arXiv:1905.02107v3 [econ.EM] 21 Aug 2019

LASSO UNDER MULTI-WAY CLUSTERING: ESTIMATION AND POST-SELECTION INFERENCE

HAROLD CHIANG AND YUYA SASAKI

Abstract. This paper studies high-dimensional regression models with lasso when data is sampled under multi-way clustering. First, we establish convergence rates for the lasso and post-lasso estimators. Second, we propose a novel inference method based on a post-double-selection procedure and show its asymptotic validity. Our procedure can be easily implemented with existing statistical packages. Simulation results demonstrate that the proposed procedure works well in finite sample. We illustrate the proposed method with a couple of empirical applications to development and growth economics.

1. Introduction

This paper studies a method of estimation and post-selection inference for regression parameters in high-dimensional linear models by lasso under multi-way clustering. The objective is motivated by recently increasing demands from applied economic research. On one hand, economists often use multi-way cluster sampled data. Examples include, but are not limited to, network data, matched employer-employee data, matched student-teacher data, scanner data where observations are double-indexed by stores and products, market share data where observations are double-indexed by market and products, and growth/development data where observations are double-indexed by ethnicity and geographical units – see Section 7 for specific applications of the last example. On the other hand, researchers also often use machine learning methods of estimation and inference for high-dimensional models in today's big data environments. There are a number of useful methods in the literature that deal with each of these two issues (multi-way clustering and

Date: First arXiv version: May 6, 2019.

Code files are available upon request from the authors.

Key words and phrases. cluster robust standard errors, high dimensions, lasso, machine learning, multi-way clustering, post-selection inference.

JEL Classification: C21, C55.

high dimensionality) separately, but the existing methods do not seem to provide a solution to dealing with both of these practically relevant issues simultaneously. In this light, we present lasso under multi-way clustering, and propose a post-selection inference method for regression parameters under this sampling assumption.

In the important branch of the literature following the seminal work by Belloni, Chen, Chernozhukov and Han (2012), post-selection inference with lasso has been widely studied under various settings by Belloni, Chernozhukov and Hansen (2014)[1], Javanmard and Montanari (2014), van de Geer, Bühlmann, Ritov & (2014), Zhang and Zhang (2014), Belloni, Chernozhukov and Kato (2015), and many others. For empirical researchers, lasso has become a powerful machine learning tool under data-rich environments. Most of the papers in this literature assume i.i.d. or independent sampling. In many empirical applications, it is sometimes more plausible to assume multi-way cluster sampling (e.g., network data, matched employer-employee data, and matched student-teacher data). Building upon Belloni, Chen, Chernozhukov and Hansen (2012), Belloni and Chernozhukov (2013) and Belloni, Chernozhukov and Hansen (2014), this paper generalizes lasso and post-double-selection procedure by allowing for multi-way cluster sampling. To our best knowledge, the present paper is the first in the literature of high-dimensional models to consider lasso under multi-way cluster sampling.

The influential work by Cameron, Gelbach and Miller (2011) proposes multi-way cluster-robust inference methods for linear and nonlinear regression models – also see Cameron and Miller (2015, Section V) for a survey. Formal analysis of asymptotic properties and bootstrap validity under multi-way clustering is studied by Menzel (2017) using the Aldous-Hoover representation – see Kallenberg (2005, Chapter 7) for example. Under the assumptions of separable exchangeability, the method of Menzel (2017) covers both degenerate and non-degenerate cases. Using the same representation, while focusing on the non-degenerate cases, Davezies, D'Haultfoeuille and Guyonvarch (2018) develop empirical process theory under multi-way cluster sampling which applies to a large class of econometric models. Building upon the asymptotic framework of these two papers, MacKinnon, Nielsen and Webb (2019) propose several wild bootstrap procedures for linear regression models, and examine their finite-sample performances under several different cluster sampling scenarios. In this

[1]See also Belloni, Chernozhukov and Hansen (2011).

paper, we take advantage of the innovations by these preceding papers to develop a multi-way cluster-robust inference method for high-dimensional models. To our best knowledge, the present paper is the first in this literature on multi-way clustering to consider high-dimensional models.

The rest of this paper is organized as follows. Section 2 introduces the model. Section 3 presents an overview of the proposed methodology. Section 4 discusses a formal asymptotic theory. Section 5 presents an extension of the baseline results to cases of heterogeneous cluster sizes. Section 6 presents simulation studies. Section 7 presents an empirical illustration with development and growth economics. Section 8 concludes. The appendix contains mathematical proofs and auxiliary lemmas.

2. The Model

Consider the high-dimensional regression model

$$Y_{ij} = D_{ij}\alpha + X_{ij}'\beta + R_{ij}^Y + \varepsilon_{ij},\ \mathrm{E}[\varepsilon_{ij}|D_{ij}, X_{ij}] = 0, \tag{2.1}$$

where Y_{ij} is an observed outcome variable, $(D_{ij}, X_{ij}')'$ is an observed vector of regressors, and R_{ij}^Y is an approximation error for the unit of observation with the double index (i, j). We set α as a scalar parameter of interest. The dimension p of the nuisance parameter vector $\beta \in \mathbb{R}^p$ is potentially increasing in the sample size. Following the literature on high-dimensional post-selection inference (e.g., Belloni, Chernozhukov and Hansen, 2014), we also consider the auxiliary projection

$$D_{ij} = X_{ij}\gamma + R_{ij}^D + v_{ij},\ \mathrm{E}[v_{ij}|X_{ij}] = 0, \tag{2.2}$$

where R^D is an approximation error. The dimension p of the nuisance parameter vector $\gamma \in \mathbb{R}^p$ is the same as that of β, and is potentially increasing in the sample size.

In the absence of two-way clustering, the system (2.1)–(2.2) would be the same as the model considered in Belloni, Chernozhukov and Hansen (2014). We first consider two-way clustering where each cell contains one observation. Section 5 presents an extension to the case of heterogeneous cluster sizes.

3. Overview of the Method

In this section, we present an overview of the proposed method, namely estimation and post-selection inference. Formal theoretical justifications are discussed in Section 4.

A researcher observes a sample $\left\{ (Y_{ij}, D_{ij}, X_{ij}') \middle| \, i \in \{1, ..., N\}, j \in \{1, ..., M\} \right\}$ of size NM. The estimation procedure consists of two steps. First, define the lasso estimates for (2.1) and (2.2) by

$$(\widehat{\alpha}, \widehat{\beta}')' = \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{j=1}^{M} (Y_{ij} - \alpha D_{ij} - X_{ij}'\beta)^2 + \lambda_1 \|(\alpha, \beta')'\|_1, \tag{3.3}$$

$$\text{and} \qquad \widehat{\gamma} = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{j=1}^{M} (D_{ij} - X_{ij}'\gamma)^2 + \lambda_2 \|\gamma\|_1, \tag{3.4}$$

respectively, for some regularization parameters λ_1 and λ_2, valid choices of which are discussed in the statement of Theorem 1 ahead. Denote the supports of the lasso estimates by $\widehat{I}_1 = \text{support}(\widehat{\beta})$ and $\widehat{I}_2 = \text{support}(\widehat{\gamma})$, and let $\widehat{I} = \widehat{I}_1 \cup \widehat{I}_2$. In the second step, define the post-double-selection lasso estimate $\widetilde{\alpha}$ by

$$(\widetilde{\alpha}, \widetilde{\beta}') = \underset{\text{support}(\beta) \subset \widehat{I}}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{j=1}^{M} (Y_{ij} - \alpha D_{ij} - X_{ij}'\beta)^2. \tag{3.5}$$

Let $\underline{C} = N \wedge M$, $\mu_N = \underline{C}/N$ and $\mu_M = \underline{C}/M$. Under suitable conditions to be formally stated in Section 4, we have the asymptotic normality

$$\sigma^{-1} \sqrt{\underline{C}} (\widetilde{\alpha} - \alpha) \rightsquigarrow N(0, 1),$$

where the asymptotic variance is given by $\sigma^2 = Q^{-1} \Gamma Q^{-1}$ with

$$\begin{aligned} Q =& \mathrm{E}[v_{11}^2], \\ \Gamma =& \bar{\mu}_N \Gamma_N + \bar{\mu}_M \Gamma_M = \bar{\mu}_N \mathrm{E}[v_{11} \varepsilon_{11} v_{12} \varepsilon_{12}] + \bar{\mu}_M \mathrm{E}[v_{11} \varepsilon_{11} v_{21} \varepsilon_{21}], \end{aligned}$$

and $\bar{\mu}_N$ and $\bar{\mu}_M$ denoting the limits of μ_N and μ_M, respectively.

The asymptotic variance is estimated by the sample counterpart $\widehat{\sigma}^2 = \widehat{Q}^{-1}\widehat{\Gamma}\widehat{Q}^{-1}$, where

$$\widehat{Q} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\widehat{v}_{ij}^2,$$

$$\widehat{\Gamma} = \frac{\underline{C}}{(NM)^2}\sum_{i=1}^{N}\sum_{1\leq j,j'\leq M}\widehat{v}_{ij}\widehat{\varepsilon}_{ij}\widehat{\varepsilon}_{ij'}\widehat{v}_{ij'} + \frac{\underline{C}}{(NM)^2}\sum_{1\leq i,i'\leq N}\sum_{j=1}^{M}\widehat{v}_{ij}\widehat{\varepsilon}_{ij}\widehat{\varepsilon}_{i'j}\widehat{v}_{i'j},$$

$\widehat{v}_{ij} = D_{ij} - X_{ij}'\widehat{\gamma}$, and $\widehat{\varepsilon}_{ij} = Y_{ij} - \widehat{\alpha}D_{ij} - X_{ij}'\widehat{\beta}$.

In summary, we propose to report the post-double-selection lasso estimate $\widetilde{\alpha}$ as an estimate of α with its standard error given by $\widehat{\sigma}/\sqrt{\underline{C}}$. The α^*-level confidence interval can be constructed as $\left[\widetilde{\alpha} + \Phi^{-1}(\alpha^*/2)\widehat{\sigma}/\sqrt{\underline{C}}, \widetilde{\alpha} + \Phi^{-1}(1-\alpha^*/2)\widehat{\sigma}/\sqrt{\underline{C}}\right]$, where Φ^{-1} denotes the quantile function of the standard normal distribution.

4. Asymptotic Theory

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 non-decreasing in n and $M(n)N(n)$ is increasing in n. For simplicity, each size of intersection n_{ij} is assumed to be uniformly bounded by a positive integer $\bar{n}$ that is independent of n. With this said, we will suppress the index notation and write (N, M) for simplicity. We fix a number of notations. For each n, let P_n denote the law with respect to sample size (N, M) – note that we allow the dimension p of X_{ij} to grow with n. Let $a := p \vee (NM)$. Also recall the notations $\underline{C} = N \wedge M$, $\mu_N = \underline{C}/N$, and $\mu_M = \underline{C}/M$ from Section 3. We use the short-hand notation $[k] = \{1, ..., k\}$ and $[k]^c = \mathbb{N}\backslash[k]$ for any $k \in \mathbb{N}$. For a sequence $(t_{ij})_{i\in[N],j\in[M]}$, denote $\|t_{ij}\|_n = \sqrt{\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}t_{ij}^2}$. Thus, $\|X_{ij}'\delta\|_n = \sqrt{\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\delta' X_{ij}X_{ij}'\delta}$ is the prediction norm of δ. Let $\|A\|_\infty = \max_{k,l}|A_{k,l}|$ denote the max norm of matrix A. We write $a \lesssim b$ to mean $a \leqslant cb$ for some $c > 0$ that does not depend on n. We also write $a \lesssim_{\mathrm{P}} b$ to mean $a = O_P(b)$. We write $Z_{ij} = (Y_{ij}, D_{ij}, X_{ij}')'$ for the $(p+2)$-dimensional random vector in data. Throughout, we assume that this random vector Z_{ij} is Borel measurable – see Kallenberg (2005, pp. 304). With these notations, we state the following four assumptions.

Assumption 1 (Sampling)**.** *Suppose that* $\underline{C} \to \infty$, $\mu_N \to \bar{\mu}_N \geqslant 0$, *and* $\mu_M \to \bar{\mu}_M \geqslant 0$.

