_0} \left\{ \sum_{j \in I_k} \left( \tilde{\psi}_{j,m} - E_{P_n} \tilde{\psi}_{j,m} \right) \right\}^2 \right] \]
\[ = \frac{C}{|I_k|^2} \sum_{e \in E_1, (j_j') \in E_{B_0}} E_{P_n} \left[ \sum_{m=1}^{d_0} \left( \tilde{\psi}_{j,m} - E_{P_n} \tilde{\psi}_{j,m} \right) \left( \tilde{\psi}_{j',m} - E_{P_n} \tilde{\psi}_{j',m} \right) \right] \]
\[ + \frac{C}{|I_k|^2} \sum_{r=2}^{\ell} \sum_{e \in E_r, (j_j') \in E_{B_0}} E_{P_n} \left[ \sum_{m=1}^{d_0} \left( \tilde{\psi}_{j,m} - E_{P_n} \tilde{\psi}_{j,m} \right) \left( \tilde{\psi}_{j',m} - E_{P_n} \tilde{\psi}_{j',m} \right) \right] \]
\[ = \frac{C}{|I_k|^2} \sum_{e \in E_1, (j_j') \in E_{B_0}} E_{P_n} \left[ \left( \tilde{\psi}_j - E_{P_n} \tilde{\psi}_j \right) \left( \tilde{\psi}_{j'} - E_{P_n} \tilde{\psi}_{j'} \right) \right] + R' + 0 \]
\[ \leq E_{P_n} \left[ \left\| \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \tilde{\eta}) - \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta_0) \right\|^2 \right] \]
\[ \leq E_{P_n} \left[ \left\| \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \tilde{\eta}) - \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta_0) \right\|^2 \right] \]
\[ = \sup_{\eta \in T_n} E_{P_n} \left[ \left\| \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta) - \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta) \right\|^2 \right] = (r_n')^2, \]
where the first inequality follows from Cauchy-Schwartz’s inequality, the second-to-last equality is due to Assumption \[4\] and the last equality is due to Assumption \[6\](iii). Using the similar argument for \(R\), we have \(R' = O(C^{-1})\).

Hence, \(I_{3,k} = O_{P_n}(r'_n)\). To bound \(I_{4,k}\), let
\[
f_k(r) := E_{P_n}\left[ \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta_0 + r(\hat{\eta}_k - \eta_0)) \right] - E_{P_n}\left[ \sum_{i \in [N_0]} \psi(W_{i,0}; \theta_0, \eta_0) \right], \quad r \in [0, 1].
\]
An application of the mean value expansion coordinate-wise gives
\[
f_k(1) = f_k(0) + f'_k(0) + f''_k(\tilde{r})/2,
\]
where \(\tilde{r} \in (0, 1)\). Note that \(f_k(0) = 0\) under Assumption \[5\](i), and
\[
\|f_k'(0)\| = \left\| \partial_\eta E_{P_n}\left[ \sum_{i \in [N_j]} \psi(W; \theta_0, \eta_0)[\hat{\eta}_k - \eta_0] \right] \right\| \leq \lambda_n
\]
under Assumption \[5\](iv). Moreover, under Assumption \[6\](iii), on the event \(E_n\), we have
\[
\|f''_k(\tilde{r})\| \leq \sup_{r \in (0,1)} \|f''_k(r)\| \leq \lambda_n.
\]
This completes a proof of \[D.2\].

