Journal of Business & Economic Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Multiway Cluster Robust Double/Debiased Machine Learning

Harold D. Chiang, Kengo Kato, Yukun Ma & Yuya Sasaki

To cite this article: Harold D. Chiang, Kengo Kato, Yukun Ma & Yuya Sasaki (2021): Multiway Cluster Robust Double/Debiased Machine Learning, Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2021.1895815

To link to this article: https://doi.org/10.1080/07350015.2021.1895815

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2021, VOL. 00, NO. 0, 1–11
https://doi.org/10.1080/07350015.2021.1895815

Multiway Cluster Robust Double/Debiased Machine Learning*

Harold D. Chiang[a], Kengo Kato[b], Yukun Ma[c], and Yuya Sasaki[c]

[a]Department of Economics, University of Wisconsin-Madison, Madison, WI; [b]Department of Statistics and Data Science, Cornell University, Ithaca, NY; [c]Department of Economics, Vanderbilt University, Nashville, TN

ABSTRACT

This article 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 nonrobust ones in the demand model.

ARTICLE HISTORY
Received March 2020
Accepted February 2021

KEYWORDS
Double/debiased machine learning, Multiway clustering, Multiway cross-fitting

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 the 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 markets 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 iid 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 article by Chernozhukov et al. (CCDDHNR, 2018a). They provided a general DML toolbox for estimation and inference for structural parameters with high-dimensional and/or infinite-dimensional nuisance parameters. In that article, the estimation method and properties of the estimator are presented under the typical microeconometric assumption of iid 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 (2018a). To our best knowledge, the present article 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) conducted formal analyses of bootstrap validity under multiway cluster sampling robustly accounting for non-degenerate and degenerate cases. Chen, Linton, and Van Keilegom (2018) developed 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 articles. To our knowledge, the present article 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, for example, 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 (2018a) 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.

CONTACT Harold D. Chiang *Chiang: hdchiang@wisc.edu* Department of Economics, University of Wisconsin-Madison, William H. Sewell Social Science Building 1180 Observatory Drive Madison, WI 53706-1393.

*First arXiv date: September 8, 2019.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/UBES*.

This framework is further generalized in different directions by Belloni et al. (2017) and Belloni et al. (2018). CCDDHNR (2018a) combined 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 iid 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, for example, 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, Gelbach, and Miller (2011) first provided a multiway cluster robust variance estimator for linear regression models without imposing additional parametric assumptions on the intra-cluster correlation structure. This variance estimator has significantly reshaped the landscape of econometric practices in applied microeconomics in the past decade.[1] In contrast to the popularity among empirical researchers, theoretical justification of the validity of this type of procedures was lagging behind. The first rigorous treatment of asymptotic properties of multiway cluster robust estimators are established by Menzel (2017) using the Aldous-Hoover representation under the assumption of separable exchangeability. The asymptotic theory of Menzel (2017) covers both non-degenerate and degenerate cases. Focusing on non-degenerate situations, Chen, Linton, and Van Keilegom (2018) further extended this approach to a general empirical process theory.[2] Using this asymptotic framework, MacKinnon, Nielsen, and Webb (2019) studied linear regression models 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 with Gaussian limiting distributions.

Despite of the popularity of both machine learning and cluster robust inference among empirical researchers, relatively limited cluster robust inference results exist for machine learning based methods. Inference for machine learning based methods with one-way clustering is studied by Belloni et al. (2016), Kock (2016), Kock and Tang (2019), Semenova et al. (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) investigated the performance of lasso and post-lasso in the partially linear model setting of Belloni, Chernozhukov, and Hansen (2014b) under multiway cluster sampling. To our best knowledge, there is no general machine learning based procedures with known validity under multiway cluster sampling environments.

[1] As of July 31, 2020, Cameron, Gelbach, and Miller (2011) has received over 2,700 citations. The majority of these citations came from applied economic articles.

[2] See also Davezies, D'Haultfoeuille, and Guyonvarch (2019) for further generalization of the empirical process theory for dyadic data under joint exchangeability assumption.

2. Setup

We begin with an introduction of the data structure of multiway clustering (Section 2.1) and the econometric structure of Neyman orthogonal score (Section 2.2).

2.1. Multiway Clustering

Suppose that the researcher observes a sample $\{W_{ij} | i \in \{1,\dots,N\}, j \in \{1,\dots,M\}\}$ of double-indexed observations of size NM. It is two-way clustered if units in the cluster $\{W_{ij}\}_{j=1}^{M}$ are dependent for any given $i \in \{1,\dots,N\}$ and units in the cluster $\{W_{ij}\}_{i=1}^{N}$ are dependent for any given $j \in \{1,\dots,M\}$.

To fix ideas, consider the case of market share data that consist of N markets and M products. For any given market i, the M observations $\{W_{ij}\}_{j=1}^{M}$ across different products are subject to a common demand shock in the ith market, and this common shock can induce dependence within the ith market cluster $\{W_{ij}\}_{j=1}^{M}$. Likewise, for any given product j, the N observations $\{W_{ij}\}_{i=1}^{N}$ across different markets are subject to a common supply shock from the producer of product j, and this common shock can induce dependence within the jth product cluster $\{W_{ij}\}_{i=1}^{N}$. In this way, the fundamentals of economics, namely supply and demand, may well cause two ways of dependence or two-way clustering in market share data. In addition to the market share data, similar two-way dependence structures are shared by network data, matched employer-employee data, matched student-teacher data, and scanner data among others.

