Many-covariate and cluster robust estimation and inference

Stanislav Anatolyev[*]
CERGE-EI and NES

Cheuk Fai Ng[†]
Cambridge University

Abstract

Empirical economists often use regression models employing large sets of covariates and presuming clustered data dependence. We provide inference methods for linear regressions with covariates whose number may be comparable to sample size, and observations that are clustered into possibly heterogeneous clusters. We present a leave-cluster-out-crossfit (LCOC) method of constructing an OLS asymptotic variance estimator, which extends leave-one-out variance estimation for independent data to clustered data, and which is robust to many covariates and heteroskedasticity. We show consistency of the LCOC estimator and asymptotic normality of the standardized OLS estimator. We demonstrate finite-sample properties of LCOC in simulations in comparison with available alternatives. Finally, we provide two empirical illustrations, where LCOC is applied to existing studies of effects of high school achievement awards and an impact of legalized abortion on crime reduction.

Keywords: linear regression, heteroskedasticity, clustered sample, many covariates, leave-out estimation, variance estimation.

JEL classification codes: C12, C13, C21

[*] Corresponding author. Address: CERGE-EI, Politických vězňů 7, 11121 Prague 1, Czech Republic. E-mail: stanislav.anatolyev@cerge-ei.cz.

[†] Address: University of Cambridge, Faculty of Economics, Sidgwick Avenue, Cambridge CB3 9DD, United Kingdom. E-mail: cfn24@cam.ac.uk.

1 Introduction

In linear regression models, it is common to assume that the observations can be partitioned into clusters, where observations within one cluster are correlated, while being independent across clusters. The usual approach to estimation and inference is to estimate the regression model without explicitly controlling for within-cluster error correlation, and then using what is known as cluster-robust inference (e.g., cluster-robust standard errors), see White (1984), Liang and Zeger (1986), Arellano (1987) and Bester, Conley, and Hansen (2011). The popularity of cluster-robust standard errors among applied researchers soared after Rogers (1993) incorporated this technique into Stata. Comprehensive reviews on this subject are Cameron and Miller (2015) and Imbens and Kolesár (2016).

The asymptotic theory for clustered samples remains an active area of study as researchers strive to relax required assumptions or extend applicable settings. It is typically developed under the "large G" asymptotic framework, where the number of clusters tends to infinity, though slowly, with sample size. The earliest treatment is White (1984) for balanced clusters. Carter, Schnepel, and Steigerwald (2017) extended asymptotic results to allow for unbalanced clusters. Hansen and Lee (2019) developed an asymptotic framework that accommodates both unbalanced clusters and growing cluster size. An alternative framework is the "fixed G" asymptotics, where the number of clusters is fixed. Ibragimov and Müller (2010, 2016) provided an inferential theory in this setting. Recently, it was shown by Leung (2023) that these cluster-robust methods could even be applied to network-dependent data as long as their proposed primitive conditions are met.

The other direction of relaxing the classical regression setup is to allow for numerous regressors (or covariates). Crucially, the aforementioned asymptotic results do not automatically extend to models with "many regressors," but rather are generally not robust to such features. "Many regressors" refers to a setting, where the number of covariates is comparable to the sample size; see Anatolyev (2019) for a survey on the literature about many regressors. Classical asymptotic settings such as those in White (1984) tend to require the regressor design matrix, when properly normalized, to have a well defined probability limit, specifically a positive definite matrix. This requirement is violated in a many regressor setting as the dimension of the design matrix is asymptotically growing, up to proportionality to the growing sample size. Therefore, the asymptotic properties of OLS estimation need to be reestablished in such a setting.

As Bell and McCaffrey (2002) have demonstrated in their simulations, the conventional Liang and Zeger (1986) cluster-robust variance estimator (referred to as LZ estimator hereafter) exhibits heavy bias in finite samples (for a recent discussion, see Young 2019). A substantial body of research has emerged to investigate and address this finite sample problem. One widely adopted alternative to the LZ estimator is the jackknife estimator for clustered samples developed by Cochran (1977) (referred to as JK estimator). Another popular alternative was introduced by Bell and McCaffrey (2002) (referred to as BM estimator) (see also Imbens and Kolesár 2016, and Kolesár 2023). Hansen (2025) provides an excellent review as well as a discussion on how these two estimators are connected algebraically. However, the jackknife estimator is also (positively) biased, as evidenced by our own simulation results, and this bias becomes more pronounced in the presence of many regressors, as it does not asymptotically vanish. Therefore, the existing variance estimators, including the LZ, JK and BM, would generally be inconsistent in the presence of many regressors. This gives rise to a need for a consistent cluster-robust and many-covariate-robust variance estimator.