(1) $(Z_{ij})_{(i,j)\in\mathbb{N}^2}$ *is an infinite sequence of separately exchangeable* $(p+2)$*-dimensional random vectors. That is, for any permutations* π_1 *and* π_2 *of* $\mathbb{N}$*, we have*

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

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

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

Assumption 2 (Moments)**.** *There exists a sequence* $\{B_n\}_{n=1}^{\infty}$ *of positive constants such that the following conditions hold for all* $n \in \mathbb{N}$ *for some* $q > 4$*:*

(1) $\mathrm{E}[|D_{11}|^{2q}] + \max_{k\in[p]} \mathrm{E}[|X_{11,k}|^{2q}] + \mathrm{E}[|\varepsilon_{11}|^{2q}|X_{11}, v_{11}] + \mathrm{E}[|v_{11}|^{2q}|X_{11}] \leq K$ *a.s. and* $0 < c \leq \mathrm{E}[v_{11}^2|X_{11}]$ *a.s. for positive constants,* c *and* K*, that are independent of* n*.*

(2) $\mathrm{E}[\|X_{11}\|_\infty^{2q}] \leq B_n^{2q}$ *and* $B_n\sqrt{\log a} \lesssim (N \vee M)^{1/2-1/q}$*.*

(3) $\bar{\mu}_N \mathrm{E}[v_{11}\varepsilon_{11}v_{12}\varepsilon_{12}] + \bar{\mu}_M \mathrm{E}[v_{11}\varepsilon_{11}v_{21}\varepsilon_{21}]$ *and the maximal and minimal eigenvalues of* $\mathrm{E}[X_{11}X_{11}']$ *are bounded and bounded away from zero uniformly in* n*.*

Assumption 3 (Sparsity)**.**

(1) $\|\beta\|_0 + \|\gamma\|_0 \lesssim s$*.*

(2) $\|R_{ij}^Y\|_n + \|R_{ij}^D\|_n \leq c_s \lesssim_{\mathrm{P}} \sqrt{s/\underline{C}}$*.*

(3) $\frac{s_n^2(\log(a))^2}{\underline{C}} = o(1)$*.*

Assumption 4 (Sparse Eigenvalues)**.** *There exists a sequence* $\{\ell_n\}$ *such that* $\ell_n \to \infty$ *and, with probability at least* $1 - o(1)$*,*

$$0 < c \leq \phi_{\min}(s\ell_n) \leq \phi_{\max}(s\ell_n) \leq c' < \infty$$

holds for some constants, c *and* c'*, that are independent of* n*, where*

$$\phi_{\max}(m) := \max_{1\leq\|\delta\|_0\leq m} \frac{\delta' M \delta}{\|\delta\|^2} \text{ and } \phi_{\min}(m) := \min_{1\leq\|\delta\|_0\leq m} \frac{\delta' M \delta}{\|\delta\|^2},$$

with $M = \frac{1}{NM}\sum_{i=1}^N \sum_{j=1}^M X_{ij}X_{ij}'$*, denote the maximal and minimal* m*-sparse eigenvalues.*

Remark 4.1 (Discussion of the Assumptions)**.** Assumption 1 is closely related to Assumption 1 of Davezies, D'Haultfoeuille and Guyonvarch (2018). The main difference is that we allow p to be changing with n. We remark that the exchangeability assumption is not new in econometrics – it has been used in Andrews (2005) and Menzel (2015) as well as Menzel

(2017), Davezies, D'Haultfoeuille and Guyonvarch (2018), and MacKinnon, Nielsen and Webb (2019). Assumption 2 is standard in the literature on post-selection inference with lasso. Parts (1) and (2) require an existence of higher order moments of key objects. Note that common assumptions in high-dimensional literature, such as sub-gaussianity or boundedness, are not required. They can be replaced by some higher level conditions similar to Condition RF of Belloni, Chen, Chernozhukov and Hansen (2012).[2] Part (3) of Assumption 2 requires that the asymptotic variance is bounded away from zero.[3] Assumption 3 is a direct generalization of Condition ASTE (iii) and (iv) of Belloni, Chernozhukov and Hansen (2014). Finally, Assumption 4 is analogous to Condition SE of Belloni, Chernozhukov and Hansen (2014), which is standard in the high-dimensional literature. It only imposes small diagonal submatrices to be well behaved. △

4.1. **Independentization via Hájek Projection.** In this section, we show that an empirical process in multi-way clustered samples can be represented as a sum of independent variables via Hájek projection. Furthermore, its variance can be shown to be approximated by covariances of observed variables.

For any $f : \text{support}(Z) \to \mathbb{R}$, we let

$$\mathbb{G}_C f := \sqrt{\underline{C}}\Big\{\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} f(Z_{ij}) - \mathrm{E}[f(Z_{11})]\Big\}$$

denote its empirical process.

Lemma 1 (Independentization via Hájek Projection)**.** *If Assumption 1 holds and* $f : \text{support}(Z) \to \mathbb{R}$ *satisfies* $\mathrm{E}f^2(Z_{11}) < K$ *for a finite constant* K *that is independent of* n*, then there exist i.i.d. uniform random variables* U_{i0} *and* U_{0j} *such that the Hájek projection* $H_n f$ *of* $\mathbb{G}_C f$ *on*

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

[2]See their Lemma 3.

[3]Similarly to Davezies, D'Haultfoeuille and Guyonvarch (2018), we focus on non-degenerate cases in this paper. See Menzel (2017) for the studies of degenerate cases using a bootstrap-based method.

is equal to

$$H_n f = \sum_{i=1}^{N} \frac{\sqrt{C}}{N} \mathrm{E}\Big[f(Z_{i1}) - \mathrm{E} f(Z_{11}) \Big| U_{i0} \Big] + \sum_{j=1}^{M} \frac{\sqrt{C}}{M} \mathrm{E}\Big[f(Z_{1j}) - \mathrm{E} f(Z_{11}) \Big| U_{0j} \Big]$$

for each n. Furthermore,

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

holds a.s.

A proof of this lemma can be found in Appendix A.1. The first part of the lemma shows that an empirical process $\mathbb{G}_C f$ under multi-way cluster sampling can be represented as a sum of independent unobserved variables via Hájek projection $H_n f$. While U_{i0} and U_{0j} are unobserved, the second part of this lemma in turn shows that the variance of the Hájek projection can be approximated by covariances of observed variables. Note that, since $H_n f$ is a Hájek projection, the lemma implies $\frac{\mathbb{G}_C f}{\sqrt{V(\mathbb{G}_C f)}} = \frac{H_n f}{\sqrt{V(H_n f)}} + o_{\mathrm{P}}(1)$ if $\bar{\mu}_N Cov(f(Z_{11}), f(Z_{12})) + \bar{\mu}_M Cov(f(Z_{11}), f(Z_{21}))$ is bounded and bounded away from zero uniformly in n.

Our Lemma 1 can be seen as an extension to Lemma D.2 in Davezies, D'Haultfoeuille and Guyonvarch (2018). Specifically, while Davezies, D'Haultfoeuille and Guyonvarch (2018) consider a fixed data generating process over the sample size n, our Lemma 1 allows the data generating process to vary with n in particular for the sake of accommodating the increasing of dimensionality p for high-dimensional models. The lemma serves as a main building block for all the asymptotic results to be presented ahead.

4.2. **Convergence Rates of Lasso and Post-Lasso under Multi-Way Clustering.** We next show the convergence rates of the lasso estimator $(\widehat{\alpha}, \widehat{\beta}', \widehat{\gamma}')'$ and the post-lasso estimator $(\widetilde{\alpha}, \widetilde{\beta}', \widetilde{\gamma}')'$ under multi-way clustering.

Theorem 1 (Convergence Rates for Lasso and Post-Lasso under Multi-Way Clustering)**.** *If Assumptions 1, 2 (1)–(2), 3 (1)–(2), and 4 are satisfied, and* $\lambda_1, \lambda_2 = C\sqrt{(NM)^2 \log a / \underline{C}}$

for some constant $C > 1$, then

$$\|\widehat{\eta} - \eta\|_1 + \|\widehat{\gamma} - \gamma\|_1 \lesssim \sqrt{\frac{s^2 \log a}{\underline{C}}}, \qquad \|W_{ij}'(\widehat{\eta} - \eta)\|_n + \|X_{ij}(\widehat{\gamma} - \gamma)\|_n \lesssim \sqrt{\frac{s \log a}{\underline{C}}},$$

$$\|\widetilde{\eta} - \eta\|_1 + \|\widetilde{\gamma} - \gamma\|_1 \lesssim \sqrt{\frac{s^2 \log a}{\underline{C}}}, \qquad \|W_{ij}'(\widetilde{\eta} - \eta)\|_n + \|X_{ij}(\widetilde{\gamma} - \gamma)\|_n \lesssim \sqrt{\frac{s \log a}{\underline{C}}}, \quad \textit{and}$$

$$\|\widehat{\eta} - \eta\| + \|\widehat{\gamma} - \gamma\| + \|\widetilde{\eta} - \eta\| + \|\widetilde{\gamma} - \gamma\| \lesssim \sqrt{\frac{s \log a}{\underline{C}}}$$

hold, where $W_{ij} = [D_{ij}, X_{ij}']'$ and $\eta = (\alpha, \beta')'$.

A proof can be found in Appendix A.2, and is based on the previous result (Lemma 1). In the multi-way sampling, this lemma can be viewed as a counterpart of Lemma 6 and Lemma 7 in Belloni, Chen, Chernozhukov and Hansen (2012).

4.3. **Post-Selection-Inference with Post-Lasso under Multi-way Clustering.** In this section, we present the main result of this paper. The limit normal distribution of the post-double-selection lasso estimate $\widetilde{\alpha}$ is established based on the previous two results (Lemma 1 and Theorem 1).

Theorem 2 (Asymptotic Normality)**.** *If Assumptions 1, 2, 3 and 4 are satisfied, and λ_1 and λ_2 are chosen according to the statement of Theorem 1, then*

$$\sigma^{-1}\sqrt{\underline{C}}(\widetilde{\alpha} - \alpha) \rightsquigarrow N(0, 1),$$

where $\sigma^2 = Q^{-1}\Gamma Q^{-1}$, $Q = \mathrm{E}[v_{11}^2]$ and

$$\Gamma = \bar{\mu}_N \Gamma_N + \bar{\mu}_M \Gamma_M = \bar{\mu}_N \mathrm{E}[v_{11}\varepsilon_{11}v_{12}\varepsilon_{12}] + \bar{\mu}_M \mathrm{E}[v_{11}\varepsilon_{11}v_{21}\varepsilon_{21}].$$

A proof can be found in Appendix A.3. This result provides a theoretical justification for the asymptotic variance proposed in the overview in Section 3. In practice, we do not know the components, Q and Γ, of the asymptotic variance. The following subsection proposes estimators of them.

4.4. **Variance Estimation.** In this section, we propose an analog variance estimator. The components, Q and Γ, of the asymptotic variance can be estimated by

$$\widehat{Q} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\widehat{v}_{ij}^2 \qquad \text{and}$$

$$\widehat{\Gamma} = \frac{\underline{C}}{(NM)^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}\widehat{v}_{ij}\widehat{\varepsilon}_{ij}\widehat{\varepsilon}_{ij'}\widehat{v}_{ij'} + \frac{\underline{C}}{(NM)^2}\sum_{1\le i,i'\le N}\sum_{j=1}^{M}\widehat{v}_{ij}\widehat{\varepsilon}_{ij}\widehat{\varepsilon}_{i'j}\widehat{v}_{i'j},$$

respectively, where $\widehat{v}_{ij} = D_{ij} - X_{ij}'\widehat{\gamma}$ and $\widehat{\varepsilon}_{ij} = Y_{ij} - \widehat{\alpha}D_{ij} - X_{ij}'\widehat{\beta}$ are the residuals. With these component estimators, we propose that the asymptotic variance $\sigma^2 = Q^{-1}\Gamma Q^{-1}$ be estimated by $\widehat{\sigma}^2 = \widehat{Q}^{-1}\widehat{\Gamma}\widehat{Q}^{-1}$. The following theorem provides a theoretical support for this variance estimator.

Theorem 3 (Variance Estimation)**.** *If Assumptions 1, 2, 3 and 4 are satisfied, λ_1 and λ_2 are chosen according to the statement of Theorem 1, $\frac{(NM)^{1/q}B_n^2 s^3(\log a)^2}{\underline{C}^2} = o(1)$, $\frac{(NM)^{1/q}s\log a}{\underline{C}} = o(1)$, and $\|R_{ij}^D R_{ij}^Y\|_n^2 = O(1)$, then the variance estimator $\widehat{\sigma}^2 = \widehat{Q}^{-1}\widehat{\Gamma}\widehat{Q}^{-1}$ is consistent for $\sigma^2 = Q^{-1}\Gamma Q^{-1}$.*

A proof is found in Appendix A.4. In light of this result, we propose to compute the standard error by $\widehat{\sigma}/\sqrt{\underline{C}}$. Similarly, in light of this result together with Theorem 2, we propose to construct the α^*-level confidence interval by $\left[\widetilde{\alpha} + \Phi^{-1}(\alpha^*/2)\widehat{\sigma}/\sqrt{\underline{C}}, \widetilde{\alpha} + \Phi^{-1}(1-\alpha^*/2)\widehat{\sigma}/\sqrt{\underline{C}}\right]$, where Φ^{-1} denotes the quantile function of the standard normal distribution.

5. Extension: Heterogeneous Cluster Sizes

Thus far, we focus on the case where each cluster contains one observation. In this section, we presented an extension of the baseline results to situations where the numbers of observations are heterogeneous across clusters. Suppose that we have n_{ij} observations for each cell $(i,j) \in [N]\times[M]$, where n_{ij} is a random variable that is allowed to depend on $(X_{ij,\ell})_{\ell\ge 1}$. To deal with the situation of $n_{ij} = 0$, for any sequence $(t_\ell)_{\ell\ge 1}$, define $\sum_{\ell=1}^{0} t_\ell = 0$. Consider the model

$$Y_{ij,\ell} = D_{ij,\ell}\alpha + X_{ij,\ell}'\beta + R_{ij,\ell}^Y + \varepsilon_{ij,\ell},\ \mathrm{E}[\varepsilon_{ij,\ell}|D_{ij,\ell}, X_{ij,\ell}] = 0,$$

where $Y_{ij,\ell}$ is an observed outcome variable, $(D_{ij,\ell}, X'_{ij,\ell})'$ is an observed vector of regressors, and $R^Y_{ij,\ell}$ is an approximation error for the unit $\ell \in [n_{ij}]$ with the double index (i,j) indicating i-th cluster in the first clustering dimension and j-th cluster in the second clustering dimension. Using matrix notations, we can rewrite the model as

$$Y_{ij} = D_{ij}\alpha + X_{ij}\beta + R^Y_{ij} + \varepsilon_{ij},\ \mathrm{E}[\varepsilon_{ij}|D_{ij}, X_{ij}] = 0,$$

where each of $Y_{ij} = (Y_{ij,\ell})_{\ell\in[n_{ij}]}$, $D_{ij} = (D_{ij,\ell})_{\ell\in[n_{ij}]}$, $R^Y_{ij} = (R^Y_{ij,\ell})_{\ell\in[n_{ij}]}$, and $\varepsilon_{ij} = (\varepsilon_{ij,\ell})_{\ell\in[n_{ij}]}$ is of dimension $n_{ij}\times 1$, and $X_{ij} = (X'_{ij,\ell})_{\ell\in[n_{ij}]}$ is of dimension $n_{ij}\times p$. We similarly write the accompanying auxiliary projection as

$$D_{ij} = X_{ij}\gamma + R^D_{ij} + v_{ij},\ \mathrm{E}[v_{ij}|X_{ij}] = 0,$$

where R^D is of dimension $n_{ij}\times 1$ representing approximation errors, and v_{ij} is of dimension $n_{ij}\times 1$ representing projection errors.