**Step 4.** Note that
\[
E_{P_n}\left[ \left\| \frac{\sqrt{C}}{\prod C} \sum_{j_1=1}^{C_1} \cdots \sum_{j_\ell=1}^{C_\ell} \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta_0) \right\|^2 \right]
= \frac{C}{\prod C} E_{P_n}\left[ \sum_{m=1}^{d_0} \sum_{i \in [N_j]} \psi_m(W_{i,0}; \theta_0, \eta_0)^2 \right]
= \frac{C}{\prod C} \sum_{e \in \mathcal{E}_1} \sum_{(j,j')} \sum_{(j,j') \in A_e} E_{P_n}\left[ \sum_{m=1}^{d_0} \sum_{i \in [N_j]} \sum_{i' \in [N_{j'}]} \psi_m(W_{i,j}; \theta_0, \eta_0) \psi_m(W_{i',j'}; \theta_0, \eta_0) \right]
+ \frac{C}{\prod C} \sum_{r=2}^{\ell} \sum_{e \in \mathcal{E}_r} \sum_{(j,j') \in A_e} E_{P_n}\left[ \sum_{m=1}^{d_0} \sum_{i \in [N_j]} \sum_{i' \in [N_{j'}]} \psi_m(W_{i,j}; \theta_0, \eta_0) \psi_m(W_{i',j'}; \theta_0, \eta_0) \right]
\leq E_{P_n}\left[ \left\| \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta_0) \right\|^2 \right] \leq c_1^2.
\]
**Step 5.** Note that all singular values of \( J_0 \) are bounded from above by \( c_1 \) under Assumption 5 (v) and all eigenvalues of \( \Gamma \) are bounded from below by \( c_0 \) under Assumption 6 (iv). Therefore, we have \( \|\sigma^{-1}\| \leq c_1/\sqrt{c_0} \) and thus \( \|\sigma^{-1}\| = O_P(1) \). This completes a proof of (D.4). \( \square \)

**D.2 Proof of Theorem 4**

**Proof.** Step 2 of the proof of Theorem 3 proves \( \|J - J_0\| = O_p(C^{-1/2} + r_n) \) and Assumption 5 (v) implies \( \|J_0^{-1}\| \leq c_0^{-1} \). Therefore, to prove the claim of the theorem, it suffices to show

\[
\left\| \frac{1}{K^\ell} \sum_{k \in [K]^{\ell}} \frac{|I_k|}{|k|} \sum_{i=1}^\ell \sum_{j,j' \in I_k \text{ such that } i \in [N_j], j' \in [N_{j'}]} \sum_{l,j,\theta_k} \psi(W_{i,j}; \theta_k) \psi(W_{i,j'}; \theta_k) \psi(W_{i,j'}; \theta_k) \psi(W_{i,j'}; \theta_k) \right\| = O_P(\rho_n).
\]

Moreover, since \( K \) and \( d_\theta \) are constants and \( \mu_i \to \mu_i \leq 1 \), it suffices to show that for each \( k \in [K]^{\ell} \) and \( l, m \in [d_\theta] \), it holds that

\[
\left| \frac{|I_k|}{|k|^2} \sum_{j,j' \in I_k \text{ such that } l,j,\theta_k} \sum_{i \in [N_j], j' \in [N_{j'}]} \psi(W_{i,j}; \theta_k) \psi(W_{i,j'}; \theta_k) \right| = O_P(\rho_n).
\]

Denote the left-hand side of the equation as \( I_{k,lm} \). First, note that \( |I_k|/|k| = \mu_i \). We denote \( i' \) for \( I_{k'} \) such that \( |I_{k'}| = |I_k| \), and apply the triangle inequality to get

\[
I_{k,lm} \leq I_{k,lm,1} + I_{k,lm,2},
\]

where

\[
I_{k,lm,1} := \frac{1}{\prod_{i \notin i'} |I_{k_i}|^2 |I_{k_{i'}}|} \sum_{j,j' \in I_k \text{ such that } l,j,\theta_k} \left\{ \sum_{i \in [N_j], j' \in [N_{j'}]} \psi(W_{i,j}; \theta_k) \psi(W_{i,j'}; \theta_k) \psi(W_{i,j'}; \theta_k) \right\},
\]

\[
I_{k,lm,2} := \frac{1}{\prod_{i \notin i'} |I_{k_i}|^2 |I_{k_{i'}}|} \sum_{j,j' \in I_k \text{ such that } l,j,\theta_k} \left| \sum_{i \in [N_j], j' \in [N_{j'}]} \psi(W_{i,j}; \theta_k) \psi(W_{i,j'}; \theta_k) \psi(W_{i,j'}; \theta_k) \right|.
\]
We first find a bound for $T_{k,lm,2}$. Since $q > 4$, it holds that