Under the assumption of two-way clustering (a formal statement of which is postponed until Assumption 1 in Section 4), the random vector W_{ij} can be represented by the structure

$$W_{ij} = \tau(U_{i0}, U_{0j}, U_{ij})$$

for some function τ (unknown to econometricians), where $\{U_{ij} \mid (i,j) \in \mathbb{N}^2 \setminus \{(0,0)\}\}$ are independent uniform random variables (and unobservable by econometricians). This representation is called the Aldous-Hoover representation, and serves as a key mathematical device to develop asymptotic statistical theories under two-way clustering. Intuitively, U_{i0} can be considered as the ith market fixed effect representing the demand shock in this market, U_{0j} can be considered as the jth product fixed effect representing the supply shock by the produce of this product, and U_{ij} is the idiosyncratic shock. This representation also motivates the two-way fixed effect model

$$Y_{ij} = X_{ij}'\beta + \overbrace{\tau(\underbrace{\alpha_i}_{U_{i0}}, \underbrace{\gamma_j}_{U_{0j}}, U_{ij})}^{W_{ij}=\tau(U_{i0},U_{0j},U_{ij})} \quad i \in \{1,\dots,N\}, j \in \{1,\dots,M\}.$$

Without assuming additivity among α_i, γ_j and U_{ij}, this panel structure can be analyzed by allowing for the possibility of two-way cluster dependence in the sole error term W_{ij} in $Y_{ij} = X_{ij}'\beta + W_{ij}$.

Suppose that the researcher is interested in the mean $\mathrm{E}[\psi(W_{ij})]$ for some scalar-valued function ψ. If there were no

dependence in the data $\{W_{ij} | i \in \{1,\dots,N\}, j \in \{1,\dots,M\}\}$, then the standard econometric theory implies that the standard error of the sample mean $\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\psi(W_{ij})$ is $SE_0 = \sqrt{\mathbb{V}_{NM}(\psi(W_{ij}))/(NM)}$. On the other hand, if the researcher suspects the two-way dependence as described above, then he/she would instead like to employ the two-way-cluster-robust standard error

$$SE_2 = \sqrt{(\bar{\mu}_1\widehat{\Gamma}_1 + \bar{\mu}_2\widehat{\Gamma}_2)/\underline{C}}, \tag{2.1}$$

where $\underline{C} = \min\{N,M\}$, $\bar{\mu}_1 = \underline{C}/N$, $\bar{\mu}_2 = \underline{C}/M$, $\widehat{\Gamma}_1 = \frac{1}{NM^2}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{j'=1}^{M}\psi(W_{ij})\psi(W_{ij'})$ and $\widehat{\Gamma}_2 = \frac{1}{N^2M}\sum_{i=1}^{N}\sum_{i'=1}^{N}\sum_{j=1}^{M}\psi(W_{ij})\psi(W_{i'j})$. This idea of two-way cluster-robust standard errors dates back to the seminal article by Cameron, Gelbach, and Miller (2011),[3] and has been widely used in empirical economic researches as mentioned in the introduction.

2.2. Neyman Orthogonal Score

2.2.1. Motivation of Neyman Orthogonal Score

We next introduce the concept of Neyman orthogonal score as an important component of the DML. To fix ideas to this end, consider as a concrete example the partially linear IV model (cf. Okui, Small, Tan and Robins, 2012 ; CCDDHNR, 2018a, Section 4.2) translated into double-indexed data

$$Y_{ij} = D_{ij}\theta_0 + g_{10}(X_{ij}) + \epsilon_{ij}, \qquad \mathrm{E}[\epsilon_{ij}|X_{ij}, Z_{ij}] = 0, \tag{2.2}$$

$$Z_{ij} = m_0(X_{ij}) + v_{ij}, \qquad \mathrm{E}[v_{ij}|X_{ij}] = 0. \tag{2.3}$$

The researcher observes the random vector $W_{ij} = (Y_{ij}, D_{ij}, X_{ij}', Z_{ij})'$, whose four components are interpreted as the outcome, endogenous regressor, exogenous regressors, and instrumental variable, respectively. The parameter vector θ_0 is an object of interest. The functions, g_{10} and m_0, are unknown and need to be estimated by a machine learner, such as a kernel smoother, series estimator, lasso, ridge, elastic nets, and neural networks among others. However, the naïve idea of plugging machine-learned $\widehat{g}_1$ and $\widehat{m}$ in the standard IV regression estimation framework would not work well, because $\widehat{g}_1$ and $\widehat{m}$ typically converge to g_{10} and m_0, respectively, at rates that are slower than the desired rate for an estimator of θ_0, which is of order $1/\sqrt{\underline{C}}$ if observations are two-way cluster dependent – see the expression in Equation (2.1) in Section 2.1.

A Neyman orthogonal score is a useful device to mitigate these effects of slow convergence rates of machine learners. To see this, consider the function ψ defined by

$$\psi(w;\theta,\eta) = (y - (d - g_2(x))\theta - g_1(x))(z - m(x)), \tag{2.4}$$

where $w = (y,d,x,z)$ and $\eta = (g_1,g_2,m)$. If we set $\eta_0 = (g_{10},g_{20},m_0)$ where $g_{10}(X) = \mathrm{E}[Y|X]$, $g_{20}(X) = \mathrm{E}[D|X]$ and $m_0(X) = \mathrm{E}[Z|X]$, then it holds under the model (2.2) and (2.3) that

$$\mathrm{E}[\psi(W_{ij};\theta_0,\eta_0)] = 0.$$

Furthermore, if the projections take the linear forms, $g_{10}(x) = p(x)'\beta_0$, $g_{20}(x) = p(x)'\gamma_0$ and $m_0(x) = p(x)'\xi_0$ with possibly high-dimensional basis $p(x)$, for simplicity, then it also holds under the model (2.2) and (2.3) that

$$\frac{\partial}{\partial\beta}\mathrm{E}[\psi(W_{ij};\theta_0,\eta)]\Big|_{\eta=\eta_0} = 0,$$

$$\frac{\partial}{\partial\gamma}\mathrm{E}[\psi(W_{ij};\theta_0,\eta)]\Big|_{\eta=\eta_0} = 0,$$

$$\frac{\partial}{\partial\xi}\mathrm{E}[\psi(W_{ij};\theta_0,\eta)]\Big|_{\eta=\eta_0} = 0.$$

These three 'orthogonality' equations imply that the moment condition $\mathrm{E}[\psi(W_{ij};\theta_0,\eta_0)] = 0$ is insensitive to local perturbations of η in a neighborhood of η_0. Because of this insensitivity, even slowly converging errors of the machine learner $\widehat{\eta} = (\widehat{g}_1,\widehat{g}_2,\widehat{m})$, such as the lasso estimator of (β_0,γ_0,ξ_0), will exercise only ignorable effects on $\mathrm{E}[\psi(W_{ij};\theta_0,\widehat{\eta})]$, thus allowing for estimation of θ_0 with the desired rate $1/\sqrt{\underline{C}}$ of convergence. With these properties, this function ψ is said to be a Neyman orthogonal score function for the model (2.2) and (2.3) – see Okui et al. (2012) and CCDDHNR (2018a).