Under this setting, the first step of estimation procedure consists of

$$(\widehat{\alpha}, \widehat{\beta}')' = \underset{\alpha,\beta}{\mathrm{argmin}} \sum_{i=1}^N\sum_{j=1}^M\sum_{\ell=1}^{n_{ij}} (Y_{ij,\ell} - \alpha D_{ij,\ell} - X'_{ij,\ell}\beta)^2 + \lambda_1\|(\alpha,\beta')'\|_1$$

$$\text{and} \qquad \widehat{\gamma} = \underset{\gamma}{\mathrm{argmin}} \sum_{i=1}^N\sum_{j=1}^M\sum_{\ell=1}^{n_{ij}} (D_{ij,\ell} - X'_{ij,\ell}\gamma)^2 + \lambda_2\|\gamma\|_1.$$

In turn, the second-step estimates are obtained by

$$(\widetilde{\alpha}, \widetilde{\beta}') = \underset{\mathrm{support}(\beta)\subset\widehat{I}}{\mathrm{argmin}} \sum_{i=1}^N\sum_{j=1}^M\sum_{\ell=1}^{n_{ij}} (Y_{ij,\ell} - \alpha D_{ij,\ell} - X'_{ij,\ell}\beta)^2.$$

The asymptotic variance estimator for $\widetilde{\alpha}$ is given by $\widehat{\sigma}^2 = \widehat{Q}^{-1}\widehat{\Gamma}\widehat{Q}^{-1}$, where

$$\widehat{Q} = \frac{1}{NM}\sum_{i=1}^N\sum_{j=1}^M \widehat{v}'_{ij}\widehat{v}_{ij},$$

$$\widehat{\Gamma} = \frac{C}{(NM)^2}\sum_{i=1}^N \sum_{1\le j,j'\le M} \widehat{v}'_{ij}\widehat{\varepsilon}_{ij}\widehat{\varepsilon}'_{ij'}\widehat{v}_{ij'} + \frac{C}{(NM)^2}\sum_{1\le i,i'\le N}\sum_{j=1}^M \widehat{v}'_{ij}\widehat{\varepsilon}_{ij}\widehat{\varepsilon}'_{i'j}\widehat{v}_{i'j},$$

$\widehat{v}_{ij} = D_{ij} - X_{ij}\widehat{\gamma}$, and $\widehat{\varepsilon}_{ij} = Y_{ij} - \widehat{\alpha}D_{ij} - X_{ij}\widehat{\beta}$.

We now formally state assumptions for the extended theory to support the asymptotic validity of this procedure. Define $W_{ij} = (n_{ij}, (Z_{ij,\ell})_{\ell\ge 1})$ and $\ddot{M} = \frac{1}{NM}\sum_{i=1}^N\sum_{j=1}^M X'_{ij}X_{ij}$.

Assumption 5 (Sampling)**.** *Suppose that* $\underline{C}\to\infty$, $\mu_N \to \bar{\mu}_N \geqslant 0$, *and* $\mu_M \to \bar{\mu}_M \geqslant 0$.

(1) $(W_{ij})_{(i,j)\in\mathbb{N}^2}$ *is an infinite sequence of separately exchangeable random processes.*

(2) $(W_{ij})_{(i,j)\in\mathbb{N}^2}$ *is dissociated.*

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

(4) $\mathrm{E}[n_{ij}]>0$ *and* $n_{ij}\leq\bar{n}$ *for a positive finite constant* $\bar{n}$ *independent of* n*.*

Assumption 6 (Moments)**.** *There exists a sequence* $\{B_n\}_{n=1}^{\infty}$ *of positive constants such that the following conditions hold for all* $n\in\mathbb{N}$ *for some* $q>4$*:*

(1) $\mathrm{E}[\max_{\ell\in[n_{ij}]}|D_{11,\ell}|^{2q}]+\max_{k\in[p]}\mathrm{E}[\max_{\ell\in[n_{ij}]}|X_{11,\ell,k}|^{2q}]+\mathrm{E}[\max_{\ell\in[n_{ij}]}|\varepsilon_{11,\ell}|^{2q}|X_{11,\ell},v_{11,\ell}]+\mathrm{E}[|v_{11,\ell}|^{2q}|X_{11,\ell}]\leq K$ *a.s. and* $0<c\leq\mathrm{E}[\max_{\ell\in[n_{ij}]}v_{11,\ell}^2|X_{11,\ell}]$ *a.s. for positive constants,* c *and* K*, that are independent of* n*.*

(2) $\mathrm{E}[\max_{\ell\in[n_{ij}]}\|X_{11}\|_\infty^{2q}]\leq B_n^{2q}$ *and* $B_n\sqrt{\log a}\lesssim(N\vee M)^{1/2-1/q}$*.*

(3) $\bar{\mu}_N\mathrm{E}[v_{11}'\varepsilon_{11}\varepsilon_{12}'v_{12}]+\bar{\mu}_M\mathrm{E}[v_{11}'\varepsilon_{11}\varepsilon_{21}'v_{21}]$ *and the maximal and minimal eigenvalues of* $\mathrm{E}[X_{11}'X_{11}]$ *are bounded and bounded away from zero uniformly in* n*.*

Assumption 7 (Sparsity)**.**

(1) $\|\beta\|_0+\|\gamma\|_0\lesssim s$*.*

(2) $\sqrt{(NM)^{-1}\sum_{i=1}^N\sum_{j=1}^M\sum_{\ell\in[n_{ij}]}(R_{ij,\ell}^Y)^2}+\sqrt{(NM)^{-1}\sum_{i=1}^N\sum_{j=1}^M\sum_{\ell\in[n_{ij}]}(R_{ij,\ell}^D)^2}\leq c_s\lesssim_{\mathrm{P}}\sqrt{s/\underline{C}}$*.*

(3) $\frac{s_n^2(\log(a))^2}{\underline{C}}=o(1)$*.*

Assumption 8 (Sparse Eigenvalues)**.** *There exists a sequence* $\{\ell_n\}$ *such that* $\ell_n\to\infty$ *and, with probability at least* $1-o(1)$*,*

$$0<c\leq\phi_{\min}(s\ell_n)\leq\phi_{\max}(s\ell_n)\leq c'<\infty$$

holds for some constants, c *and* c'*, that are independent of* n*, where*

$$\phi_{\max}(m):=\max_{1\leq\|\delta\|_0\leq m}\frac{\delta'\ddot{M}\delta}{\|\delta\|^2}\text{ and }\phi_{\min}(m):=\min_{1\leq\|\delta\|_0\leq m}\frac{\delta'\ddot{M}\delta}{\|\delta\|^2}.$$

The following statement provides a theoretical guarantee for the estimation and inference procedure for the extended model outlined above.

Corollary 1. *If Assumptions 5, 6, 7 and 8 are satisfied, and* λ_1 *and* λ_2 *are chosen according to the statement of Theorem 1, then*

$$\sigma^{-1}\sqrt{\underline{C}}(\widetilde{\alpha}-\alpha)\rightsquigarrow N(0,1),$$

where $\sigma^2 = Q^{-1}\Gamma Q^{-1}$ *with*

$$\begin{aligned} Q =& \mathrm{E}[v_{11}' v_{11}], \\ \Gamma =& \bar{\mu}_N \mathrm{E}[v_{11}' \varepsilon_{11} v_{12}' \varepsilon_{12}] + \bar{\mu}_M \mathrm{E}[v_{11}' \varepsilon_{11} v_{21}' \varepsilon_{21}]. \end{aligned}$$

Furthermore, if $\frac{(NM)^{1/q} B_n^2 s^3 (\log a)^2}{\underline{C}^2} = o(1)$, $\frac{(NM)^{1/q} s \log a}{\underline{C}} = o(1)$, *and* $\|R_{ij}^D R_{ij}^Y\|_n^2 = O(1)$, *then the variance estimator* $\widehat{\sigma}^2$ *is consistent for* σ^2.

A proof of Corollary 1 closely follows that of the results in Section 4, and are therefore omitted. The key difference is that we now apply Aldous-Hoover representation on W_{ij} rather than on Z_{ij}.[4]

6. Simulation Studies

In this section, we present simulation studies of finite-sample performance of the proposed method of estimation and post-selection inference. We compare the performance of our method against existing alternatives from the lasso literature that do not account for multi-way clustering.

6.1. **Simulation Setup.**

We consider the linear model

$$Y_{ij} = D_{ij}\alpha + X_{ij}'\beta + \varepsilon_{ij}.$$

The parameter values are fixed at $(\alpha, \beta')' = \left(0.5, 0.5^2, \cdots, 0.5^{\dim(X)+1}\right)'$. The random vector $(D_{ij}, X_{ij}', \varepsilon_{ij})$ is constructed by

$$\begin{aligned} (D_{ij}, X_{ij}) =& (1 - \omega_1^x - \omega_2^x)\upsilon_{ij}^x + \omega_1^x \upsilon_i^x + \omega_2^x \upsilon_j^x \qquad \text{and} \\ \varepsilon_{ij} =& (1 - \omega_1^\varepsilon - \omega_2^\varepsilon)\upsilon_{ij}^\varepsilon + \omega_1^\varepsilon \upsilon_i^\varepsilon + \omega_2^\varepsilon \upsilon_j^\varepsilon \end{aligned}$$

[4]For more insights on this extension, see Section 3.1 of Davezies, D'Haultfoeuille and Guyonvarch (2019).

with two-way clustering weights (ω_1^x, ω_2^x) and $(\omega_1^\varepsilon, \omega_2^\varepsilon)$, where υ_{ij}^x, υ_i^x, and υ_j^x are independently generated according to

$$\upsilon_{ij}^x, \upsilon_i^x, \upsilon_j^x \sim N\left(0, \begin{pmatrix} \rho^0 & \rho^1 & \cdots & \rho^{\dim(X)-1} & \rho^{\dim(X)} \\ \rho^1 & \rho^0 & \cdots & \rho^{\dim(X)-2} & \rho^{\dim(X)-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho^{\dim(X)-1} & \rho^{\dim(X)-2} & \cdots & \rho^0 & \rho^1 \\ \rho^{\dim(X)} & \rho^{\dim(X)-1} & \cdots & \rho^1 & \rho^0 \end{pmatrix}\right),$$

and $\upsilon_{ij}^\varepsilon$, υ_i^ε, and υ_j^ε are independently generated according to

$$\upsilon_{ij}^\varepsilon, \upsilon_i^\varepsilon, \upsilon_j^\varepsilon \sim N(0,1).$$

Note that the weights (ω_1^x, ω_2^x) and $(\omega_1^\varepsilon, \omega_2^\varepsilon)$ specify the extent of dependence in two-way clustering in (D_{ij}, X_{ij}') and ε_{ij}, respectively. Also, the parameter ρ specifies the extent of collinearity among the high-dimensional covariates (D_{ij}, X_{ij}'). We set $(\omega_1^x, \omega_2^x) = (0.25, 0.25)$, $(\omega_1^\varepsilon, \omega_2^\varepsilon) = (0.25, 0.25)$, and $\rho = 0.50$.

6.2. **Alternative Variance Estimators.** We compare the performance of our multi-way cluster-robust variance estimator with two existing alternative benchmarks. One is the heteroskedasticity robust variance estimator (such as the one in Belloni, Chernozhukov and Hansen (2014)) without accounting for cluster sampling, i.e., Γ is estimated by

$$\widehat{\Gamma}_{HC} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\widehat{v}_{ij}^2\widehat{\varepsilon}_{ij}^2.$$

We will refer to this variance estimator $\widehat{Q}^{-1}\widehat{\Gamma}_{HC}\widehat{Q}^{-1}$ as the '0-Way' estimator. The other is the one-way cluster-robust variance estimator (similar to those of Belloni, Chernozhukov and Hansen and Kozbu (2016) and Kock (2016)) clustered at one (e.g., second) dimension, i.e., Γ is estimated by

$$\widehat{\Gamma}_{CR} = \frac{1}{NM^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}\widehat{v}_{ij}\widehat{\varepsilon}_{ij}\widehat{v}_{ij'}\widehat{\varepsilon}_{ij'}.$$

We will refer to this variance estimator $\widehat{Q}^{-1}\widehat{\Gamma}_{CR}\widehat{Q}^{-1}$ as the '1-Way' estimator.

6.3. **Results.** Table 1 summarizes simulation results. The first two columns indicate the two-way sample sizes (N, M). The third column indicates the dimension (Dim) of $(\alpha, \beta')'$. The next four columns report simulation statistics for $\widetilde{\alpha}$. These statistics include the average (Avg), bias (Bias), standard deviation (SD), and root mean square error (RMSE). The last three columns report 95% coverage frequencies of α based on three variance estimators. The first is the heteroskedasticity robust variance estimator (0-Way). The second is the one-way cluster-robust variance estimator (1-Way). The third is our multi-way cluster-robust variance estimator (2-Way). The results are based on 25,000 Monte Carlo iterations for each row in the table.

In view of the statistics columns, observe that the post-double-selection lasso estimate $\widetilde{\alpha}$ behaves well in larger sample sizes (e.g., $N, M \geqslant 20$) both in terms of bias and variance. Next, observe the 95% coverage frequencies by the three alternative variance estimators. Both the 0-Way and 1-Way variance estimators significantly underestimate the variances of the post-double-selection lasso estimate $\widetilde{\alpha}$. On the other hand, the coverage frequency based on our 2-Way variance estimator approaches the nominal probability (95%) as the sample size increases. These results demonstrate that, when the true sampling process entails multi-way clustering, traditional variance estimators may bias the inference and our multi-way cluster-robust variance estimator performs robustly well.

7. Empirical Illustrations

In this section, we illustrate our proposed method with applications to a couple of empirical studies. There is a sequence of recent growth and development economic studies using empirical data that are clustered at ethnic and geographical levels (e.g., Nunn and Wantchekon, 2011; Michalopoulos and Papaioannou, 2013, 2014, 2016; Gershman, 2016; Anderson, 2018; Dickens, 2018). The next two subsections present how our method can enrich the model flexibility and robustness of such studies, focusing on the cases of Nunn and Wantchekon (2011) and Michalopoulos and Papaioannou (2013).

7.1. **Slave Trade and Mistrust in Africa.** Nunn and Wantchekon (2011) analyze the effects of slave trade on mistrust in Africa, controlling for various demongraphic and geographical covariates including age, age squared, ethnic fractionalization, gender, urban residence, occupation, religion, and living conditions as well as country fixed effects in their

baseline model. Estimates of these effects are obtained by running regressions with a sample that pools n_{ij} individuals $\ell \in [n_{ij}]$ in ethnic group i and districts j across the cells $(i,j) \in [N] \times [M]$ of $N(= 185)$ ethnic groups and $M(= 1257)$ districts. Standard errors are computed by the two-way cluster-robust method of Cameron, Gelbach and Miller (2011) for the ethnic group and district as two ways of clustering.