$$E_P[T_{k,lm,2}] = \prod_{i \neq i'} |I_{k_i}|^4 |I_{k_{i'}}|^2 E_P \left[ \sum_{j, j' \in I_k} \sum_{l, l' \in I_{j'}} \sum_{i \in [N_j]} \sum_{t' \in [N_j]} \psi_l(W_{s,j; \theta_0, \eta_0}) \psi_{m}(W_{r,j'; \theta_0, \eta_0}) \right. \left. - E_P \left[ \sum_{i=1}^{N_1} \sum_{t=1}^{N_2} \psi_l(W_{s,1; \theta_0, \eta_0}) \psi_{m}(W_{r,t; \theta_0, \eta_0}) \right] \right]^2 \right]$$

$$\leq \prod_{i \neq i'} |I_{k_i}|^4 |I_{k_{i'}}|^2 E_P \left[ \sum_{j, j' \in I_k} \sum_{l, l' \in I_{j'}} \sum_{i \in [N_j]} \sum_{t' \in [N_j]} \sum_{i'' \in [N_j']} \sum_{t'' \in [N_j']} \psi_l(W_{s,f; \theta_0, \eta_0}) \psi_{m}(W_{r,g'; \theta_0, \eta_0}) \psi_{m}(W_{r,g''; \theta_0, \eta_0}) \right]$$

$$+ \sum_{i \neq i'} |I_{k_i}|^4 |I_{k_{i'}}|^2 E_P \left[ \sum_{j, j' \in I_k} \sum_{l, l' \in I_{j'}} \sum_{i \in [N_j]} \sum_{t' \in [N_j]} \sum_{i'' \in [N_j']} \sum_{t'' \in [N_j']} \psi_l(W_{s,f; \theta_0, \eta_0}) \psi_{m}(W_{r,g'; \theta_0, \eta_0}) \psi_{m}(W_{r,g''; \theta_0, \eta_0}) \right]$$

$$+ o\left(\frac{|I_k|^{-1}}{c} + 0\right) \leq \frac{1}{|I_k|} E_P \left[ \sum_{i \in [N_0]} \psi(W_{s,0; \theta_0, \eta_0}) \right] \lesssim c_1^4/C = O(C^{-1}).$$

Now, to bound $T_{k,lm,1}$, 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 $\psi_{j,h} := \psi_l(W_{s,j; \theta_0, \eta_0})$ and $\tilde{\psi}_{j,h} := \psi_l(W_{s,j; \tilde{\theta}, \tilde{\eta}_k})$ for $h \in \{l, m\}$ and apply the above identity with $a = \sum_{i \in [N_j]} \psi_{j,l} - \sum_{i \in [N_j]} \psi_{j,l}$, $b = \sum_{i \in [N_j]} \psi_{j',m} - \sum_{i \in [N_j]} \psi_{j',m}$, $a+\delta a = \sum_{i \in [N_j]} \tilde{\psi}_{j,l} - \sum_{i \in [N_j]} \tilde{\psi}_{j',m}$, $b+\delta b = \sum_{i \in [N_j]} \tilde{\psi}_{j',m}$, $r = \left| \sum_{i \in [N_j]} \tilde{\psi}_{j,l} - \sum_{i \in [N_j]} \psi_{j,l} \right| \vee \left| \sum_{i \in [N_j]} \tilde{\psi}_{j',m} - \sum_{i \in [N_j]} \psi_{j',m} \right|$ and $c = \left| \sum_{i \in [N_j]} \tilde{\psi}_{j,l} \right| \vee \left| \sum_{i \in [N_j]} \tilde{\psi}_{j',m} \right|.$
\[
\mathcal{I}_{k,lm} = \left| \prod_{i \neq l} \frac{1}{|I_{k_l}|} \sum_{j,j' \in I_k} \sum_{l_j = l_{j'}} \left\{ \sum_{i \in [N_j]} \sum_{i' \in [N_{j'}]} \hat{\psi}_{j,l} \hat{\psi}_{j',m} - \sum_{i \in [N_j]} \sum_{i' \in [N_{j'}]} \psi_{j,l} \psi_{j',m} \right\} \right|
\]

\[
\leq \prod_{i \neq l} \frac{1}{|I_{k_l}|} \sum_{j,j' \in I_k} \sum_{l_j = l_{j'}} \left| \sum_{i \in [N_j]} \sum_{i' \in [N_{j'}]} \hat{\psi}_{j,l} \hat{\psi}_{j',m} - \sum_{i \in [N_j]} \sum_{i' \in [N_{j'}]} \psi_{j,l} \psi_{j',m} \right|
\]