2.2.2. The Definition of Neyman Orthogonal Score

While the above discussion is specific to the example of model (2.2) and (2.3) and is informal, we now present a formal and general definition of Neyman orthogonal score according to CCDDHNR (2018a). The structural model is assumed to entail the moment restriction

$$\mathrm{E}[\psi(W_{11};\theta_0,\eta_0)] = 0 \tag{2.5}$$

for some score ψ that depends on a low-dimensional parameter vector $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ and a nuisance parameter $\eta \in T$ for a convex subset T of a normed linear space.

Let $\widetilde{T} = \{\eta - \eta_0 : \eta \in T\}$, and define the pathwise derivative map $D_r : \widetilde{T} \to \mathbb{R}^{d_\theta}$ by $D_r[\eta - \eta_0] := \partial_r\Big\{\mathrm{E}[\psi(W_{11};\theta_0,\eta_0 + r(\eta - \eta_0))]\Big\}$ for all $r \in [0,1)$. Also denote its limit by $\partial_\eta\mathrm{E}\psi(W_{11};\theta_0,\eta_0)[\eta - \eta_0] := D_0[\eta - \eta_0]$. We say that the Neyman orthogonality condition holds at (θ_0,η_0) with respect to a nuisance realization set $\mathcal{T}_n \subset T$ if the score ψ satisfies Equation (2.5), the pathwise derivative $D_r[\eta - \eta_0]$ exists for all $r \in [0,1)$ and $\eta \in \mathcal{T}_n$, and the orthogonality equation

$$\partial_\eta\mathrm{E}\psi(W_{11};\theta_0,\eta_0)[\eta - \eta_0] = 0 \tag{2.6}$$

holds for all $\eta \in \mathcal{T}_n$.

As an extended definition, we also say that the λ_n Neyman near-orthogonality condition holds at (θ_0,η_0) with respect to a nuisance realization set $\mathcal{T}_n \subset T$ if the score ψ satisfies Equation (2.5), the pathwise derivative $D_r[\eta - \eta_0]$ exists for all $r \in [0,1)$ and $\eta \in \mathcal{T}_n$, and the orthogonality equation

$$\sup_{\eta\in\mathcal{T}_n}\left|\left|\partial_\eta\mathrm{E}\psi(W_{11};\theta_0,\eta_0)[\eta - \eta_0]\right|\right| \le \lambda_n \tag{2.7}$$

holds for all $\eta \in \mathcal{T}_n$ for some positive sequence $\{\lambda_n\}_n$ such that $\lambda_n = o(\min\{N,M\}^{-1/2})$. This definition relaxes the requirement of Neyman orthogonality, and allows for the derivative to be nonzero.

[3] While the formula that we display here consists of two terms, $\bar{\mu}_1\Gamma_1$ and $\bar{\mu}_2\Gamma_2$, the formula suggested by Cameron, Gelbach, and Miller (2011) consists of one additional term to deduct double-counted terms by $\bar{\mu}_1\widehat{\Gamma}_1$ and $\bar{\mu}_2\widehat{\Gamma}_2$. Since this additional term (or the effect of double counting) is asymptotically negligible, we omit it here for the sake of succinctness.

2.2.3. Construction of Neyman Orthogonal Score

Although orthogonal scores are readily available for certain models, not all scores are orthogonal. The recent literature provides recipes to construct an orthogonal score from a possibly nonorthogonal score. We refer interested readers to Chernozhukov et al. (2018b) as well as CCDDHNR (2018a). Here, for the convenience of readers, we provide a recipe according to CCDDHNR (2018a, Section 2.2).

Let $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ and $\beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta}$, where $\mathcal{B}$ is convex, be the target and nuisance parameters. Although β is finite-dimensional, its dimension can be high. Suppose that the true parameter solves the optimization problem

$$\sup_{\theta\in\Theta,\beta\in\mathcal{B}} \mathrm{E}[\ell(W_{11};\theta,\beta)]$$

for a known criterion function ℓ. In this setting, with the nuisance parameter redefined by

$$\eta = (\beta', \mathrm{vec}(J_{\theta\beta}J_{\beta\beta}^{-1})')' \in T = \mathcal{B} \times \mathbb{R}^{d_\theta d_\beta},$$

the function ψ defined by

$$\psi(w;\theta,\eta) = \partial_\theta \ell(w;\theta,\beta) - J_{\theta\beta}J_{\beta\beta}^{-1}\frac{\partial}{\partial\beta}\ell(w;\theta,\beta)$$

is a Neyman orthogonal score, where $J_{\theta\beta}$ and $J_{\beta\beta}$ are given by

$$\begin{pmatrix} J_{\theta\theta} & J_{\theta\beta} \\ J_{\beta\theta} & J_{\beta\beta} \end{pmatrix} = \frac{\partial}{\partial(\theta'\beta')}\mathrm{E}\left[\frac{\partial}{\partial(\theta'\beta')'}\ell(W_{11};\theta,\beta)\right]\Bigg|_{\theta=\theta_0,\beta=\beta_0}.$$

2.3. Example: Demand Analysis

Section 2.1 introduces multiway clustering and Section 2.2 introduces Neyman orthogonal score, with emphases on market share data and the partially linear IV model. In this section, we present a framework of demand analysis as a concrete example that highlights both of these two points together.

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

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

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

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

In this framework, Lu, Shi, and Tao (2019, Equation (9)) derived the partially linear equation

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

for estimation of θ_0, where $Y_{ij} = \log(S_{ij}) - \log(S_{0j})$ denotes the observed log share of product i relative to the log of the outside share. Since D_{ij} usually consists of the log of the endogenous price of product i in market j, researchers often use instruments Z_{ij} such that $E[\epsilon_{ij}|X_{ij}, Z_{ij}] = 0$. This yields the reduced-form Equation (2.2), together with the innocuous nonparametric projection equation (2.3). The Neyman orthogonal score (2.4) is readily available for estimation of θ_0 in this demand model.