With our proposed method that is applicable to both high-dimensional models and multi-way clustering, they could consider even more flexible model specifications, for example, allowing for higher orders of age rather than just the quadratic specification and interactions of the age polynomials with various other dummy variables. We present estimates with standard errors under such extended models with flexible specifications, demonstrate that qualitatively similar results continue to be obtained without substantial loss of statistical significance, and thus confirm further robustness of the main empirical findings by Nunn and Wantchekon (2011).

Consider the model

$$Y_{ij,\ell} = D_{ij,\ell}\alpha + X'_{ij,\ell}\beta + R^Y_{ij,\ell} + \varepsilon_{ij,\ell},\ \mathrm{E}[\varepsilon_{ij,\ell}|D_{ij,\ell}, X_{ij,\ell}] = 0,$$

where $Y_{ij,\ell}$ denotes a measure of trust, $D_{ij,\ell}$ denotes an intensity measure of slave trade, $X_{ij,\ell}$ contains polynomial basis elements of age up to degree 10, ethnic fractionalization, gender, urban residence, occupation, religion, living conditions, the interactions of the polynomial basis of age with all the dummy variables, and country fixed effects, consisting of 597 dimensions of covariates in total. Note that the total number of regressors ($p + 1 = 598$) is much larger than the effective sample size ($\underline{C} = N \wedge M = 185$) of two-way clustering in this extended setting.

Table 2 summarize the estimates of the effects of slave trade on mistrust as measured by the "trust of neighbors," corresponding to Table 1 of Nunn and Wantchekon (2011). The last two columns in the table show the original estimates obtained under the prototypical model by Nunn and Wantchekon (2011, Table 1) and corresponding lasso estimates obtained under more flexible model specification by our method. Across all the measures of slave exports, the original estimates and our lasso estimates are similar with similar levels of statistical significance. These results demonstrate that, even for flexible model specifications entailing high-dimensional covariates, the proposed method allows to produce

qualitatively similar results without extensive loss of significance, and we can thus confirm further robustness of the main empirical findings by Nunn and Wantchekon (2011).

7.2. **Pre-Colonial Institutions and Regional Developments in Africa.** Michalopoulos and Papaioannou (2013) analyze the effects of pre-colonial institutions on contemporary regional developments in Africa, controlling for various population, locational and geographic covariates including population density, distance to capital, distance to sea coast, distance to border, water area, land area, elevation, land suitable for agriculture, ecological suitability, petrolium, and diamond mine as well as country fixed effects in their baseline model. Estimates of these effects are obtained by running regressions with a sample that pools n_{ij} populated pixels $\ell \in [n_{ij}]$ in ethnic group i and country j across the cells $(i,j) \in [N] \times [M]$ of $N(=93)$ ethnic groups and $M(=48)$ countries. Standard errors are computed by the two-way cluster-robust method of Cameron, Gelbach and Miller (2011) for the ethnic group and district as two ways of clustering.

With our proposed method that is applicable to both high-dimensional models and multi-way clustering, they could consider even more flexible model specifications, for example, allowing for interactions of all combinations of geographical covariates and locational covariates. We present estimates with standard errors under such extended models with flexible specifications, demonstrate that qualitatively similar results continue to be obtained without substantial loss of statistical significance, and thus confirm further robustness of the main empirical findings by Michalopoulos and Papaioannou (2013).

Consider the model

$$Y_{ij,\ell} = D_{ij,\ell}\alpha + X'_{ij,\ell}\beta + R^Y_{ij,\ell} + \varepsilon_{ij,\ell},\ \mathrm{E}[\varepsilon_{ij,\ell}|D_{ij,\ell}, X_{ij,\ell}] = 0,$$

where $Y_{ij,\ell}$ denotes a regional development measured by night light intensity, $D_{ij,\ell}$ denotes an intensity measure of pre-colonial ethinic institutions, $X_{ij,\ell}$ contains population density, interactions of all combinations of locational controls (distance to capital, distance to sea coast, and distance to border), interactions of all combinations of geographical controls (water area, land area, elevation, land suitable for agriculture, ecological suitability, petrolium, and diamond mine), and country fixed effects, consisting of 82 or 83 dimensions of covariates in total. Note that the total number of regressors ($p+1 = 83$ or 84) is much larger than the effective sample size ($\underline{C} = N \wedge M = 48$) of two-way clustering in this extended setting.

Table 3 summarize the estimates of the effects of pre-colonial institutions on regional development as measured by the "light density," corresponding to parts of Table 3 of Michalopoulos and Papaioannou (2013). The last two columns in the table show the original estimates obtained under the prototypical model by Michalopoulos and Papaioannou (2013, Table 3) and corresponding lasso estimates obtained under more flexible model specification by our method. Across all the measures of pre-colonial institutions and all specifications, the original estimates and our lasso estimates are similar with similar levels of statistical significance. These results demonstrate that, even for flexible model specifications entailing high-dimensional covariates, the proposed method allows to produce qualitatively similar results without extensive loss of significance, and we can thus confirm further robustness of the main empirical findings by Michalopoulos and Papaioannou (2013).

8. Conclusion

In this paper, we investigate high-dimensional regression models when data is sampled under multi-way clustering. We establish the convergence rates for the lasso and post-lasso estimators under multi-way clustering. We then propose an inference method based on a post-double-selection procedure and show that it is asymptotically valid under multi-way clustering. Simulation studies demonstrate that the proposed procedure works well in finite sample under multi-way clustering in comparison with existing alternatives. We demonstrate that our method can enrich the flexibility of regression models and robustness of empirical results through a couple of empirical applications in growth and development economics.

Indeed, both multi-way clustering and high dimensionality are two important issues which concern applied research. The existing literature provide solutions to each of multi-way clustering and high-dimensionality separately. To our best knowledge, the literature does not seem to provide a solution to both of these issues simultaneously. In this paper, we filled this void in the literature.

Appendix A. Mathematical Proofs

Throughout, the symbol $=$ stands for $\overset{a.s.}{=}$. We use the notations $Y = [Y_{11}, ..., Y_{NM}]'$, $X = [X_{11}, ..., X_{NM}]'$, $D = [D_{11}, ..., D_{NM}]'$, $\mathcal{E} = [\varepsilon_{11}, ..., \varepsilon_{NM}]'$, $V = [v_{11}, ..., v_{NM}]'$, $R^Y = [R^Y_{11}, ..., R^Y_{NM}]'$, $R^D = [R^D_{11}, ..., R^D_{NM}]'$, $g = X\beta + R^Y$, and $m = X\gamma + R^D$. For any $A \subset [p]$, let $X_A = \{X_j : j \in A\}$, where X_j denotes the j-th the columns of X. Also define the projection operator by

$$\mathcal{P}_A = X_A(X_A'X_A)^- X_A',$$

and the orthogonal projection operator by $\mathcal{M}_A = I - \mathcal{P}_A$.

A.1. **Proof of Lemma 1.**

Proof. Our proof strategy closely follows that of Lemma D.2 in Davezies, D'Haultfoeuille and Guyonvarch (2018), except that we care about allowing the data generating process to vary with n to accommodate the increasing dimensionality p.

Under Assumption 1 (1) and (2), Lemma C.1 (a version of Aldous-Hoover representation) of Davezies, D'Haultfoeuille and Guyonvarch (2018) implies that, for each n, there exists a measurable function τ_n such that

$$\{Z_{ij}\}_{(i,j)\in\mathbb{N}^2} = \{\tau_n(U_{i0}, U_{0j}, U_{ij})\}_{(i,j)\in\mathbb{N}^2} \tag{A.6}$$

holds, where $\left\{\{U_{i0}\}_{i\in\mathbb{N}}, \{U_{0j}\}_{j\in\mathbb{N}}, \{U_{ij}\}_{(i,j)\in\mathbb{N}^2}\right\}$ are i.i.d. uniform$(0,1)$ random variables.

The Hájek projection $H_n f$ of $\mathbb{G}_C f$ on the set $\mathcal{G}_n$ is characterized by

$$\mathrm{E}\Big[(\mathbb{G}_C f - H_n f)\cdot g(U_n)\Big] = 0 \quad \text{for any } g(U_n) \in \mathcal{G}_n,$$

where $U_n = (U_{i0}, U_{0j})_{i\in[N], j\in[M]}$. Thus, for any U_c with $c = (c_1, c_2) \in \mathcal{I}_n = \{(i,0), (0,j) : i \in [N], j \in [M]\}$, we have

$$\mathrm{E}[\mathbb{G}_C f | U_c] = \mathrm{E}[H_n f | U_c].$$

Because the range of H_n is a closed subspace, we have

$$H_n f = \sum_{i=1}^{N} \mathrm{E}[H_n f | U_{i0}] + \sum_{j=1}^{M} \mathrm{E}[H_n f | U_{0j}].$$

It follows from the above two equations that

$$H_n f = \sum_{i=1}^{N} \mathrm{E}[\mathbb{G}_C f | U_{i0}] + \sum_{j=1}^{M} \mathrm{E}[\mathbb{G}_C f | U_{0j}].$$

Now, fix $c \in \mathcal{I}_n$ and let $e(c) = (\mathbb{1}\{c_1 > 0\}, \mathbb{1}\{c_2 > 0\})$. By the independence of $\big\{\{U_{i0}\}_{i\in\mathbb{N}}, \{U_{0j}\}_{j\in\mathbb{N}}, \{U_{ij}\}_{(i,j)\in\mathbb{N}^2}\big\}$, Z_{ij} and U_c are independent whenever $c \neq (i,j) \odot e(c)$, where $\odot$ denotes the Hadamard product. Thus, $\mathrm{E}\Big[f(Z_{ij}) - \mathrm{E}f(Z_{11})\Big|U_c\Big] = \mathrm{E}\Big[f(Z_{ij}) - \mathrm{E}f(Z_{11})\Big] = 0$. Therefore,

$$\begin{aligned}
\mathrm{E}[\mathbb{G}_C f | U_c] =& \frac{\sqrt{\underline{C}}}{NM} \sum_{i=1}^{N}\sum_{j=1}^{M} \mathrm{E}\Big[f(Z_{ij}) - \mathrm{E}f(Z_{11})\Big|U_c\Big] \\
=& \frac{\sqrt{\underline{C}}}{NM} \sum_{i=1}^{N}\sum_{j=1}^{M} \mathbb{1}\{(i,j) \odot e(c) = c\} \mathrm{E}\Big[f(Z_{ij}) - \mathrm{E}f(Z_{11})\Big|U_c\Big].
\end{aligned}$$

The representation (A.6) implies, for all (i,j) such that $(i,j) \odot e(c) = c$,

$$\mathrm{E}\Big[f(Z_{ij}) - \mathrm{E}f(Z_{11})\Big|U_c\Big] = \mathrm{E}\Big[f(Z_{c\vee\mathbf{1}}) - \mathrm{E}f(Z_{11})\Big|U_c\Big],$$

i.e. the index outside the support of c can be changed to 1. Now, suppose $c_k = 0$, $k \in \{1,2\}$. The representation (A.6) again gives

$$\frac{\sqrt{\underline{C}}}{NM} \sum_{i=1}^{N}\sum_{j=1}^{M} \mathbb{1}\{(i,j) \odot e(c) = c\} \mathrm{E}\Big[f(Z_{c\vee\mathbf{1}}) - \mathrm{E}f(Z_{11})\Big|U_c\Big] = \frac{\sqrt{\underline{C}} C_k}{NM} \mathrm{E}\Big[f(Z_{c\vee\mathbf{1}}) - \mathrm{E}f(Z_{11})\Big|U_c\Big],$$

where $C_1 = N$, $C_2 = M$. Therefore,

$$\begin{aligned}
H_n f =& \sum_{c\in\mathcal{I}_n} \frac{\sqrt{\underline{C}} C_k}{NM} \mathrm{E}\Big[f(Z_{c\vee\mathbf{1}}) - \mathrm{E}f(Z_{11})\Big|U_c\Big] \\
=& \sum_{i=1}^{N} \frac{\sqrt{\underline{C}}}{N} \mathrm{E}\Big[f(Z_{i1}) - \mathrm{E}f(Z_{11})\Big|U_{i0}\Big] + \sum_{j=1}^{M} \frac{\sqrt{\underline{C}}}{M} \mathrm{E}\Big[f(Z_{1j}) - \mathrm{E}f(Z_{11})\Big|U_{0j}\Big],
\end{aligned}$$

and each term in the two summands is independent from the others. This establishes the first claim of the lemma.

We next show that the variance of $H_n f$ can be calculated as

$$\begin{aligned}
V(H_n f) =& \mu_N V(\mathrm{E}[f(Z_{11})|U_{10}]) + \mu_M V(\mathrm{E}[f(Z_{11})|U_{01}]) \\
=& \mu_N Cov(f(Z_{11}), f(Z_{12})) + \mu_M Cov(f(Z_{11}), f(Z_{21})).
\end{aligned}$$

To see this, note that

$$\begin{aligned}V(\mathrm{E}[f(Z_{11})|U_{10}]) =& Cov(\mathrm{E}[f(Z_{11})|U_{10}], \mathrm{E}[f(Z_{12})|U_{10}]) \\ =& Cov(f(Z_{11}), f(Z_{12})) - \mathrm{E}[Cov(f(Z_{11}), f(Z_{12})|U_{10})] \\ =& Cov(f(Z_{11}), f(Z_{12})) - \mathrm{E}[Cov\{f(\,\tau_n(U_{10}, U_{01}, U_{11})\,), f(\,\tau_n(U_{10}, U_{02}, U_{12})\,)|U_{10}\}] \\ =& Cov(f(Z_{11}), f(Z_{12})) - 0,\end{aligned}$$

where the first equality follows from the representation (A.6), the second from the law of total covariance, the third from the representation (A.6), and the last from the fact that $\big\{\{U_{i0}\}_{i\in\mathbb{N}}, \{U_{0j}\}_{j\in\mathbb{N}}, \{U_{ij}\}_{(i,j)\in\mathbb{N}^2}\big\}$ are independent. Analogous lines of calculations yield $V(\mathrm{E}[f(X_{11})|U_{01}]) = Cov(f(X_{11}), f(X_{21}))$. Also, a direct calculation using Assumption 1 (1) and (2) shows

$$\begin{aligned}V(\mathbb{G}_C f) =& \frac{\underline{C}}{(NM)^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M} Cov(f(Z_{ij}), f(Z_{ij'})) + \frac{\underline{C}}{(NM)^2}\sum_{1\le i,i'\le N}\sum_{j=1}^{M} Cov(f(Z_{ij}), f(Z_{i'j})) \\ & - \frac{\underline{C}}{(NM)^2}\sum_{i=1}^{N}\sum_{j=1}^{M} V(f(Z_{ij})) \\ =& \mu_N Cov(f(Z_{11}), f(Z_{12})) + \mu_M Cov(f(Z_{11}), f(Z_{21})) + O\Big(\frac{1}{\underline{C}}\Big),\end{aligned}$$

since $\underline{C}/NM \le 1/\underline{C}$ and $\mathrm{E}f^2$ is bounded over n. This establishes the second claim of the lemma. ∎

A.2. **Proof of Theorem 1.**

Proof. We will focusing on the result for $\widetilde{\gamma}$ since the results for $\widetilde{\eta}$ will follow analogously. The proof is divided into four steps. The conclusions from the first two steps give the rates for lasso. The third step provides bounds for the rates of the post-lasso in terms of convergence rates of lasso. The fourth step provides the ℓ_2-norm rate.