\[
\leq \prod_{i \neq l} \frac{1}{|I_{k_l}|} \sum_{j,j' \in I_k} \sum_{l_j = l_{j'}} \left( \left| \sum_{i \in [N_j]} \psi_{j,l} \right| \vee \left| \sum_{i' \in [N_{j'}]} \psi_{j',m} \right| + \left| \sum_{i \in [N_j]} \hat{\psi}_{j,l} \right| \vee \left| \sum_{i' \in [N_{j'}]} \hat{\psi}_{j',m} \right| \right)
\]

\[
\times \left( \left( \sum_{i \in [N_j]} \psi_{j,l} \right) \vee \left( \sum_{i' \in [N_{j'}]} \psi_{j',m} \right) + \left( \sum_{i \in [N_j]} \hat{\psi}_{j,l} \right) \vee \left( \sum_{i' \in [N_{j'}]} \hat{\psi}_{j',m} \right) \right)
\]

\[
\leq \prod_{i \neq l} \frac{1}{|I_{k_l}|} \sum_{j,j' \in I_k} \sum_{l_j = l_{j'}} \left( \left| \sum_{i \in [N_j]} \psi_{j,l} \right| \vee \left| \sum_{i' \in [N_{j'}]} \psi_{j',m} \right| \right)
\]

\[
\times \left( \left( \sum_{i \in [N_j]} \psi_{j,l} \right) \vee \left( \sum_{i' \in [N_{j'}]} \psi_{j',m} \right) + \left( \sum_{i \in [N_j]} \hat{\psi}_{j,l} \right) \vee \left( \sum_{i' \in [N_{j'}]} \hat{\psi}_{j',m} \right) \right)
\]

where the second to the last inequality follows the Cauchy-Schwartz's inequality and Minkowski's
inequality. Notice that
\[
\sum_{j,j' \in I_k} \sum_{i \in [N]} \psi_{j,i} \mid \sum_{i' \in [N_j^1]} \psi_{j',i,m} \mid^2 \leq \max_{1 \leq i \leq \ell} \{ |I_k| \} \sum_{j_1 = 1}^{C_1} \ldots \sum_{j_\ell = 1}^{C_\ell} \left\| \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta_0) \right\|^2,
\]
\[
\sum_{j,j' \in I_k} \sum_{i \in [N]} \hat{\psi}_{j,i} - \sum_{i' \in [N_j]} \psi_{j,i} \mid \sum_{i' \in [N_j]} \hat{\psi}_{j',i,m} - \sum_{i' \in [N_j]} \psi_{j',i,m} \mid^2 \leq \max_{1 \leq i \leq \ell} \{ |I_k| \} \sum_{j_1 = 1}^{C_1} \ldots \sum_{j_\ell = 1}^{C_\ell} \left\| \sum_{i \in [N_j]} \psi(W_{i,j}; \bar{\theta}, \tilde{\eta}_k) - \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta_0) \right\|^2.
\]
Thus, the above bound for $I_{k,lm,1}$ implies that
\[
I^2_{k,lm,1} \lesssim R_n \times \left( \frac{1}{|I_k|} \sum_{j \in I_k} \left\| \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta_0) \right\|^2 + R_n \right),
\]
where
\[
R_n := \frac{1}{|I_k|} \sum_{j \in I_k} \left\| \sum_{i \in [N_j]} \psi(W_{i,j}; \bar{\theta}, \tilde{\eta}_k) - \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta_0) \right\|^2.
\]
Notice that
\[
\frac{1}{|I_k|} \sum_{j \in I_k} \left\| \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta_0) \right\|^2 = O_P(1),
\]
which is implied by Markov’s inequality and the calculations
\[
E_P \left[ \frac{1}{|I_k|} \sum_{j \in I_k} \left\| \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta_0) \right\|^2 \right] = E_P \left[ \left\| \sum_{i=1}^{N_0} \psi(W_{i,0}; \theta_0, \eta_0) \right\|^2 \right] \leq c_1^2
\]
under Assumptions $4$ and $6$ (ii). Finally, to bound $R_n$, using Assumption $3$ (ii),
\[
R_n \lesssim \frac{1}{|I_k|} \sum_{j \in I_k} \left\| \sum_{i \in [N_j]} \psi^a(W_{i,j}; \tilde{\eta}_k)(\bar{\theta} - \theta_0) \right\|^2 + \frac{1}{|I_k|} \sum_{j \in I_k} \left\| \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \tilde{\eta}_k) - \sum_{i \in [N_j]} \psi(W_{i,j}; \theta_0, \eta_0) \right\|^2.
\]
The first term on RHS is bounded by
\[
\left( \frac{1}{|I_k|} \sum_{j \in I_k} \left\| \sum_{i \in [N_j]} \psi^a(W_{i,j}; \tilde{\eta}_k) \right\|^2 \right) \times \| \bar{\theta} - \theta_0 \|^2 = O_P(1) \times O_P(C^{-1}) = O_P(C^{-1})
\]
43
due to Assumption 6 (ii), Markov’s inequality, and Theorem 3. Furthermore, given that \((W_i,j)_{j \in I_k}\) satisfies \(\hat{\eta}_k \in \mathcal{T}_n\),