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 clustering as discussed in Section 2.1. Specifically, for each product i, the cluster $\{W_{ij}\}_{j=1}^M$ is likely dependent through a supply shock by the producer of product i. Similarly, for each market j, the cluster $\{W_{ij}\}_{i=1}^N$ is likely dependent through a demand shock in market j. □

As discussed in Section 2.1, assuming two-way clustering for the error term ϵ_{ij} effectively allows for two-way fixed effects in the structure. In other words, the error term can be understood to be a possibly nonlinear and nonadditive function $\epsilon_{ij} = \tau_\epsilon(\alpha_i, \gamma_j, U_{ij})$ of product fixed effect α_i, market fixed effect γ_j and idiosyncratic shock U_{ij}. This structure generalizes the common additive two-way fixed effect panel models.[4]

3. Procedure

Now that the setup has been introduced in Section 2, we proceed with presenting the DML procedure. We first review the DML based on the conventional cross-fitting that works under iid sampling. After illustrating limitations of the conventional cross-fitting under multiway clustering, we then move on to presenting our proposed multiway cross-fitting algorithm and the multiway cluster robust DML based on this algorithm.

3.1. Conventional Cross-Fitting

3.1.1. The Role of Cross-Fitting under Random Sampling

For convenience of reviewing the conventional cross-fitting, consider an iid sample $\{W_i\}_{i=1}^n$ just within the current section. With the Neyman orthogonal score (2.4) where x is assumed to be a scalar for simplicity of writing, we can write the asymptotic linear representation for $\sqrt{n}(\widehat{\theta} - \theta_0)$ as

$$\sqrt{n}(\widehat{\theta} - \theta_0) = A^* + B^* + C^* + o_p(1), \tag{3.1}$$

where A^* is asymptotically normal, B^* consists of terms like $\frac{B}{\sqrt{n}}\sum_{i=1}^n(\widehat{g}_1(X_i) - g_{10}(X_i))(\widehat{m}(X_i) - m_0(X_i))$, and C^* consists of terms like $\frac{C}{\sqrt{n}}\sum_{i=1}^n(\widehat{g}_1(X_i) - g_{10}(X_i))v_i$.[5] See CCDDHNR (2018a, pages C4-C5) for an analogous discussion in the case of Robinson's (1988) model.

The typical term $\frac{B}{\sqrt{n}}\sum_{i=1}^n(\widehat{g}_1(X_i) - g_{10}(X_i))(\widehat{m}(X_i) - m_0(X_i))$ in B^* consists of products of machine learning errors, and is asymptotically negligible if each of these machine learning errors $\widehat{g}_1(X_i) - g_{10}(X_i)$ and $\widehat{m}(X_i) - m_0(X_i)$ vanishes at a rate that is faster than the order of $n^{-1/4}$ (yet possibly slower than the order of $n^{-1/2}$). This property owes to the Neyman orthogonality of ψ. The typical term $\frac{C}{\sqrt{n}}\sum_{i=1}^n(\widehat{g}_1(X_i) - g_{10}(X_i))v_i$ in C^*, on the other hand, is asymptotically negligible by the law of iterated expectations and Markov inequality, if the sample that is used to obtain $\widehat{g}_1$ is different from and independent of the

[4] The term "fixed effect" also usually entails the endogeneity, which concerns the identification problem. This additional feature is also accommodated if the instrument Z_{ij} is exogenous to ϵ_{ij}.

[5] See Appendix F in the supplementary appendix for a full influence function representation with concrete expressions.

sample that is used to evaluate the sum. It is this argument about the C^* term that leads us to the idea of cross-fitting under an iid sampling.

Randomly split $\{1,\dots,n\}$ into two folds, I_1 and I_2. Let $\widehat{g}_{11}$ (respectively, $\widehat{g}_{12}$) denote a machine learner for g_{10} using the subsample $\{W_i\}_{i\in I_2}$ (respectively, $\{W_i\}_{i\in I_1}$). Since the sample is iid, $\{W_i\}_{i\in I_1}$ and $\{W_i\}_{i\in I_2}$ are independent subsamples. Therefore, $\widehat{g}_{11}$ is independent of $\{W_i\}_{i\in I_1}$, and thus $\frac{\sqrt{n}}{n/2}\sum_{i\in I_1}(\widehat{g}_{11}(X_i)-g_{10}(X_i))v_i$ can be made asymptotically negligible by the aforementioned argument. Similarly, $\frac{\sqrt{n}}{n/2}\sum_{i\in I_2}(\widehat{g}_{12}(X_i)-g_{10}(X_i))v_i$ can be made asymptotically negligible. Each of these two sums uses only a half of the whole sample of size n. To improve the efficiency, we want to use the whole sample by averaging the estimates associated with the two folds. The average

$$\frac{1}{2}\left[\frac{\sqrt{n}}{n/2}\sum_{i\in I_1}(\widehat{g}_{11}(X_i)-g_{10}(X_i))v_i + \frac{\sqrt{n}}{n/2}\sum_{i\in I_2}(\widehat{g}_{12}(X_i)-g_{10}(X_i))v_i\right]$$

is the split-half counterpart of the typical term in C^*. The asymptotic negligibility of each of the two subsample means imply that of this whole sample mean.

Extending this heuristic idea, the general cross-fitting under an iid sampling proceeds as follows. With a fixed positive integer $K>1$, randomly partition $\{1,\dots,n\}$ into K folds $\{I_1,\dots,I_K\}$. Note that, for each $k\in\{1,\dots,K\}$, observations in I_k are independent of its complement $I_k^c=\{1,\dots,N\}\setminus I_k$ and vice versa by the iid sampling assumption. For each $k\in\{1,\dots,K\}$, obtain an estimate

$$\widehat{\eta}_k=\widehat{\eta}\left((W_i)_{i\in I_k^c}\right)$$

of the nuisance parameter η by some machine learning method using only the subsample of those observations with $i\in I_k^c$. In turn, we define $\widetilde{\theta}$ as the solution to

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|I_k|}\sum_{i\in I_k}\psi(W_i;\widetilde{\theta},\widehat{\eta}_k)=0, \tag{3.2}$$

where $|I_k|$ denotes the cardinality of the set I_k. Here, it is crucial that $\widehat{\eta}_k$ is independent of I_k in each of the K inner sums to vanish terms in C^*.