Step 1. The oracle inequality follows directly from Lemma 6 of Belloni, Chen, Chernozhukov and Hansen (2012), which is applicable under Assumption 3 (1)–(2). This implies that we have the following bounds for lasso estimator:

$$\|\widehat{\gamma}-\gamma\|_1 \lesssim \frac{\sqrt{s}\lambda_2}{NM}+\frac{NMc_s^2}{\lambda_2} \quad \text{and}$$
$$\|X_{ij}'(\widehat{\eta}-\eta)\|_n \lesssim \frac{s\lambda_2}{NM}+c_s,$$

conditionally on the regularized event $\lambda_2/NM \geq c\|(NM)^{-1}\sum_{i=1}^N\sum_{j=1}^M X_{ij}v_{ij}\|_\infty$ for some constant $c>1$. Apply Assumption 3 (2) and use the choice of λ_2 to obtain the desired results in the first line.

Step 2. We now claim that, if we set $\lambda_2 = O\Big(\sqrt{(NM)^2\log a/\underline{C}}\Big)$, then the regularized event

$$\max_{k\in[p]}\Big|\frac{1}{NM}\sum_{i=1}^N\sum_{j=1}^M X_{ij,k}v_{ij}-\mathrm{E}[X_{11,k}v_{11}]\Big| \lesssim \frac{1}{c}\sqrt{\frac{\log a}{\underline{C}}}=\frac{\lambda_2}{NM}. \tag{A.7}$$

realizes with probability at least $1-C(\log\underline{C})^{-1}$.

First notice that the left-hand side can be bounded as

$$\begin{aligned}
&\max_{k\in[p]}\Big|\frac{1}{NM}\sum_{i=1}^N\sum_{j=1}^M X_{ij,k}v_{ij}-\mathrm{E}[X_{11,k}v_{11}]\Big| \\
\leq&\max_{k\in[p]}\Big|\frac{1}{NM}\sum_{i=1}^N\sum_{j=1}^M X_{ij,k}v_{ij}-\frac{M}{NM}\sum_{i=1}^N\mathrm{E}[X_{i1,k}v_{i1}|U_{i0}]-\frac{N}{NM}\sum_{j=1}^M\mathrm{E}[X_{1j,k}v_{1j}|U_{0j}]\Big| \\
&+\max_{k\in[p]}\Big|\frac{M}{NM}\sum_{i=1}^N\mathrm{E}[X_{i1,k}v_{i1}|U_{i0}]-\mathrm{E}[X_{11,k}v_{11}]\Big|+\max_{k\in[p]}\Big|\frac{N}{NM}\sum_{j=1}^M\mathrm{E}[X_{1j,k}v_{1j}|U_{0j}]-\mathrm{E}[X_{11,k}v_{11}]\Big| \\
=&(1)+(2)+(3)
\end{aligned}$$

where $E[X_{11,k}v_{11}]=0$ is used. Part (1) is $O_{\mathrm{P}}(\frac{1}{\underline{C}})$ by Lemma 1 under Assumptions 1 and 2 (1). Using Lemma 2 (see Appendix B ahead), we can show (2) $=O_{\mathrm{P}}(\sqrt{\frac{\log a}{N}})$. To see this, note that Assumption 2 (1) implies that $\sigma^2 := \max_{k\in[p]}\frac{1}{N}\sum_{i=1}^N\mathrm{E}(\mathrm{E}[X_{i1}v_{i1}|U_{i0}])^2$ is

uniformly bounded, and Assumption 2 (1)–(2) suggests

$$\begin{aligned}
B^2 =&:\mathrm{E}[\max_{i\in[N]}\max_{k\in[p]}(\mathrm{E}[X_{i1.k}v_{i1}|U_{i0}])^2] \\
\leq&\mathrm{E}[\max_{i\in[N]}\|X_{i1}v_{i1}\|_\infty^2] \\
\leq&\Big(\mathrm{E}[\max_{i\in[N]}\|X_{i1}v_{i1}\|_\infty^q]\Big)^{2/q} \\
\leq&N^{2/q}\Big(\mathrm{E}[\|X_{11}\|_\infty^q|v_{11}|^q]\Big)^{2/q} \\
\leq&N^{2/q}\Big(\sqrt{\mathrm{E}[\|X_{11}\|_\infty^{2q}]}\sqrt{\mathrm{E}|v_{11}|^{2q}]}\Big)^{2/q} = N^{2/q}B_n^2O(1),
\end{aligned}$$

where the first inequality follows from the property of projection, the second inequality follows from Jensen's inequality, and the fourth inequality follows from Cauchy-Schwartz inequality. Thus, Lemma 2 (see Appendix B ahead) implies

$$|(2)| \lesssim_{\mathrm{P}} \sqrt{\frac{\log a}{N}} + \frac{B_n \log a}{N^{1-1/q}} \lesssim \sqrt{\frac{\log a}{N}}.$$

Finally, $(3) = O_{\mathrm{P}}(\sqrt{\frac{\log a}{M}})$ follows analogously. Therefore, $(2)+(3) = O_{\mathrm{P}}(\sqrt{\frac{\log a}{\underline{C}}})$.

Step 3. We now derive bounds for performance of post-lasso:

$$\|m_{ij} - X_{ij}\widetilde{\gamma}\|_n \lesssim_{\mathrm{P}} \sqrt{\frac{s\log a}{\underline{C}}} + \frac{\|(I-\mathcal{P}_{\widehat{I}_2})m\|}{\sqrt{NM}} \tag{A.8}$$

where $m_{ij} = X_{ij}'\gamma + R_{ij}^D$. This part of proof closely follows the proof of Lemma 7 in Belloni, Chen, Chernozhukov and Hansen (2012) with some minor modifications. First note that $m - X_{\widehat{I}_2}\widetilde{\gamma} = (I-\mathcal{P}_{\widehat{I}_2})m - \mathcal{P}_{\widehat{I}_2}V$. This implies

$$\|m - X_{\widehat{I}_2}\widetilde{\gamma}\| \leq \|(I-\mathcal{P}_{\widehat{I}_2})m\| + \|\mathcal{P}_{\widehat{I}_2}V\|.$$

By the definition of $\mathcal{P}_{\widehat{I}_2}$ and the operator norm,

$$\|\mathcal{P}_{\widehat{I}_2}V\| \leq \|X_{\widehat{I}_2}/\sqrt{NM}(X_{\widehat{I}_2}'X_{\widehat{I}_2}/NM)^{-1}\|\,\|X_{\widehat{I}_2}'V/\sqrt{MM}\|$$

and

$$\|X_{\widehat{I}_2}/\sqrt{NM}(X_{\widehat{I}_2}'X_{\widehat{I}_2}/NM)^{-1}\| \leq \sqrt{1/\phi_{\min}(s+\widetilde{m}_2)},$$

where $\widetilde{m}_2 = |\widehat{I}_2 \setminus T_2|$, $T_2 = \mathrm{support}(\gamma)$. Thus under Assumption 4, we obtain

$$\|\mathcal{P}_{\widehat{I}_2}V\| \leq \sqrt{1/\phi_{\min}(s+\widetilde{m}_2)}\,\|X_{\widehat{I}_2}'V/\sqrt{MM}\| \leq \sqrt{\frac{s+\widetilde{m}_2}{\phi_{\min}(s+\widetilde{m}_2)}}\,\|XV/\sqrt{NM}\|_\infty \lesssim_{\mathrm{P}} \sqrt{\frac{s\log a}{\underline{C}}},$$

where the last inequality follows from equation (A.7) and Lemma 3 (see Appendix B ahead). This shows (A.8).

By (A.7), the same argument as that of Lemma 7 in Belloni, Chen, Chernozhukov and Hansen (2012) establishes

$$\frac{\|(I - \mathcal{P}_{\widehat{I}_2})m\|}{\sqrt{NM}} \lesssim \frac{\sqrt{s}\lambda_2}{NM} + c_s.$$

Therefore, (A.8) can be rewritten as

$$\|m_{ij} - X_{ij}\widetilde{\gamma}\|_n \lesssim_{\mathrm{P}} \sqrt{\frac{s \log a}{\underline{C}}} + \frac{\sqrt{s}\lambda_2}{NM} + c_s.$$

Next, applying Lemma 3 (see Appendix B ahead), we have

$$\|\widetilde{\gamma} - \gamma\|_1 \leq \sqrt{\|\widetilde{\gamma} - \gamma\|_0}\, \|\widetilde{\gamma} - \gamma\| \leq \sqrt{s + \widetilde{m}_2}\, \frac{\|X_{ij}(\widetilde{\gamma} - \gamma)\|_n}{\sqrt{\phi_{\min}(s + \widetilde{m}_2)}}$$

under Assumption 3 (1) and 4.

Combining the above bounds, the choice of λ_2, and Assumption 3 (2), we obtain

$$\|\widetilde{\gamma} - \gamma\|_1 \lesssim_{\mathrm{P}} \sqrt{\frac{s^2 \log a}{\underline{C}}} \qquad \text{and}$$

$$\|X_{ij}(\widetilde{\gamma} - \gamma)\|_n \lesssim_{\mathrm{P}} \sqrt{\frac{s \log a}{\underline{C}}}.$$

Step 4. The ℓ_2-norm rates are implied by the prediction norm rates, Assumption 4, and Lemma 3 (see Appendix B ahead). ∎

A.3. **Proof of Theorem 2.**

Proof. Our proof follows parallel steps to Steps 1-6 in the proof of Theorem 1 in Belloni, Chernozhukov and Hans (2014). However, due to the two-way cluster sampling, most of the probabilistic bounds are established differently.

We use the short-hand notation

$$\widetilde{b}_Z(A) := \underset{b \in \mathbb{R}^p : b_j = 0 \forall j \in A^c}{\operatorname{argmin}} \|Z - X'b\|^2$$

for any vector $Z \in \mathbb{R}^n$.

Step 1 Write $\widetilde{\alpha} = [D'\mathcal{M}_{\widehat{I}}D/NM]^{-1}D'\mathcal{M}_{\widehat{I}}Y/NM$ and thus we can write

$$\sqrt{\underline{C}}(\widetilde{\alpha} - \alpha) = [D'\mathcal{M}_{\widehat{I}}D/NM]^{-1} \cdot \sqrt{\underline{C}}D'\mathcal{M}_{\widehat{I}}(g + \mathcal{E})/NM =: (II)^{-1} \cdot (I).$$

By Steps 2 and 3 to be presented below, we obtain

$$(II) = V'V/NM + o_{\mathrm{P}}(1) \text{ and } (I) = \sqrt{\underline{C}}V'\mathcal{E}/NM + o_{\mathrm{P}}(1).$$

Also note that $V'V/NM = \mathrm{E}[v_{11}^2] + o_{\mathrm{P}}(1)$ by Lemma 1 and Assumption 1–2, which can be shown following the same arguments as those in Step 3 of the proof for Theorem 3. Under Assumption 2 (1), $\mathrm{E}[v_{11}^2]$ is bounded and bounded away from zero uniformly in n. Therefore $(II)^{-1} = \mathrm{E}[v_{11}^2]^{-1} + o_{\mathrm{P}}(1)$.