\[
E_P \left[ \left\| \sum_{i \in [N]} \psi(W_{i,j}; \theta_0, \hat{\eta}_k) - \sum_{i \in [N]} \psi(W_{i,j}; \theta_0, \eta_0) \right\|^2 \right| I_k^c
\]

\[
\leq \sup_{\eta \in \mathcal{T}_n} E_P \left[ \left\| \sum_{i \in [N]} \psi(W_{i,j}; \theta_0, \eta) - \sum_{i \in [N]} \psi(W_{i,j}; \theta_0, \eta_0) \right\|^2 \right| I_k^c
\]

due to Assumptions 4 and 6 (iii). Also, the event \(\hat{\eta}_k \in \mathcal{T}_n\) happens with probability \(1 - o(1)\), we have

\[R_n = O_P(C^{-1} + (r'_n)^2)\]. Thus we conclude that

\[\mathcal{I}_{kn,1} = O_P(C^{-1/2} + r'_n).\]

This completes the proof.

\[\square\]

E Additional Lemma

In this section, we establish a multiway generalization of Lemma 1. For any \(r = 1, \ldots, \ell\), we let \(\mathcal{I}_r(\mathbf{C}) = \{\mathbf{c} = j \odot \mathbf{e} : \mathbf{e} \in \mathcal{E}_r, 1 \leq j \leq \mathbf{C}\}\) and \(\mathbf{e}_m = \{\mathbf{e} \in \{0,1\}^{\ell} : \sum_{i=1}^\ell e_i = m\}\), with \(\odot\) the Hadamard product on \(\mathbb{R}^\ell\).

For each \(n \in \mathbb{N}\), let \((N^n_j, (W^n_{i,j})_{1 \leq i \leq N^n_j})_{j \geq 1}\) be a set of random variables. For any \(f : \text{supp}(W^n) \to \mathbb{R}^d\) for a fixed \(d \in \mathbb{N}\), let us define the multiway empirical process

\[
G_{n,f} := \sqrt{C} \left\{ \frac{1}{\prod_i c_i} \sum_{j=1}^\ell \sum_{i=1}^{C_i} \sum_{e \in \mathcal{E}_r} f(W^n_{i,j}) - E_P \left[ \sum_{i \in [N^n_i]} f(W^n_{i,1}) \right] \right\}.
\]

**Lemma 3** (Independentization via Hájek Projections). For each \(n \in \mathbb{N}\), suppose that \(\left( N^n_j, (W^n_{i,j})_{1 \leq i \leq N^n_j} \right)_{j \geq 1}\) satisfies Assumption 4. Let \(\mathcal{F}_n, |\mathcal{F}_n| = d\), be a family of functions \(f : \text{supp}(W^n) \to \mathbb{R}\) that satisfies