With this cross-fitting operation, the solution $\widetilde{\theta}$ allows for the asymptotic linear representation (3.1) with the B^* and C^* terms asymptotically negligible. Thus, this DML estimator $\widetilde{\theta}$ (CCDDHNR, 2018a) enjoys the root-n asymptotic normality solely based on the A^* term under the iid sampling.

3.1.2. Limitation of the Conventional Cross-Fitting under Multiway Clustering

Let us now turn back to the case of using the two-way cluster dependent data as in Section 2. In this case, the random splitting of the sample $\{W_{ij}\,|\,i\in\{1,\dots,N\},j\in\{1,\dots,M\}\}$ of size NM into K folds will not ensure the independence between observation in I_k from I_k^c for any $k\in\{1,\dots,K\}$.

An alternative idea might be to randomly split the clusters in one of the cluster dimensions, say the N market clusters $\{\{W_{1j}\}_{j=1}^M,...,\{W_{Nj}\}_{j=1}^M\}$, into K folds $\{I_1,\dots,I_K\}$. However, this will not ensure the independence between observation in I_k from I_k^c either. To see this, note that I_k contains observations from the set of products $j\in\{1,\dots,M\}$ as well as I_k^c contains observations from the same set of products $j\in\{1,\dots,M\}$. A common supply shock induced by producers of these products may well cause dependence between observation in I_k and those in I_k^c.

3.2. The Multiway Cross-Fitting and the Multiway Cluster Robust DML

In light of the non-applicability of the conventional cross-fitting under multiway clustering pointed out in Section 3.1.2, we now propose a novel multiway cross-fitting algorithm and the multiway cluster robust DML based on it. For any $r\in\mathbb{N}$, we use the notation $[r]=\{1,\dots,r\}$. With a fixed positive integer $K>1$, randomly partition $[N]$ into K parts $\{I_1,\dots,I_K\}$ and $[M]$ into K parts $\{J_1,\dots,J_K\}$. For each $(k,\ell)\in[K]^2$, obtain an estimate

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

of the nuisance parameter η by some machine learning method 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 $\widetilde{\theta}$, the multiway double/debiased machine learning (multiway DML) estimator for θ_0, as the solution to

$$\frac{1}{K^2}\sum_{(k,\ell)\in[K]^2}\mathbb{E}_{n,k\ell}[\psi(W;\widetilde{\theta},\widehat{\eta}_{k\ell})]=0, \tag{3.3}$$

where, with $|I_k|$ and $|J_\ell|$ denoting the cardinalities of I_k and J_ℓ, respectively, $\mathbb{E}_{n,k\ell}[f(W)]=\frac{1}{|I_k||J_\ell|}\sum_{(i,j)\in I_k\times J_\ell}f(W_{ij})$ denotes the subsample empirical expectation using only 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 $\widehat{\eta}_{k\ell}$ is computed using the subsample of those observations with multiway indices $(i,j)\in([N]\setminus I_k)\times([M]\setminus J_\ell)$, and in turn the score term $\mathbb{E}_{n,k\ell}[\psi(W;\cdot,\widehat{\eta}_{k\ell})]$ 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 $N=M=40$ (with each side of a small square block representing 10 clusters), where the cross-fitting repeats for $K^2(=2^2=4)$ times.

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

Remark 2 (Higher Cluster Dimensions). When we have α-way clustering for an integer $\alpha>2$, the above algorithm

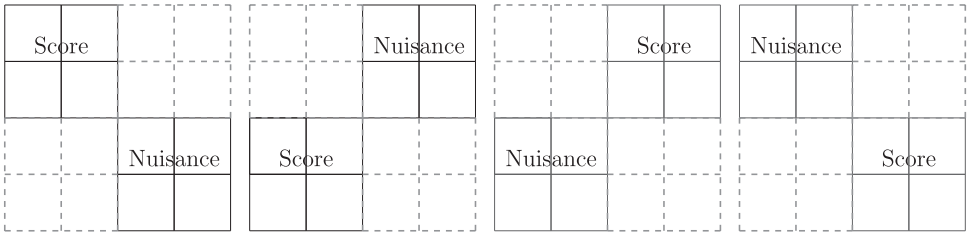

Figure 1. An illustration of 2^2-fold cross-fitting.

can be easily generalized into a K^α-fold multiway DML estimator. See Appendix C in the supplementary appendix for a generalization. □

Recalling the notation $\underline{C} = N \wedge M$, we propose to estimate the asymptotic variance of $\sqrt{\underline{C}}(\widetilde{\theta} - \theta_0)$ by

$$\widehat{\sigma}^2 = \widehat{J}^{-1}\widehat{\Gamma}(\widehat{J}^{-1})', \tag{3.4}$$

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

$$\widehat{\Gamma} = \frac{1}{K^2}\sum_{(k,\ell)\in[K]^2}\left\{\frac{|I|\wedge|J|}{(|I||J|)^2}\sum_{i\in I_k}\sum_{j,j'\in J_\ell}\psi(W_{ij};\widetilde{\theta},\widehat{\eta}_{k\ell})\psi(W_{ij'};\widetilde{\theta},\widehat{\eta}_{k\ell})' + \frac{|I|\wedge|J|}{(|I||J|)^2}\sum_{i,i'\in I_k}\sum_{j\in J_\ell}\psi(W_{ij};\widetilde{\theta},\widehat{\eta}_{k\ell})\psi(W_{i'j};\widetilde{\theta},\widehat{\eta}_{k\ell})'\right\} \text{ and} \tag{3.5}$$

$$\widehat{J} = \frac{1}{K^2}\sum_{(k,\ell)\in[K]^2}\mathbb{E}_{n,k\ell}\left[\frac{\partial}{\partial\theta}\psi(W;\theta,\widehat{\eta}_{k\ell})\Big|_{\theta=\widetilde{\theta}}\right], \tag{3.6}$$

accounting for multiway cluster dependence, similarly to the existing multiway cluster robust formula (2.1). For a d_θ-dimensional vector r, the $(1-a)$ confidence interval for the linear functional $r'\theta_0$ can be constructed by

$$\mathrm{CI}_a := [r'\widetilde{\theta} \pm \Phi^{-1}(1-a/2)\sqrt{r'\widehat{\sigma}^2 r/\underline{C}}],$$

where Φ denotes the standard normal CDF. Summarizing all the above procedures, we present the following step-by-step guidance for implementation of the multiway cluster robust DML.