Under Assumption 2 (3), σ^2 is bounded and bounded away from zero. Setting $W_{ij} := \sigma^{-1}v_{ij}\varepsilon_{ij}$ and $Z_{ij} \overset{f}{\mapsto} W_{ij}$, we have $\mathrm{E}f(Z_{11}) = 0$ and

$$\mathbb{G}_C f = \frac{\sqrt{C}}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} W_{ij} = \sigma^{-1}\sqrt{\underline{C}}(\widetilde{\alpha} - \alpha) + o_{\mathrm{P}}(1).$$

Apply Lemma 1 under Assumption 1 and 2 (1) to obtain the Hájek projection

$$H_n f = \sum_{i=1}^{N}\frac{\sqrt{\underline{C}}}{N}\mathrm{E}[f(Z_{i1})|U_{i0}] + \sum_{j=1}^{M}\frac{\sqrt{\underline{C}}}{M}\mathrm{E}[f(Z_{1j})|U_{0j}]$$

of $\mathbb{G}_C f$, where terms in each summand are independent and two summands are independent of each other. We now check Lyapunov's conditions. First, note that Assumption 2 (1) guarantees that the third moments of both summands are bounded uniformly in n. Second, the second part of Lemma 1 implies that

$$\begin{aligned}\lim_{n\to\infty} V(H_n f) &= \bar{\mu}_N V(\mathrm{E}[f(Z_{11})|U_{10}]) + \bar{\mu}_M V(\mathrm{E}[f(Z_{11})|U_{01}]) \\ &= \bar{\mu}_N \mathrm{E}[f(Z_{11})f(Z_{12})] + \bar{\mu}_M \mathrm{E}[f(Z_{11})f(Z_{21})] = \Gamma \in (c, \infty)\end{aligned}$$

a.s. for $c > 0$, where the last inequalities follow from Assumption 2 (3). Therefore, we apply Lyapunov's CLT to obtain

$$H_n f \rightsquigarrow N\Big(0, \bar{\mu}_N V(\mathrm{E}[f(Z_{11})|U_{10}]) + \bar{\mu}_M V(\mathrm{E}[f(Z_{11})|U_{01}])\Big).$$

The first equality in the variance equation of Lemma 1 yields

$$V(\mathbb{G}_C f) = \bar{\mu}_N \mathrm{E}[f(Z_{11})f(Z_{12})] + \bar{\mu}_M \mathrm{E}[f(Z_{11})f(Z_{21})] + o_{\mathrm{P}}(1),$$

where the right-hand side is asymptotically positive and bounded away from zero. Therefore,

$$\sqrt{\underline{C}}(\widetilde{\alpha}-\alpha)=\mathbb{G}_C f \rightsquigarrow N(0,\sigma^2).$$

Step 2 Use $D=m+V$ to decompose

$$\begin{aligned}(I)=&\sqrt{\underline{C}}V'\mathcal{E}/NM+\sqrt{\underline{C}}m'\mathcal{M}_{\widehat{I}}g/NM+\sqrt{\underline{C}}m'\mathcal{M}_{\widehat{I}}\mathcal{E}/NM+\sqrt{\underline{C}}V'\mathcal{M}_{\widehat{I}}g/NM-\sqrt{\underline{C}}V'\mathcal{P}_{\widehat{I}}\mathcal{E}/NM\\ =&\sqrt{\underline{C}}V'\mathcal{E}/NM+(1a)+(1b)+(1c)-(1d).\end{aligned}$$

By Steps 5 and 6 to be presented below, we have

$$|(1a)|\lesssim\sqrt{\underline{C}}\|\mathcal{M}_{\widehat{I}}m/\sqrt{NM}\|\cdot\|\mathcal{M}_{\widehat{I}}g/\sqrt{NM}\|\lesssim_{\mathrm{P}}\sqrt{\frac{s^2(\log a)^2}{\underline{C}}}.$$

Using the decompositions $m=X\gamma+R^D$, $m'\mathcal{P}_{\widehat{I}}=\widetilde{b}_m'(\widehat{I})X'$ and $m'\mathcal{M}_{\widehat{I}}\mathcal{E}=(R^D)'\mathcal{E}-(\widetilde{b}_m(\widehat{I})-\gamma)'X'\mathcal{E}$, one has

$$|(1b)|\le\sqrt{\underline{C}}|R^{D\prime}\mathcal{E}/NM|+\sqrt{\underline{C}}|(\widetilde{b}_m(\widehat{I})-\gamma)X'\mathcal{E}/NM|\lesssim_{\mathrm{P}}\sqrt{\frac{s^2(\log a)^2}{\underline{C}}},$$

because, under Assumptions 2 (1) and 3 (2),

$$\sqrt{\underline{C}}|R^{D\prime}\mathcal{E}/NM|\le\sqrt{\underline{C}}\sqrt{R^{D\prime}R^D/NM}\cdot O_{\mathrm{P}}\Big(\sqrt{\frac{1}{NM}\mathrm{E}\|\mathcal{E}\|^2}\Big)\lesssim_{\mathrm{P}}\sqrt{\frac{s}{\underline{C}}}$$

and

$$\begin{aligned}\sqrt{\underline{C}}|(\widetilde{b}_m(\widehat{I})-\gamma)X'\mathcal{E}/NM|&\le\sqrt{\underline{C}}\|\widetilde{b}_m(\widehat{I})-\gamma\|_1\|X'\mathcal{E}/NM\|_\infty\\ &\lesssim_{\mathrm{P}}\sqrt{\underline{C}}\sqrt{\frac{s^2\log a}{\underline{C}}}\cdot\sqrt{\frac{\log a}{\underline{C}}}=\sqrt{\frac{s^2(\log a)^2}{\underline{C}}},\end{aligned}$$

where $\|\widetilde{b}_m(\widehat{I})-\gamma\|_1\lesssim_{\mathrm{P}}\sqrt{\frac{s^2\log a}{\underline{C}}}$ follows from Step 5 and $\|X'\mathcal{E}/NM\|_\infty\lesssim_{\mathrm{P}}\sqrt{\frac{\log a}{\underline{C}}}$ follows from Step 4. Third, using the same argument as above for (1b) following Steps 4 and 6 and $g=X\beta+R^Y$, we have

$$|(1c)|\le\sqrt{C}|R^{Y\prime}V/\sqrt{NM}|+\sqrt{\underline{C}}|(\widetilde{b}_g(\widehat{I})-\beta)'X'V/\sqrt{NM}|\lesssim_{\mathrm{P}}\sqrt{\frac{s}{\underline{C}}}+\sqrt{\frac{s^2(\log a)^2}{\underline{C}}}.$$

Finally,

$$|(1d)|\le\sqrt{\underline{C}}|\widetilde{b}_V(\widehat{I})'X'\mathcal{E}/NM|\le\sqrt{\underline{C}}\|\widetilde{b}_V(\widehat{I})\|_1\|X'\mathcal{E}/NM\|_\infty\lesssim_{\mathrm{P}}\sqrt{\frac{s^2(\log a)^2}{\underline{C}}}$$

following equation (A.7) in the proof of Theorem 1, and

$$\|\widetilde{b}_V(\widehat{I})\|_1 \leq \sqrt{\widehat{s}}\|\widetilde{b}_V(\widehat{I})\| \leq \sqrt{\widehat{s}}\|(X_{\widehat{I}}'X_{\widehat{I}}/NM)^{-1}X_{\widehat{I}}'V/NM\|$$
$$\lesssim_{\mathrm{P}} \frac{\sqrt{\widehat{s}}}{\phi_{\min}(C\widehat{s})}\sqrt{\widehat{s}}\|X_{\widehat{I}}'V/NM\|_\infty \lesssim_{\mathrm{P}} \sqrt{\frac{s^2\log a}{\underline{C}}}$$

following Step 4, Lemma 3, and Assumption 4.

Step 3 We can write

$$\begin{aligned}(II) =&(m+V)'\mathcal{M}_{\widehat{I}}(m+V)/NM\\ =&V'V/NM + m'\mathcal{M}_{\widehat{I}}m/NM + 2m'\mathcal{M}_{\widehat{I}}V/NM - V'\mathcal{P}_{\widehat{I}}V/NM\\ =&V'V/NM + (2a) + (2b) - (2c).\end{aligned}$$

We have $|(2a)| \lesssim_{\mathrm{P}} \sqrt{\frac{s^2(\log a)^2}{\underline{C}}}$ by Step 5, $|(2b)| \lesssim_{\mathrm{P}} \sqrt{\frac{s^2(\log a)^2}{\underline{C}}}$ by a similar argument to bounding $|(1b)|$, and $|(2c)| \lesssim_{\mathrm{P}} \sqrt{\frac{s^2(\log a)^2}{\underline{C}}}$ by a similar argument to bounding $|(1d)|$.

Step 4 In this step, we show that the following regularized events hold with probability $1-o(1)$:

$$(a)\ \sqrt{\underline{C}}\|X'\mathcal{E}/NM\|_\infty \lesssim_{\mathrm{P}} \sqrt{\log a} \text{ and } (b)\ \sqrt{\underline{C}}\|X'V/NM\|_\infty \lesssim_{\mathrm{P}} \sqrt{\log a}.$$

This claim follows from similar lines of argument to those showing equation (A.7) in the proof of Theorem 1 under Assumptions 1 and 2 (1)–(2).

Step 5 In this step, we show

$$(a)\ \|\mathcal{M}_{\widehat{I}}m/\sqrt{NM}\| \lesssim_{\mathrm{P}} \sqrt{\frac{s\log a}{\underline{C}}} \text{ and } (b)\ \|\widetilde{b}_m(\widehat{I}) - \gamma\| \lesssim_{\mathrm{P}} \sqrt{\frac{s\log a}{\underline{C}}}.$$

First, by applying Theorem 1 under Assumptions 1, 2 (1)–(2), 3(1)–(2), and 4, and by following the same argument as the one in Step 5 of Belloni, Chernozhukov and Hansen (2014), we have

$$\|\mathcal{M}_{\widehat{I}}m/\sqrt{NM}\| \leq \|\mathcal{M}_{\widehat{I}_2}m/\sqrt{NM}\| \leq \|(X\widetilde{b}_D(\widehat{I}_2) - m)/\sqrt{NM}\| \lesssim_{\mathrm{P}} \sqrt{\frac{s\log a}{\underline{C}}},$$

where the first inequality follows from $\widehat{I}_2 \subset \widehat{I}$ and the second follows from the fact that $\widetilde{b}_m(\widehat{I}_2)$ minimizes $\|m - X_{\widehat{I}_2}'b\|$. Second, the reverse triangle inequality yields

$$\left|\|X(\widetilde{b}_m(\widehat{I}) - \gamma)/\sqrt{NM}\| - \|R_m/\sqrt{NM}\|\right| \lesssim_{\mathrm{P}} \|\mathcal{M}_{\widehat{I}}m/\sqrt{NM}\|,$$

and, by Assumption 3 (2), $\|R_m/\sqrt{NM}\| \lesssim_{\mathrm{P}} \sqrt{s/\underline{C}}$. Thus, by using Lemma 3 with Assumptions 3 (1) and 4, we obtain

$$\begin{aligned}\|\widetilde{b}_m(\widehat{I}) - \gamma\| \lesssim_{\mathrm{P}} & \sqrt{\phi_{\min}(\widehat{s}+s)}\|\widetilde{b}_m(\widehat{I}) - \gamma\| \\ \leq & \|X(\widetilde{b}_m(\widehat{I}) - \gamma)/\sqrt{NM}\| \lesssim_{\mathrm{P}} \sqrt{\frac{s\log a}{\underline{C}}}.\end{aligned}$$

Step 6 Finally, we can show

$$(a)\ \sqrt{\underline{C}}\|\mathcal{M}_{\widehat{I}}g/NM\| \lesssim_{\mathrm{P}} \sqrt{\frac{s\log a}{\underline{C}}} \text{ and } (b)\ \|\widetilde{b}_g(\widehat{I}) - \beta\| \lesssim_{\mathrm{P}} \sqrt{\frac{s\log a}{\underline{C}}}.$$

following similar lines of argument to those of Step 5 under Assumptions 1, 2, 3, and 4. ■

A.4. **Proof of Theorem 3.**

Proof. First, note that we have the following decomposition

$$\begin{aligned}&|\widehat{Q}^{-1}\widehat{\Gamma}\widehat{Q}^{-1} - Q^{-1}\Gamma Q^{-1}| \\ \lesssim &|\widehat{Q}^{-1} - Q^{-1}||\widehat{Q}^{-1} + Q^{-1}||\widehat{\Gamma}| + |\widehat{\Gamma} - \Gamma||Q^{-1}|^2,\end{aligned}$$

where $|Q^{-1}|$ is bounded away from zero uniformly by Assumption 2(1). The rest of this proof is divide into 5 steps. In Steps 1 and 2, we obtain a bound for $|\widehat{\Gamma} - \Gamma|$. In Steps 3 and 4, we obtain a bound for $|\widehat{Q}^{-1} - Q^{-1}|$. Finally, Step 5 shows a bound for $|\widehat{\Gamma}|$ and $|\widehat{Q}^{-1} + Q^{-1}|$.

Step 1. We derive a bound for $|\widetilde{\Gamma} - \Gamma|$, where

$$\widetilde{\Gamma} = \frac{\mu_N}{NM^2}\sum_{i=1}^{N}\sum_{1\leq j,j'\leq M} v_{ij}\varepsilon_{ij}v_{ij'}\varepsilon_{ij'} + \frac{\mu_M}{N^2M}\sum_{1\leq i,i'\leq N}\sum_{j=1}^{M} v_{ij}\varepsilon_{ij}v_{i'j}\varepsilon_{i'j}.$$

We first claim that

$$\frac{\mu_N}{NM^2}\sum_{i=1}^{N}\sum_{1\leq j,j'\leq M} v_{ij}\varepsilon_{ij}v_{ij'}\varepsilon_{ij'} = \bar{\mu}_N\mathrm{E}[v_{11}\varepsilon_{11}v_{12}\varepsilon_{12}] + o_{\mathrm{P}}(1). \tag{A.9}$$

Note that, for each n, for any $i, \iota \in [N]$ and $j, k, l, m \in [M]$, we have

$$\begin{aligned}Cov\Big(v_{ij}\varepsilon_{ij}v_{ik}\varepsilon_{ik}, v_{\iota l}\varepsilon_{\iota l}v_{\iota m}\varepsilon_{\iota m}\Big) \leq & \max_{i\in[N],j,j'\in[M]} V(v_{ij}\varepsilon_{ij}v_{ij'}\varepsilon_{ij'}) \\ = & \max_{i\in[N],j,j'\in[M]}\Big\{\mathrm{E}[(v_{ij}\varepsilon_{ij}v_{ij'}\varepsilon_{ij'})^2] - (\mathrm{E}[v_{ij}\varepsilon_{ij}v_{ij'}\varepsilon_{ij'}])^2\Big\}.\end{aligned}$$

Using Cauchy-Schwartz's inequality with Assumptions 1 (1) and 2 (1), the first term in the variance can be bounded as

$$\begin{aligned}\mathrm{E}[(v_{ij}\varepsilon_{ij}v_{ij'}\varepsilon_{ij'})^2] \leq& \sqrt{\mathrm{E}[v_{ij}^4\varepsilon_{ij}^4]\mathrm{E}[v_{ij'}^4\varepsilon_{ij'}^4]} \\ \leq& \sqrt{\mathrm{E}v_{11}^8\mathrm{E}\varepsilon_{11}^8} = O(1)\end{aligned}$$

uniformly over n. Under Assumptions 1 (1) and 2 (1), similar calculations can be carried out to the square-root of the second term to obtain

$$\mathrm{E}[v_{ij}\varepsilon_{ij}v_{ij'}\varepsilon_{ij'}] \leq \sqrt{\mathrm{E}[v_{ij}^2\varepsilon_{ij}^2]\mathrm{E}[v_{ij'}^2\varepsilon_{ij'}^2]} \leq \sqrt{\mathrm{E}[v_{11}^4]\mathrm{E}[\varepsilon_{11}^4]} = O(1)$$

uniformly over n. This shows that, for any n, for any $i, \iota \in [N]$ and $j, k, l, m \in [M]$, it holds that, for a $K > 0$ independent of n,

$$\left|Cov\Big(v_{ij}\varepsilon_{ij}v_{ik}\varepsilon_{ik}, v_{\iota l}\varepsilon_{\iota l}v_{\iota m}\varepsilon_{\iota m}\Big)\right| \leq K. \tag{A.10}$$