\[E \left[ \left( \sum_{i \in [N^n_i]} f(W^n_{i,1}) \right)^2 \right] < K < \infty\] for some \(K\) independent of \(n\). In addition, assume that \(C \to \infty\) and for every \(\mathbf{e} \in \mathcal{E}_1, \frac{C}{\prod_i c_i} \to \mathbf{p}_j \geq 0\), where \(i\) is the nonzero coordinate of \(\mathbf{e}\). Then there exists a family of mutually independent standard uniform r.v.’s \((U_{c})_{c \geq 0}\) such that the \(H_n f\), the Hájek projection of \(G_{n,f}\) on the set of statistics of the form \(\sum_{c \in \mathcal{L}_r(\mathbf{C})} g_c(U_{c})\) (with \(g_c(U_{c})\) square integrable, satisfies

\[
H_n f = \sum_{c \in \mathcal{L}_r(\mathbf{C})} \frac{\sqrt{C}}{\prod_{i,e \neq 0} c_i} \left( E \left[ \sum_{i=1}^{N^n_{e,x}} f(W^n_{i,e,x}) \right] U_{c} - E \left[ \sum_{i \in [N^n_i]} f(W^n_{i,1}) \right] \right).
\]

(E.1)
In addition, it holds uniformly over $\mathcal{F}_n$ that

$$V(G_n f) = V(H_n f) + O(C^{-1}) = \sum_{e \in \varepsilon_1} \tilde{\mu}_e \text{Cov} \left( \sum_{i=1}^{N_i^n} f(W_{i,1}^n), \sum_{i=1}^{N_{2-e}^n} f(W_{i,2-e}^n) \right) + O(C^{-1}).$$

**Proof.** Throughout the proof, we drop the superscript $n$ for simplicity. Under Assumption $\text{H} (i)$ and (ii), for each $n$, one can apply Lemma 7.35 of Kallenberg (2006) and obtain a measurable function $\tau_n$ such that

$$(N_j, (W_{i,j})_{1 \leq i \leq N_j})_{j \geq 1} = (\tau_n(U_{j \in e})1_{e \in \varepsilon_1})_{j \geq 1} \quad (E.2)$$

where $(U_c)_{c \geq 0}$ denote a family of mutually independent uniform random variables on $[0, 1]$.

The rest of our proof closely follows that of Lemma D.2 in Davezies et al. (2018) with $r = r' = 1$.

The Hájek projection $H_n f$ is characterized by

$$E \left[ (G_n f - H_n f) \times \sum_{c \in \mathcal{I}_1(C)} g_c(U_c) \right] = 0 \text{ for any } (g_c)_{c \in \mathcal{I}_1(C)} \in \left( L^f([0; 1]) \right)^{|\mathcal{I}_1(C)|}.$$

As a result, we have

$$E [G_n f | U_c] = E [H_n f | U_c] \text{ for any } c \in \mathcal{I}_1(C).$$

Because the range $H_n$ is closed subspace of square integrable random variables,

$$H_n f = \sum_{c \in \mathcal{I}_1(C)} \mathbb{E} (H_n f | U_c).$$

Next

$$H_n f = \sum_{c \in \mathcal{I}_1(C)} \mathbb{E} (G_n f | U_c).$$

Note that for any $c \in \mathcal{I}_1(C)$, $c \wedge 1$ is the unique element $\varepsilon_1$ such that $c = j \circ e$ for some $j$ (note that $j$ is not unique). Moreover, for any $c \in \mathcal{I}_1(C)$ independence between the $U'$s ensures that

$$\sum_{i \in [N_j]} f(W_{i,j}) \perp U_c \text{ if } j \circ e \neq c. \text{ This implies}$$

$$\mathbb{E} (G_n f | U_c) = \frac{\sqrt{C}}{\Pi C} \sum_{1 \leq j \leq C} \mathbb{E} \left[ \sum_{i \in [N_j]} f(W_{i,j}) - \mathbb{E} \left[ \sum_{i \in [N_1]} f(W_{i,1}) \right] | U_c \right]$$

$$= \frac{\sqrt{C}}{\Pi C} \sum_{1 \leq j \leq C} 1 \{ j \circ e = c \} \mathbb{E} \left[ \sum_{i \in [N_j]} f(W_{i,j}) - \mathbb{E} \left[ \sum_{i \in [N_1]} f(W_{i,1}) \right] | U_c \right].$$