Algorithm 1 (K^2-fold Multiway Cluster Robust DML). Let $K = 2$.

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 $(k,\ell) \in [K]^2$: obtain an estimate $\widehat{\eta}_{k\ell} = \widehat{\eta}\left((W_{ij})_{(i,j)\in([N]\setminus I_k)\times([M]\setminus J_\ell)}\right)$ of the nuisance parameter η by some machine learning method using only the subsample of those observations with multiway indices (i,j) in $([N]\setminus I_k)\times([M]\setminus J_\ell)$.
3. Solve the equation $\frac{1}{K^2}\sum_{(k,\ell)\in[K]^2}\mathbb{E}_{n,k\ell}[\psi(W;\widetilde{\theta},\widehat{\eta}_{k\ell})] = 0$ for θ to obtain the multiway DML estimate $\widetilde{\theta}$.
4. Let the multiway DML asymptotic variance estimator be given by $\widehat{\sigma}^2 = \widehat{J}^{-1}\widehat{\Gamma}(\widehat{J}^{-1})'$ where $\widehat{\Gamma}$ and $\widehat{J}$ are give in (3.5) and (3.6), respectively.
5. Report the estimate $\widetilde{\theta}$, its standard error $\sqrt{\widehat{\sigma}^2/\underline{C}}$, and/or the $(1-a)$ confidence interval

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

For the partially linear IV model introduced in Section 2, Algorithm 1 with more details about the procedure in each step is as follows.

Algorithm 2 (K^2-fold Multiway Cluster Robust DML for Partially Linear IV Model with Lasso). Let $K = 2$.

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 $\widehat{g}_{1,k\ell}(x) = x'\widehat{\beta}_{k\ell}$ using observations from $I_k^c \times J_\ell^c$.
   b. Run a lasso of D on X to obtain $\widehat{g}_{2,k\ell}(x) = x'\widehat{\gamma}_{k\ell}$ using observations from $I_k^c \times J_\ell^c$.
   c. Run a lasso of Z on X to obtain $\widehat{m}_{k\ell}(x) = x'\widehat{\xi}_{k\ell}$ using observations from $I_k^c \times J_\ell^c$.
3. Solve the equation

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

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

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

where

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

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

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

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

4. 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 index by a single index $n \in \mathbb{N}$ as $(N, M) = (N(n), M(n))$ where $M(n)$ and $N(n)$ are nondecreasing 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 $\{\mathcal{P}_n\}_n$ be a sequence of sets of probability laws of $\{W_{ij}\}_{ij}$ – note that we allow for increasing dimensionality of W_{ij} in the sample size n. Let $P = P_n \in \mathcal{P}_n$ denote the law with respect to sample size (N, M). Throughout, we assume that this random vector W_{ij} is Borel measurable. Recalling the notation $\underline{C} = N \wedge M$, define $\mu_N = \underline{C}/N$ and $\mu_M = \underline{C}/M$, and suppose that $\mu_N \to \bar{\mu}_N$, $\mu_M \to \bar{\mu}_M$. We write $a \lesssim b$ to mean $a \leq cb$ for some $c > 0$ that does not depend on n. We also write $a \lesssim_P b$ to mean $a = O_P(b)$. For any finite dimensional vector v, $||v||$ denotes the ℓ_2 or Euclidean norm of v. For any matrix A, $||A||$ denotes the induced ℓ_2-norm of the matrix. For any set B, $|B|$ denotes the cardinality of the set.

We state the following assumption on multiway clustered sampling.

Assumption 1 (Sampling). Suppose $\underline{C} \to \infty$. 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 π_1 and π_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 $(W_{ij})_{i\in[N],j\in[M]}$.

The separate exchangeability in part (i) is similar to the identical distribution assumption. It means that the N markets $i \in \{1, \ldots, N\}$ and the M products $j \in \{1, \ldots, M\}$ have no identities, and are supposed to have *ex ante* identical distributions. Therefore, shuffling the labels of the markets and labels of the products *separately* does not affect the joint distribution of data[6]. While we maintain this assumption that is similar to the identical distribution assumption, we do relax independence assumption. The dissociation in part (ii) implies that observations are independent if they do not share the same market or the same products. On the other hand, if two observations share either the same market or the same product, then they are allowed to be arbitrarily dependent, as discussed in Section 2.1.

[6]For example, this imposes $(W_{11}, W_{12}) \overset{d}{=} (W_{21}, W_{22})$. But it does not impose that (W_{11}, W_{12}) and (W_{11}, W_{22}) to be identically distributed.

Consider linear Neyman orthogonal scores ψ of the form

$$\psi(w; \theta, \eta) = \psi^a(w; \eta)\theta + \psi^b(w; \eta), \text{ for all } w \in \text{supp}(\mathrm{W}),\ \theta \in \Theta,\ \eta \in \mathrm{T}. \tag{4.1}$$

A generalization to nonlinear score follows from linearization with Gateaux differentiability as in Section 3.3 of CCDDHNR (2018a). We focus on linear scores as they cover a wide range of applications.

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

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

i. The true parameter value θ_0 satisfies (2.5).

ii. ψ is linear in the sense that it satisfies (4.1).

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

iv. ψ satisfies either the Neyman orthogonality condition (2.6) or more generally the Neyman λ_n near orthogonality condition at (θ_0, η_0) with respect to a nuisance realization set $\mathcal{T}_n \subset T$ as

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

v. The identification condition holds as the singular values of the matrix $J_0 := \mathrm{E}_P[\psi^a(W_{00}; \eta_0)]$ are between c_0 and c_1.