With this bound of the covariance, we can bound the variance as

$$\begin{aligned}&V\Big(\frac{1}{NM^2}\sum_{i=1}^{N}\sum_{1\leq j,j'\leq M} v_{ij}\varepsilon_{ij}v_{ij'}\varepsilon_{ij'}\Big) \\ =&Cov\Big(\frac{1}{NM^2}\sum_{i=1}^{N}\sum_{1\leq j,j'\leq M} v_{ij}\varepsilon_{ij}v_{ij'}\varepsilon_{ij'}, \frac{1}{NM^2}\sum_{i=1}^{N}\sum_{1\leq j,j'\leq M} v_{ij}\varepsilon_{ij}v_{ij'}\varepsilon_{ij'}\Big) \\ =&\frac{1}{N^2M^4}\sum_{i=1}^{N}\sum_{1\leq j,k,l,m\leq M} Cov\Big(v_{ij}\varepsilon_{ij}v_{ik}\varepsilon_{ik}, v_{il}\varepsilon_{il}v_{im}\varepsilon_{im}\Big) \\ &+\frac{2}{N^2M^4}\sum_{j=1}^{M}\sum_{1\leq i,i'\leq N}\sum_{1\leq k,l\leq M} Cov\Big(v_{ij}\varepsilon_{ij}v_{ik}\varepsilon_{ik}, v_{i'j}\varepsilon_{i'j}v_{i'l}\varepsilon_{i'l}\Big) + o\Big(\frac{1}{\underline{C}}\Big) \\ =&O\Big(\frac{1}{\underline{C}}\Big) = o(1)\end{aligned}$$

uniformly over n, where the second equality follows from Assumption 1 (2) and counting the number of terms in each summand, and the third equality is due to (A.10). Applying Chebyshev's inequality, it follows that

$$\begin{aligned}&\mathrm{P}\Big(\Big|\frac{1}{NM^2}\sum_{i=1}^{N}\sum_{1\leq j,j'\leq M}(v_{ij}\varepsilon_{ij}v_{ij'}\varepsilon_{ij'} - \mathrm{E}[v_{ij}\varepsilon_{ij}v_{ij'}\varepsilon_{ij'}])\Big| > \epsilon\Big) \\ \leq&\frac{\sup_n V\Big(\frac{1}{NM^2}\sum_{i=1}^{N}\sum_{1\leq j,j'\leq M} v_{ij}\varepsilon_{ij}v_{ij'}\varepsilon_{ij'}\Big)}{\epsilon^2} = \frac{1}{\epsilon^2}\cdot o(1).\end{aligned} \tag{A.11}$$

for any $\epsilon > 0$.

Also, under Assumption 2 (1), the first result in Lemma D.10 of Davezies, D'Haultfoeuille and Guyonvarch (2018) ensures

$$\mathrm{E}\Big[\Big|\frac{\mu_N}{NM^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M} v_{ij}\varepsilon_{ij}v_{ij'}\varepsilon_{ij'} - \frac{\mu_N}{NM(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{j'\neq j} v_{ij}\varepsilon_{ij}v_{ij'}\varepsilon_{ij'}\Big|\Big] = o(1)$$

uniformly over n. Furthermore, under Assumptions 1 (1) and 2 (1), we have

$$\mathrm{E}\Big[\frac{1}{NM(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{j'\neq j} v_{ij}\varepsilon_{ij}v_{ij'}\varepsilon_{ij'}\Big] = \mathrm{E}[v_{11}\varepsilon_{11}v_{12}\varepsilon_{12}].$$

Combining these with (A.11), we obtain (A.9). A symmetric argument also shows

$$\frac{\mu_M}{N^2M}\sum_{1\le i,i'\le N}\sum_{j=1}^{M} v_{ij}\varepsilon_{ij}v_{i'j}\varepsilon_{i'j} = \bar{\mu}_M\mathrm{E}[v_{11}\varepsilon_{11}v_{21}\varepsilon_{21}] + o_{\mathrm{P}}(1).$$

Therefore, we obtain $|\widetilde{\Gamma} - \Gamma| = o_{\mathrm{P}}(1)$.

Step 2. In this step we bound $|\widehat{\Gamma} - \widetilde{\Gamma}|$, where $\widetilde{\Gamma}$ is defined in Step 1. Consider the decomposition

$$\underbrace{\frac{\mu_N}{NM^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}\Big(\widehat{v}_{ij}\widehat{\varepsilon}_{ij}\widehat{v}_{ij'}\widehat{\varepsilon}_{ij'} - v_{ij}\varepsilon_{ij}v_{ij'}\varepsilon_{ij'}\Big)}_{(1)} +$$

$$\underbrace{\frac{\mu_N}{N^2M}\sum_{1\le i,i'\le N}\sum_{j=1}^{M}\Big(\widehat{v}_{ij}\widehat{\varepsilon}_{ij}\widehat{v}_{i'j}\widehat{\varepsilon}_{i'j} - v_{ij}\varepsilon_{ij}v_{i'j}\varepsilon_{i'j}\Big)}_{(2)}.$$

Recall $\varepsilon_{ij} = Y_{ij} - Z_{ij}'\eta - R_{ij}^Y$, $\widehat{\varepsilon}_{ij} = Y_{ij} - Z_{ij}'\widetilde{\eta}$, $v_{ij} = D_{ij} - X_{ij}'\gamma - R_{ij}^D$, $\widehat{v}_{ij} = D_{ij} - X_{ij}'\widetilde{\gamma}$, and thus, $\widehat{\varepsilon}_{ij} - \varepsilon_{ij} = Z_{ij}'(\widetilde{\eta} - \eta) - R_{ij}^Y$ and $\widehat{v}_{ij} - v_{ij} = X_{ij}'(\widetilde{\gamma} - \gamma) - R_{ij}^D$. We can further decompose (1) as

$$\begin{aligned}(1) =& \frac{\mu_N}{NM^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}\Big((\widehat{v}_{ij} - v_{ij})\widehat{\varepsilon}_{ij}\widehat{v}_{ij'}\widehat{\varepsilon}_{ij'}\Big) + \frac{\mu_N}{NM^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}\Big(v_{ij}(\widehat{\varepsilon}_{ij} - \varepsilon_{ij})\widehat{v}_{ij'}\widehat{\varepsilon}_{ij'}\Big)\\ &+ \frac{\mu_N}{NM^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}\Big(v_{ij}\varepsilon_{ij}(\widehat{v}_{ij'} - v_{ij'})\widehat{\varepsilon}_{ij'}\Big) + \frac{\mu_N}{NM^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}\Big(v_{ij}\varepsilon_{ij}v_{ij'}(\widehat{\varepsilon}_{ij'} - \varepsilon_{ij'})\Big)\\ =& (1a) + (1b) + (1c) + (1d).\end{aligned}$$

Under Assumption 2 (1), we first bound

$$\begin{aligned}(1a) \lesssim_{\mathrm{P}} & \Big| \frac{\mu_N}{NM^2} \sum_{i=1}^{N} \sum_{1 \le j,j' \le M} \Big((\widehat{v}_{ij} - v_{ij}) \widehat{\varepsilon}_{ij} \widehat{v}_{ij'} \widehat{\varepsilon}_{ij'} \Big) \Big| \\ \le & \Big| \frac{\mu_N}{NM^2} \sum_{i=1}^{N} \sum_{1 \le j,j' \le M} X'_{ij} (\widehat{\gamma} - \gamma) \widehat{\varepsilon}_{ij} \widehat{v}_{ij'} \widehat{\varepsilon}_{ij'} \Big| + \Big| \frac{\mu_N}{NM^2} \sum_{i=1}^{N} \sum_{1 \le j,j' \le M} R^D_{ij} \widehat{\varepsilon}_{ij} \widehat{v}_{ij'} \widehat{\varepsilon}_{ij'} \Big| \\ = & (1aa) + (1ab).\end{aligned}$$

We obtain

$$\begin{aligned}(1aa) = & \Big| \frac{\mu_N}{NM^2} \sum_{i=1}^{N} \sum_{1 \le j,j' \le M} X'_{ij} (\widehat{\gamma} - \gamma) \widehat{\varepsilon}_{ij} \widehat{v}_{ij'} \widehat{\varepsilon}_{ij'} \Big| \\ \le & \Big| \frac{\mu_N}{NM^2} \sum_{i=1}^{N} \sum_{1 \le j,j' \le M} X'_{ij} (\widehat{\gamma} - \gamma) \cdot \Big(Z'_{ij} (\widehat{\eta} - \eta) + \varepsilon_{ij} - R^Y_{ij} \Big) \\ & \cdot \Big(X'_{ij'} (\widehat{\gamma} - \gamma) + v_{ij'} - R^D_{ij'} \Big) \cdot \Big(Z'_{ij'} (\widehat{\eta} - \eta) + \varepsilon_{ij'} - R^Y_{ij'} \Big) \Big| = o_{\mathrm{P}}(1),\end{aligned}$$

where the last equality follows from triangle inequality, Cauchy-Schwartz's inequality, Theorem 1, Assumptions 2 (1)–(2) and 3 (2), and the rate conditions in the statement of the theorem. To see this, note that, under Assumption 2 (2), Theorem 1, and the rate condition in the theorem, we have

$$\begin{aligned} & \sqrt{\frac{\underline{C}}{N^2M^2} \sum_{i=1}^{N} \sum_{1 \le j,j' \le M} \Big(X'_{ij} (\widehat{\gamma} - \gamma) Z'_{ij} (\widehat{\eta} - \eta) \Big)^2} \\ \le & \sqrt{\frac{\underline{C}}{NM} \max_{i \in [N], j \in [M]} \|Z_{ij}\|_\infty^2 \|\widehat{\eta} - \eta\|_1^2 M \|X'_{ij} (\widehat{\gamma} - \gamma)\|_n^2} \\ \lesssim_{\mathrm{P}} & \sqrt{\frac{(NM)^{1/q} B_n^2 s^3 (\log a)^2}{\underline{C} N}} = o(1).\end{aligned}$$

Furthermore, by Theorem 1 and Assumptions 2 (1) and 3 (2), we have

$$\begin{aligned}
&\sqrt{\frac{\underline{C}}{N^2M^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}\Big(X_{ij}'(\widehat{\gamma}-\gamma)\varepsilon_{ij}\Big)^2}\\
\le&\sqrt{\frac{\underline{C}}{NM}\max_{i\in[N],j\in[M]}|\varepsilon_{ij}|^2M\|X_{ij}'(\widehat{\gamma}-\gamma)\|_n^2}\\
\lesssim_{\mathrm{P}}&\sqrt{\frac{(NM)^{1/q}s\log a}{N}}=o(1).
\end{aligned}$$

The rest of the terms can be shown to be of smaller orders using similar arguments. Finally, the rate condition from the statement of the theorem gives

$$\sqrt{\frac{\underline{C}}{N^2M^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}\Big(R_{ij}^DR_{ij}^Y\Big)^2}\le\sqrt{\|R_{ij}^DR_{ij}^Y\|_n^2}=O(1).$$

Similarly, using Cauchy-Schwartz's inequality, Theorem 1, Assumptions 2 (1)–(2) and 3 (2), and the addition rate conditions in the statement of the theorem, we obtain

$$(1ab)=\Big|\frac{\mu_N}{NM^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}R_{ij}^D\widehat{\varepsilon}_{ij}\widehat{v}_{ij'}\widehat{\varepsilon}_{ij'}\Big|=o_{\mathrm{P}}(1).$$

These results yield $(1a)=o_{\mathrm{P}}(1)$. Following analogous but simpler arguments, we can show that $(1b)$, $(1c)$ and $(1d)$ are $o_{\mathrm{P}}(1)$. This shows $(1)=o_{\mathrm{P}}(1)$. Similar lines of argument under the same set of assumptions show $(2)=o_{\mathrm{P}}(1)$.

Step 3. In this and the next steps, we bound $|\widehat{Q}^{-1}-Q^{-1}|$. Note that

$$|\widehat{Q}-Q|=\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\widehat{v}_{ij}^2-\mathrm{E}[v_{11}^2]=\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(v_{ij}^2-\mathrm{E}[v_{11}^2])+\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(\widehat{v}_{ij}^2-v_{ij}^2).\tag{A.12}$$

The current step bounds the first term on the right-hand side, and Step 4 below bounds the second term on the right-hand side. With the notation $f(Z_{ij})=v_{ij}^2$, the first term on right-hand size becomes

$$\frac{1}{\sqrt{\underline{C}}}\mathbb{G}_Cf=\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(v_{ij}^2-\mathrm{E}[v_{11}^2]).$$

Applying Lemma 1 under Assumptions 1 and 2 (1) suggests that its Hájek projection equals

$$\begin{aligned}\frac{1}{\sqrt{\underline{C}}}H_n f &= \frac{1}{N}\sum_{i=1}^{N}\mathrm{E}[v_{i1}^2 - \mathrm{E}v_{11}^2|U_{i0}] + \frac{1}{M}\sum_{i=1}^{N}\mathrm{E}[v_{1j}^2 - \mathrm{E}v_{11}^2|U_{0j}] \\ &= O_{\mathrm{P}}\Big(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\Big) = O_{\mathrm{P}}\Big(\frac{1}{\sqrt{\underline{C}}}\Big),\end{aligned}$$

where the second equality follows from Lyapunov's CLT applied under Assumption 2 (1) – note that Assumption 2 implies the third moments for both terms of the right-hand side to be bounded. Assumption 2 and the second claim in Lemma 1 imply that $V(H_n f) = V(\mathbb{G}_C f) + O(\underline{C}^{-1})$. Since $H_n f$ is a projection of $\mathbb{G}_C f$, we obtain $\frac{1}{\sqrt{\underline{C}}}\mathbb{G}_C f = O_{\mathrm{P}}(\underline{C}^{-1/2}) = o_{\mathrm{P}}(1)$.