45
The representation of \((N_j, (W_{i,j})_{1 \leq i \leq N})_{j \geq 1}\) in terms of the \(U\)’s implies that
\[
\mathbb{E} \left[ \sum_{i=1}^{N_j} f(W_{i,j}) - \mathbb{E} \left[ \sum_{i \in [N_1]} f(N_1, W_{i,1}) \right] \right] | U_e = \mathbb{E} \left[ \sum_{i=1}^{N_{e,1}} f(W_{i,e\vee 1}) - \mathbb{E} \left[ \sum_{i \in [N_1]} f(W_{i,1}) \right] \right] | U_e
\]
for any \(j\) such that \(j \odot e = e\). Moreover,
\[
\mathbb{E} (G_n f | U_e) = \frac{\sqrt{C}}{\Pi C} \sum_{1 \leq j \leq C} \mathbb{1} \{ j \odot e = e \} \mathbb{E} \left[ \sum_{i=1}^{N_{e,1}} f(W_{i,e\vee 1}) - \mathbb{E} \left[ \sum_{i \in [N_1]} f(W_{i,1}) \right] \right] | U_e
\]
\[
= \frac{\sqrt{C}}{\Pi C} \sum_{i \in e, \epsilon \neq 0} C_i \mathbb{E} \left[ \sum_{i=1}^{N_{e,1}} f(W_{i,e\vee 1}) - \mathbb{E} \left[ \sum_{i \in [N_1]} f(W_{i,1}) \right] \right] | U_e
\]
\[
= \frac{\sqrt{C}}{\Pi C} \sum_{i \in e, \epsilon \neq 0} C_i \left( \mathbb{E} \left[ \sum_{i=1}^{N_{e,1}} f(W_{i,e\vee 1}) \right] | U_e - \mathbb{E} \left[ \sum_{i \in [N_1]} f(W_{i,1}) \right] \right).
\]

It follows that
\[
H_n f = \sum_{e \in \mathcal{E}_1} \frac{\sqrt{C}}{\Pi e_i = 1 C_i} \left( \mathbb{E} \left[ \sum_{i=1}^{N_{e,1}} f(W_{i,e\vee 1}) \right] | U_e - \mathbb{E} \left[ \sum_{i \in [N_1]} f(W_{i,1}) \right] \right).
\]

This shows the first claim of the lemma.

Since \(\mathcal{F}_n\) is a finite family, we are left to prove that for each \(f \in \mathcal{F}_n\),
\[
V(G_n f) = V(H_n f) + O(C^{-1}) = \sum_{i=1}^{\ell} \bar{\mu}_i \text{Cov} \left( \sum_{i=1}^{N_1} f(W_{i,1}), \sum_{i=1}^{N_2} f(W_{i,2}) \right) + O(C^{-1}),
\]
where \(2_i\) denotes the \(\ell\)-tuple vector with \(2\) in each entry but for \(1\) in the \(i\)-th entry. Note that
\[
V(H_n f) = \sum_{e \in \mathcal{E}_1} \frac{C}{\Pi e_i = 1 C_i} V \left( \mathbb{E} \left[ \sum_{i \in [N_1]} f(W_{i,1}) \right] | U_e \right).
\]  
(3.3)

To conclude, it suffices to show that for each \(e \in \mathcal{E}_1\),
\[
V \left( \mathbb{E} \left[ \sum_{i \in [N_1]} f(W_{i,1}) \right] | U_e \right) = \text{Cov} \left( \sum_{i \in [N_1]} f(W_{i,1}), \sum_{i=1}^{N_2-e} f(W_{i,2-e}) \right).
\]

As \((N_j, (W_{i,j})_{1 \leq i \leq N})_{j \geq 1} = \left\{ \tau \left( (U_j \odot e)_{e \in \mathcal{U}, \epsilon_i = 1 \epsilon_r} \right) \right\}_{j \geq 1}\) with i.i.d. \(U\)’s, we have \(\mathbb{E} \left[ \sum_{i \in [N_1]} f(W_{i,1}) \right] | U_e = \mathbb{E} \left[ \sum_{i \in [N_1]} f(W_{i,j}) \right] | U_e\) for any \(j\) such that \(j \odot e = 1 \odot e = e\). Because \(2 - e \odot e = e\), we have
\[
\forall \left( \mathbb{E} \left[ \sum_{i \in [N_1]} f(W_i, 1) \mid U_e \right] \right) = \text{Cov} \left( \mathbb{E} \left[ \sum_{i \in [N_1]} f(W_i, 1) \mid U_e \right], \mathbb{E} \left[ \sum_{i = 1}^{N_2-e} f(W_i, 2-e) \mid U_e \right] \right). \quad \text{For any } e \in \varepsilon_1, \text{ we have } 2-e \neq 1. \text{ The independence of the } U \text{'s ensures}
\]