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

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

ii. The following moment conditions hold:

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

iii. The following conditions on the rates r_n, r_n' and λ_n' hold:

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

iv. All eigenvalues of the matrix

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

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 (see Menzel 2017; Chen, Linton, and Van Keilegom 2018; Chiang and Sasaki 2019). Menzel (2017) did not invoke the dissociation, and follows an alternative approach to inference. Our theory could also proceed without this assumption and instead with conditional inference similarly to Menzel (2017). The other articles assume both the separate exchangeability and dissociation, and conduct unconditional inference as in this article. 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 (2018a). 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 (2018a). It imposes some high level conditions on the quality of the nuisance parameter estimator Part (iv) ensures a non-degenerate limit distribution under the rate of $\sqrt{\underline{C}}$. □

Remark 4 (*Partial Distributions*). Since the separate exchangeability in Assumption 1 (i) implies that the marginal distributions of $\{W_{ij}\}_{(i,j)\in\mathbb{N}^2}$ are identical, the stated conditions in Assumptions 2 and 3 (i)–(iii) based on W_{00} apply to those based on W_{ij} for all i and j as long as W_{ij} is independent of $\mathcal{T}_n$, which depends on a part of data through $\widehat{\eta}((W_{ij})_{(i,j)\in I^c\times J^c})$. We state Assumptions 2 and 3 (i)–(iii) based on the independent copy W_{00} rather than arbitrary W_{ij} to avoid additional explanation of the potential dependence relationship between W_{ij} and $\mathcal{T}_n$. Similarly, Assumption 3 (iv) is stated based on W_{11}, W_{12} and W_{21} due to the dependence structure implied by Assumption 1(ii). This differs from CCDDHNR (2018a). W_{11}, W_{12} and W_{21} can be replaced by W_{ij}, $W_{ij'}$ and $W_{i'j}$, respectively, for any $i \neq i'$, $j \neq j'$ because of the identical distribution implied by separate exchangeability. □

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

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

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

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

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

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

$$\sigma^2 := J_0^{-1}\Gamma(J_0^{-1})'. \tag{4.2}$$

A proof of this theorem is provided in Appendix A.1 in the supplementary appendix.

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

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

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

A proof of this theorem is provided in Appendix A.2 in the supplementary appendix.

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

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

$$\mathrm{CI}_a := [r'\widetilde{\theta} \pm \Phi^{-1}(1-a/2)\sqrt{r'\widehat{\sigma}^2 r/\underline{C}}]$$

satisfies

$$\sup_{P\in\mathcal{P}_n} |P_P(r'\theta_0 \in \mathrm{CI}_a) - (1-a)| \to 0.$$

As in Section 3.4 of CCDDHNR (2018a), 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.

5. Simulation Studies

5.1. Simulation Setup

Consider the partially linear IV model introduced in Section 2. We specifically focus on the following high-dimensional linear representations

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

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

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

Table 1. Simulation results based on 5,000 Monte Carlo iterations.

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

NOTES: 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).

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

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

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

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

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

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

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

5.2. Results

Monte Carlo simulations are conducted with 2500 iterations for each set. Table 1 reports simulation results. The first four columns in the table indicate the data generating process (N, M, $\underline{C}$, and dim(X)). The next column indicates the integer K for our K^2-fold cross-fitting method. We use $K = 2$ and 3 in the simulations for the displayed results, since $2^2 (\approx 5)$ and $3^2 (\approx 10)$ are close to the common numbers of folds used in cross-fitting in practice. The next column indicates the machine learning method for estimation of $\widehat{\eta}_{k\ell}$. We use the ridge, elastic net, and lasso. The last four columns of the table report Monte Carlo simulation statistics, including the bias (Bias), standard deviation (SD), root mean square error (RMSE), and coverage frequency for the nominal probability of 95% (Cover).

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

Comparing the results across the three machine learning methods, we observe that the ridge entails larger bias and smaller variance relative to the elastic net and lasso in finite sample. This makes the coverage frequency of the ridge less accurate compared with the elastic net and lasso. 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 (2018a) for iid sampling—see their Remark 3.1 where they recommended 4- or 5-fold cross-fitting.

6. Empirical Illustration: Demand Analysis with Market Share Data

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

$$Y_{ij} = D_{ij}\theta_0 + g_0(X_{ij}) + \epsilon_{ij} \tag{6.1}$$

for estimation of θ_0, where $Y_{ij} = \log(S_{ij}) - \log(S_{0j})$ denotes the observed log share of product i relative to the log of the outside share 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 $\mathrm{E}_P[\epsilon_{ij}|X_{ij}, Z_{ij}] = 0$. Such instruments often consist of observed attributes of other products in the market.

The implied Equation (6.1) together with this mean independence assumption yields the reduced-form model (2.2). Furthermore, we write the innocuous nonparametric projection equation (2.3). Therefore, we apply Algorithm 2 in Section 3.2 for the two-way cluster robust DML estimation of θ_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. (At first glance, it may appear that $\underline{C} = N \wedge M = 20$ is too small for an application of asymptotic theories. While the distribution indeed concentrates at the rate of $1/\sqrt{\underline{C}}$, the effective sample size from the view point of the CLT is in fact $N + M$, which is $20 + 557$ in this application. This is because we apply the CLT to the Hájek projection (see Lemma 1 in the supplementary appendix) which consists of $N+M$ independent summands.) 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 iid 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 5 and those by CCDDHNR (2018a, Remark 3.1). To mitigate the uncertainty induced by sample splitting, we compute estimates based on the average of 10 rerandomized DML following CCDDHNR (2018a, Section 3.4) with variance estimation according to CCDDHNR (2018a, Equation 3.13) adapted to our two-way cluster-robustness.

Table 2. Estimates and standard errors of the coefficient θ_0 of log price in the demand model.

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

NOTES: 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 10 rerandomized DML.

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.

7. Conclusion

In this article, 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 (2018a). The proposed method covers a large class of econometric models as is the case with CCDDHNR (2018a), 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 nonlinear orthogonal scores as in CCDDHNR (2018a; Section 3.3). Second, whereas we focused on unconditional moment restrictions, it may be possible and will be 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.