Step 4. To bound the second term on the RHS of equation (A.12), note that

$$\begin{aligned}\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(\widehat{v}_{ij}^2 - v_{ij}^2) =& \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}[(X_{ij}\gamma)^2 - (X_{ij}\widehat{\gamma})^2] + \frac{2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}D_{ij}X_{ij}'(\widehat{\gamma}-\gamma) \\ &+ \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(R_{ij}^D)^2 - \frac{2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}D_{ij}R_{ij}^D + \frac{2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}R_{ij}^D X_{ij}'\gamma \\ =& (3a) + (3b) + (3c) + (3d) + (3e).\end{aligned}$$

The first term can be bounded by

$$\begin{aligned}|(3a)| =& \Big|\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}[(X_{ij}\gamma)^2 - (X_{ij}\widehat{\gamma})^2]\Big| \\ \lesssim_{\mathrm{P}}& \sup_{\substack{\|\delta\|=1 \\ \|\delta\|_0\le Cs}}\delta'\Big(\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}X_{ij}X_{ij}'\Big)\delta\cdot\|\widehat{\gamma}-\gamma\|^2 + 2\sqrt{\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(X_{ij}'\gamma)^2}\cdot\|\widehat{\gamma}-\gamma\| \\ \le& \sqrt{\phi_{\max}(Cs)}\|\widehat{\gamma}-\gamma\|^2 + 2O_{\mathrm{P}}\Big(\sqrt{\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\mathrm{E}(X_{ij}'\gamma)^2}\Big)\|\widehat{\gamma}-\gamma\| \lesssim_{\mathrm{P}} \sqrt{\frac{s\log a}{\underline{C}}},\end{aligned}$$

where the first inequality follows from Assumption 3 (1), Lemma 3, and Lemma 3.1 in the supplementary appendix of van de Geer, Bühlmann, Ritov and Dezeure (2014), and the third follows from Theorem 1 and Assumption 4. An application of Cauchy-Schwartz's inequality and Theorem 1 gives $(3b) = O_{\mathrm{P}}(\sqrt{s\log a/\underline{C}})$. $(3c) = s/\underline{C}$ follows from Assumption 3. Cauchy-Schwartz's inequality and Assumptions 2(1), 3(2) lead to $(3d) \le \|(NM)^{-1/2}R^D\|\sqrt{\mathrm{E}[D_{11}^2]} = O_{\mathrm{P}}(\sqrt{s/\underline{C}})O(1)$. Using the property of the projection and a

similar argument to that of $(3d)$, we conclude that $(3e) \leq \|(NM)^{-1/2}R^D\|\sqrt{\mathrm{E}[D_{11}^2]} = O_{\mathrm{P}}(\sqrt{s/\underline{C}})O(1)$. This along with the conclusion of Step 3 show $|\widehat{Q} - Q| = o_{\mathrm{P}}(1)$. Applying the continuous mapping theorem under Assumption 2(1) then gives $|\widehat{Q}^{-1} - Q^{-1}| = o_{\mathrm{P}}(1)$.

Step 5. Finally, $|\widehat{\Gamma}| \leq |\Gamma| + |\widehat{\Gamma} - \Gamma| = O_{\mathrm{P}}(1)$ following the bounds from Steps 1 and 2 and Assumption 2 (1). Similarly, $|\widehat{Q}^{-1}| \leq |Q^{-1}| + |\widehat{Q}^{-1} - Q^{-1}|$ are both bounded following Assumption 2 (1) and Steps 3 and 4. ∎

Appendix B. Auxiliary Lemmas

The following Lemma is an immediate consequence of Theorem 5.1 of Chernozhukov, Chetverikov and Kato (2014) and Lemma 8 of Chernozhukov, Chetverikov and Kato (2015).

Lemma 2 (A Concentration Inequality)**.** *Let $(X_i)_{i\in[n]}$ be p-dimensional independent random vectors and let $B = \sqrt{E[\max_{i\in[n]}\|X_i\|_\infty^2]}$ and $\sigma^2 = \max_{j\in[p]}\frac{1}{n}\sum_{i=1}^n E|X_{ij}|^2$. Then with probability at least $1 - C(\log n)^{-1}$,*

$$\max_{j\in[p]}\Big|\frac{1}{n}\sum_{i=1}^n(|X_{ij}| - E|X_{ij}|)\Big| \lesssim \sqrt{\frac{\sigma^2\log(p\vee n)}{n}} + \frac{B\log(p\vee n)}{n}.$$

The following is an immediate result of Lemma 10 of Belloni, Chen, Chernozhukov and Hansen (2012) with $n = NM$, $\lambda = CNM\sqrt{\log a/\underline{C}}$ for $C > 1$ and $c_s = \sqrt{s/\underline{C}}$.

Lemma 3 (Sparsity Bound for Lasso)**.** *Consider lasso estimator (3.3) and suppose Assumption 3 (1)–(2) and 4. Then suppose $\lambda_2/NM \geq c\|(NM)^{-1}\sum_{i=1}^N\sum_{j=1}^M X_{ij}v_{ij}\|_\infty$ w.p. $1 - o(1)$, then denote $\widehat{s} = \mathrm{support}(\widehat{\gamma})$, we have $\widehat{s} \lesssim_{\mathrm{P}} s$. Similar result holds for lasso estimator (3.4) as well.*

References

Anderson, Siwan. "Legal origins and female HIV." *American Economic Review*, 108, no. 6 (2018): 1407-39.

Andrews, Donald W.K. "Cross-section regression with common shocks." *Econometrica*, 73, no. 5 (2005): 1551-1585.

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen. "Sparse models and methods for optimal instruments with an application to eminent domain." *Econometrica*, 80, no. 6 (2012): 2369-2429.

Belloni, Alexandre, and Victor Chernozhukov. "Least squares after model selection in high-dimensional sparse models." *Bernoulli*, 19, no. 2 (2013): 521-547.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. "Inference for high-dimensional sparse econometric models." arXiv preprint arXiv:1201.0220 (2011).

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. "Inference on treatment effects after selection among high-dimensional controls." *Review of Economic Studies*, 81, no. 2 (2014): 608-650.

Belloni, Alexandre, Victor Chernozhukov, Christian Hansen, and Damian Kozbur. "Inference in high-dimensional panel models with an application to gun control." *Journal of Business and Economic Statistics*, 34, no. 4 (2016): 590-605.

Belloni, Alexandre, Victor Chernozhukov, and Kengo Kato. "Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems." *Biometrika*, 102, no. 1 (2015): 77-94.

Cameron, Colin A., Jonah B. Gelbach, and Douglas L. Miller. "Robust inference with multiway clustering." *Journal of Business and Economic Statistics*, 29, no. 2 (2011): 238-249.

Cameron, Colin A. and Douglas L. Miller. "A practitioner's guide to cluster-robust inference." *Journal of Human Resources* 50, no. 2 (2015): 317-372.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. "Gaussian approximation of suprema of empirical processes." *Annals of Statistics*, 42, no. 4 (2014): 1564-1597.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. "Comparison and anti-concentration bounds for maxima of Gaussian random vectors." *Probability Theory and Related Fields* 162, no. 1-2 (2015): 47-70.

Davezies, Laurent, Xavier D'Haultfoeuille, and Yannick Guyonvarch. "Asymptotic results under multiway clustering." arXiv preprint arXiv:1807.07925 (2018).

Davezies, Laurent, Xavier D'Haultfoeuille, and Yannick Guyonvarch. "Empirical Process Results for Exchangeable Arrays." arXiv preprint arXiv:1906.11293 (2019).

Dickens, Andrew. "Ethnolinguistic favoritism in African politics." *American Economic Journal: Applied Economics*, 10, no. 3 (2018): 370-402.

Gershman, Boris. "Witchcraft beliefs and the erosion of social capital: Evidence from Sub-Saharan Africa and beyond." *Journal of Development Economics*, 120 (2016): 182-208.

Javanmard, Adel and Andrea Montanari. "Confidence intervals and hypothesis testing for high-dimensional regression." *Journal of Machine Learning Research*, 15, no. 1 (2014): 2869-2909.

Kallenberg, Olav. Probabilistic symmetries and invariance principles. Springer Science and Business Media (2005).

Kock, Anders Bredahl. "Oracle inequalities, variable selection and uniform inference in high-dimensional correlated random effects panel data models." *Journal of Econometrics*, 195, no. 1 (2016): 71-85.

MacKinnon, James G., Morten Ørregaard Nielsen, and Matthew D. Webb. "Wild Bootstrap and Asymptotic Inference with Multiway Clustering." No. 1415. Queen's Economics Department Working Paper, 2019.

Menzel, Konrad. "Inference for games with many players." *Review of Economic Studies*, 83, no. 1 (2015): 306-337.

Menzel, Konrad. "Bootstrap with clustering in two or more dimensions." arXiv preprint arXiv:1703.03043 (2017).

Michalopoulos, Stelios and Elias Papaioannou. "PreColonial Ethnic Institutions and Contemporary African Development." *Econometrica*, 81, no. 1 (2013) 113-152.

Michalopoulos, Stelios, and Elias Papaioannou. "National institutions and subnational development in Africa." The *Quarterly Journal of Economics*, 129, no. 1 (2013): 151-213.

Michalopoulos, Stelios, and Elias Papaioannou. "The long-run effects of the scramble for Africa." *American Economic Review*, 106, no. 7 (2016): 1802-48.

Nunn, Nathan and Leonard Wantchekon. "The slave trade and the origins of mistrust in Africa." *American Economic Review*, 101, no. 7 (2011): 3221-3252.

van de Geer, Sara, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. "On asymptotically optimal confidence regions and tests for high-dimensional models." *Annals of Statistics*, 42, no. 3 (2014): 1166-1202.

Zhang, Cun-Hui, and Stephanie S. Zhang. "Confidence intervals for low dimensional parameters in high dimensional linear models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, no. 1 (2014): 217-242.

(Harold Chiang) Department of Economics, Vanderbilt University, United States

(Yuya Sasaki) Department of Economics, Vanderbilt University, United States

			Statistics				95% Coverage		
N	M	Dim	Avg	Bias	SD	RMSE	0-Way	1-Way	2-Way
10	10	100	0.491	-0.009	0.172	0.172	0.808	0.797	0.919
20	20	100	0.499	-0.001	0.076	0.076	0.855	0.858	0.964
40	40	100	0.501	0.001	0.045	0.045	0.792	0.848	0.959
10	10	200	0.480	-0.020	0.356	0.357	0.753	0.747	0.877
20	20	200	0.498	-0.002	0.075	0.075	0.863	0.860	0.962
40	40	200	0.500	0.000	0.041	0.041	0.830	0.859	0.962
10	10	400	0.459	-0.041	0.744	0.745	0.682	0.690	0.822
20	20	400	0.496	-0.004	0.079	0.079	0.846	0.841	0.951
40	40	400	0.500	0.000	0.037	0.037	0.869	0.876	0.970
10	10	800	0.438	-0.062	0.959	0.961	0.615	0.634	0.764
20	20	800	0.492	-0.008	0.086	0.086	0.808	0.803	0.930
40	40	800	0.499	-0.001	0.037	0.037	0.870	0.871	0.968
10	10	1600	0.394	-0.106	1.140	1.140	0.555	0.585	0.704
20	20	1600	0.487	-0.013	0.098	0.099	0.763	0.762	0.903
40	40	1600	0.498	-0.002	0.038	0.038	0.862	0.859	0.964

TABLE 1. Simulation results. The first three columns indicate the two-way sample sizes (N, M) and the dimension (Dim) of $(\alpha, \beta')'$. The next four columns report simulation statistics for $\widetilde{\alpha}$, including the average (Avg), bias (Bias), standard deviation (SD), and root mean square error (RMSE). The last three columns report 95% coverage frequencies of α with the heteroskedasticity robust variance estimator (0-Way), the one-way cluster-robust variance estimator (1-Way), and our multi-way cluster-robust variance estimator (2-Way). The data generating parameters are set to $(\omega_1^x, \omega_2^x) = (0.25, 0.25)$, $(\omega_1^\varepsilon, \omega_2^\varepsilon) = (0.25, 0.25)$, and $\rho = 0.50$. The results are based on 25,000 Monte Carlo iterations for each row.

Variables		Number of	Cluster Size		Original	Lasso
Y	D	Observations	N	M	Estimates	Estimates
Trust of	Slave	20,027	185	1,257	-0.00068	-0.00083
Neighbors	Exports				(0.00015)	(0.00022)
Trust of	Exports/	20,027	185	1,257	-0.019	-0.025
Neighbors	Area				(0.005)	(0.007)
Trust of	Exports/	17,644	157	1,214	-0.531	-0.684
Neighbors	Population				(0.147)	(0.232)
Trust of	Log Slave	20,027	185	1,257	-0.037	-0.045
Neighbors	Exports				(0.014)	(0.021)
Trust of	Log Exports/	20,027	185	1,257	-0.159	-0.210
Neighbors	Area				(0.034)	(0.050)
Trust of	Log Exports/	17,644	157	1,214	-0.743	-0.957
Neighbors	Population				(0.187)	(0.304)

TABLE 2. Estimates of the effects of slave trade on mistrust in Africa. The first two columns indicate which measures of the dependent and explanatory variables are used. The next three columns show the number of observations, the number of ethnic groups (N), and the number of districts (M). The last two columns show the original estimates obtained under the prototype model by Nunn and Wantchekon (2011, Table 1) and corresponding lasso estimates obtained under more flexible model specification by our method.

Variables		No.	Cluster Size		Population	Original	Lasso
Y	D	Obs.	N	M	Density	Estimates	Estimates
Light	Jurisdictional	682	93	48	No	0.2794	0.2266
Density	Hierarchy					(0.0852)	(0.0797)
Light	Jurisdictional	682	93	48	Yes	0.1766	0.1649
Density	Hierarchy					(0.0501)	(0.0541)
Light	Political	682	93	48	No	0.5049	0.4158
Density	Centralization					(0.1573)	(0.1451)
Light	Political	682	93	48	Yes	0.3086	0.2985
Density	Centrilization					(0.0972)	(0.1080)

TABLE 3. Estimates of the effects of pre-colonial institutions on regional development in Africa. The first two columns indicate which measures of the dependent and explanatory variables are used. The next three columns show the number of observations, the number of ethnic groups (N), and the number of districts (M). The next column indicates a control for population density. The last two columns show the original estimates obtained under the prototype model by Michalopoulos and Papaioannou (2013, Table 3) and corresponding lasso estimates obtained under more flexible model specification by our method.