\[
(U_1 \otimes e')_{e' \in U^f_{r=1, e \notin e} } \perp (U_{(2-e) \otimes e'})_{e' \in U^f_{r=1, e \notin e} } | U_e
\]

and thus \( \sum_{i=1}^{N_1} f(W_i, 1) \perp \sum_{i=1}^{N_2-e} f(W_i, 2-e) | U_e \).

Hence, for \( e \in \varepsilon_1 \)

\[
\mathbb{E} \left[ \text{Cov} \left( \sum_{i \in [N_1]} f_1(W_i, 1), \sum_{i = 1}^{N_2-e} f_2(W_i, 2-e) \mid U_e \right) \right] = 0.
\]

By the law of total covariance, we obtain

\[
\forall \left( \mathbb{E} \left[ \sum_{i \in [N_1]} f(W_i, 1) \mid U_e \right] \right) = \text{Cov} \left( \sum_{i \in [N_1]} f(W_i, 1), \sum_{i = 1}^{N_2-e} f(W_i, 2-e) \right). \]

This establishes the second claim of the lemma. \(\square\)
References

Ai, C. AND X. CHEN (2003): “Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions,” Econometrica, 71, 1795–1843.

——— (2007): “Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models with Different Conditioning Variables,” Journal of Econometrics, 141, 5–43.

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., V. CH 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. CH Chernozhukov, I. Fernández-Val, AND C. Hansen (2017): “Program evaluation and causal inference with high-dimensional data,” Econometrica, 85, 233–298.

BELLONI, A., V. CH 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. CH 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. CH 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., J. LEVINSOHN, AND A. PAKES (1995): “Automobile prices in market equilibrium,” Econometrica: Journal of the Econometric Society, 841–890.
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.

Chen, X., O. Linton, and I. Van Keilegom (2003): “Estimation of Semiparametric Models when the Criterion Function Is Not Smooth,” *Econometrica*, 71, 1591–1608.

Chen, X. and D. Pouzo (2015): “Sieve Wald and QLR Inferences on Semi/Nonparametric Conditional Moment Models,” *Econometrica*, 83, 1013–1079.

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 Postselection 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.

Kalenberg, O. (2006): *Probabilistic symmetries and invariance principles*, Springer Science & Business Media.
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.

Okui, R., D. S. Small, Z. Tan, and J. M. Robins (2012): “Doubly robust instrumental variable regression,” *Statistica Sinica*, 173–205.

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:1608.00033*.

Van der Vaart, A. (1998): *Asymptotic Statistics*, Cambridge University Press.
| $N$ | $M$ | $C$ | 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).
Table 2: Estimates and standard errors of the coefficient $\theta_0$ of log price in the demand model. The first column indicates the instrumental variable. The second column shows the results of the DML by lasso not accounting for clustering with the number $K = 4$ of folds for cross fitting. The third and fourth columns show the results of the 1-way cluster-robust DML by lasso clustered at product and market, respectively, with the number $K = 4$ of folds for cross fitting. The fifth column shows the results of the 2-way cluster-robust DML by lasso with the number $K^2 = 4$ of folds for two-way cross fitting. All the results are based on the average of ten rerandomized DML.

| Instrument ($Z_{ij}$) | 0-Way | 1-Way Product | 1-Way Market | 2-Way Product × Market |
|-----------------------|-------|---------------|--------------|------------------------|
| Horsepower/weight of other products | -5.763 | -5.719 | -5.815 | -5.659 |
| Miles/dollar of other products | -6.121 | -6.056 | -6.191 | -6.121 |
| Size of other products | -5.684 | -5.641 | -5.727 | -5.593 |
|                       | (0.460) | (0.640) | (1.024) | (1.211) |
|                       | (0.607) | (0.865) | (1.491) | (3.963) |
|                       | (0.413) | (0.565) | (0.892) | (1.015) |