Acknowledgments

We benefited from useful comments by an anonymous associate editor, two anonymous referees, seminar participants at California Institute of Technology, Southern Methodist University, Stony Brook University, University of Bristol, University of Colorado - Boulder, and University of Toronto, and participants at CeMMAP UCL/Vanderbilt Joint Conference on Advances in Econometrics, CeMMAP Workshop on Causal Learning with Interactions, and 2020 Econometric Society World Congress. All remaining errors are ours.

Funding

H. Chiang's research is supported by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison with funding from the Wisconsin Alumni Research Foundation. K. Kato is partially supported by NSF grants DMS-1952306 and DMS-2014636.

References

Ai, C., and Chen, X. (2003), "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795–1843. [11]

——— (2007), "Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models with Different Conditioning Variables," *Journal of Econometrics*, 141, 5–43. [11]

Athey, S., and Imbens, G. W. (2019): "Machine Learning Methods That Economists Should Know About," *Annual Review of Economics*, 11. [1]

Athey, S., and Wager, S. (2019): "Estimating Treatment Effects with Causal Forests: An Application," arXiv preprint arXiv:1902.07409. [2]

Belloni, A., Chernozhukov, V., Chetverikov, D., and Wei, Y. (2018), "Uniformly Valid Post-regularization Confidence Regions for Many Functional Parameters in z-estimation Framework," *The Annals of Statistics*, 46, 3643–3675. [2]

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017), "Program Evaluation and Causal Inference With High-dimensional Data," *Econometrica*, 85, 233–298. [2]

Belloni, A., Chernozhukov, V., and Hansen, C. (2014a), "High-dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives*, 28, 29–50. [1]

——— (2014b): "Inference on Treatment Effects After Selection Among High-dimensional Controls," *The Review of Economic Studies*, 81, 608–650. [2]

Belloni, A., Chernozhukov, V., Hansen, C., and Kozbur, D. (2016), "Inference in High-dimensional Panel Models With an Application to Gun Control," *Journal of Business & Economic Statistics*, 34, 590–605. [2]

Belloni, A., Chernozhukov, V., and Kato, K. (2015), "Uniform Post-selection Inference for Least Absolute Deviation Regression and Other Z-estimation Problems," *Biometrika*, 102, 77–94. [1]

Berry, S., Levinsohn, J., and Pakes, A. (1995), "Automobile Prices in Market Equilibrium," *Econometrica: Journal of the Econometric Society*, 63, 841–890. [10]

Berry, S. T. (1994), "Estimating Discrete-choice Models of Product Differentiation," *The RAND Journal of Economics*, 50, 242–262. [4,10]

Cameron, A. C., and Miller, D. L. (2015): "A Practitioner's Guide to Cluster-robust Inference," *Journal of Human Resources*, 50, 317–372. [2]

Cameron, C. A., Gelbach, J. B., and Miller, D. L. (2011), "Robust Inference With Multiway Clustering," *Journal of Business and Economic Statistics*, 29, 238 – 249. [1,2,3]

Chen, X., Linton, O., and Van Keilegom, I. (2003), "Estimation of Semiparametric Models When the Criterion Function Is Not Smooth," *Econometrica*, 71, 1591–1608. [11]

Chen, X., and Pouzo, D. (2015), "Sieve Wald and QLR Inferences on Semi/Nonparametric Conditional Moment Models," *Econometrica*, 83, 1013–1079. [11]

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018a), "Double/debiased Machine Learning for Treatment and Structural Parameters," *Econometrics Journal*, 21, C1 – C68. [1,2,3,4,5,7,8,10,11]

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and J. M. Robins (2018b), "Locally Robust Semiparametric Estimation," arXiv preprint:1712.09988. [4]

Chiang, H., and Y. Sasaki (2019), "Lasso Under Multi-way Clustering: Estimation and Post-selection Inference," arXiv preprint:1905.02107. [2,8]

Davezies, L., X. D'Haultfoeuille, and Y. Guyonvarch (2018), "Asymptotic Results under Multiway Clustering," arXiv preprint:1807.07925. [1,2,8]

——— (2019), "Empirical Process Results for Exchangeable Arrays," arXiv preprint:1906.11293. [2]

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. [2]

Kallenberg, O. (2006), *Probabilistic Symmetries and Invariance Principles*, Springer Science & Business Media. New York: Springer. [8]

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. [2]

Kock, A. B., and Tang, H. (2019): "Uniform Inference in High-Dimensional Dynamic Panel Data Models with Approximately Sparse Fixed Effects," *Econometric Theory*, 35, 295–359. [2]

Lee, S., and Ng, S. (2019), "An Econometric View of Algorithmic Subsampling," arXiv preprint:1907.01954. [1]

Lu, Z., Shi, X., and Tao, J. (2019), "Semi-Nonparametric Estimation of Random Coefficient Logit Model for Aggregate Demand," Working Paper. [4,10]

MacKinnon, J. G. (2019), "How Cluster-robust Inference is Changing Applied Econometrics," *Canadian Journal of Economics/Revue Canadienne d'Économique*, 52, 851–881. [2]

MacKinnon, J. G., Nielsen, M. O., and Webb, M. D. (2019), "Wild Bootstrap and Asymptotic Inference with Multiway Clustering," Queen's Economics Department Working Paper, No. 1415. [2]

Menzel, K. (2017), "Bootstrap with Clustering in Two or More Dimensions," arXiv preprint:1703.03043. [1,2,8]

Mullainathan, S., and Spiess, J. (2017), "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31, 87–106. [1]

Okui, R., Small, D. S., Tan, Z., and Robins, J. M. (2012), "Doubly Robust Instrumental Variable Regression," *Statistica Sinica*, 173–205. [3]

Robinson, P. M. (1988), "Root-N-consistent Semiparametric Regression," *Econometrica: Journal of the Econometric Society*, 931–954. [4]

Semenova, V., Goldman, M., Chernozhukov, V., and Taddy, M. (2018), "Orthogonal Machine Learning for Demand Estimation: High Dimensional Causal Inference in Dynamic Panels," arXiv preprint:1608.00033. [2]