A good deal of development in regression theory try to account for the presence of many regressors. Calhoun (2011) and Anatolyev (2012) were among the first attempts to robustify classical inference in linear regressions with many regressors in a conditionally homoskedastic setup. Subsequently, heteroskedasticity-robust methods and results that focus on a finite number of parameters of interest have been developed by Cattaneo, Jansson, and Newey (2018b, CJN henceforth), Kline, Saggio, and Sølvsten (2020, KSS henceforth), Jochmans (2020) and Anatolyev and Sølvsten (2023).

The present paper contributes to the literature by adopting cluster dependence in a linear model with many regressors, as none of latest published research on cluster robust standard errors treat the case of many regressors, nor mention it (see, e.g., Hansen and Lee 2019, Canay, Santos, and Shaikh 2021 and Leung 2023). There are, however, two closely related unpublished manuscripts D'Adamo (2019) and Gong (2022). D'Adamo (2019) provides a cluster analogue of CJN (2018b)'s heteroskedasticity-robust variance estimation, while Gong (2022) extends D'Adamo (2019)'s work to allow for a moderate rate of cluster growth. However, these "Hadamard" type estimators often require certain strong (though sufficient) assumptions to hold in a large sample. For example, CJN (2018b)'s estimator for independent data requires the maximal leverage (i.e. the maximal diagonal element of the projection matrix associated with covariates) to be bounded from above by $\frac{1}{2}$. As revealed by simulations in Jochmans (2020), CJN (2018b)'s estimator is frequently non-existent, even for smaller leverage values.

In the present paper, we first generalize Cattaneo, Jansson, and Newey (2018b)'s asymptotic framework to accommodate cluster dependence and provide an asymptotic normality result for a finite number of OLS coefficients in a linear model with many regressors under cluster dependence. In the same fashion as Hansen and Lee (2019), we allow for unbalanced and asymptotically unbounded clusters. Our asymptotic theory implies that there is an apparent trade-off between the growth rate of cluster size, on the one hand, and the accuracy of the auxiliary regression of the regressors of interest on the nuisance covariates.

To conduct valid inference, we introduce an unbiased estimator that we call a leave-cluster-out crossfit (LCOC) estimator. This estimator is essentially the cluster analogue of the leave-one-out crossfit (LOOC) estimator proposed by Kline, Saggio, and Sølvsten (2020) for independent data. The term "crossfiting," coined by Newey and Robins (2018), refers to a sample splitting technique aimed at bias reduction. Typically, when an estimator contains multiple naive "plug-in" parts, these plug-in parts are often correlated due to having been estimated using the same sample, and this correlation between plug-in parts may cause estimation bias. In a linear model, where conditional exogeneity is assumed and the form of dependence is known, crossfiting (i.e. sample splitting) becomes an effective method for having tight control on such types of biases.[1] Simply put, the LCOC estimator is designed by replacing the outer product of cluster-level OLS residual vectors in the LZ estimator with the outer product of the cluster-level outcome vector and LCOC residual vector. Per se, such cluster-specific covariance matrix estimates, while unbiased, are imprecise estimates of the true cluster-specific covariance matrices. However, by averaging these estimated cluster-specific covariances across all clusters, the resulting estimate of the OLS asymptotic variance is nonetheless consistent.

Our paper draws parallels with Jochmans (2020), whose estimator can be considered as an approximate LOOC estimator for independent data, or in other words, for clustered data of a fixed size of unity. We too establish consistency of a crossfit type estimator in an asymptotic framework that closely follows CJN (2018b). The only difference between Jochmans (2020)'s estimator and KSS (2020)'s LOOC estimator is that the jackknife residual is computed using the auxiliary projection matrix instead of the usual projection matrix that uses the full set of regressors. Because of the cluster structure, our consistency result requires restrictions on the Hadamard products of the projection matrix and each of its cluster blocks, plus restrictions on the maximal cluster growth rate.

To assess the finite sample performance of our LCOC estimator, we conduct multiple Monte Carlo experiments. We find that the conditional unbiasedness property allows the LCOC estimator to attain near-oracle performance in all specifications when the alternative estimators fail at least in some of them. Among other setups, we employ a partially linear model, where the conditional unbiasedness of the LCOC estimator is lost. Despite a slight drop in performance, LCOC still offers good size control and power properties in this setup as well.

Finally, we provide two empirical applications to illustrate the operation of the LCOC estimator. The first application revisits the study of Angist and Lavy (2009) about the effects of high school achievement awards. We find that most of their original results still hold even when we conduct inference using the LCOC method. The second application reconducts Donohue and Levit's (2002) investigation into the causal impact of legalized abortion on crime reduction. We find that for more serious crimes like violence and murder, their conclusions cease to be valid in a higher dimensional specification when we conduct inference using the LCOC method.

The remainder of the paper is structured as follows. Section 2 introduces the model and shows the construction of the LCOC estimator. Section 3 presents and discusses the model assumptions and establishes the LCOC estimator's properties. Section 4 reports Monte Carlo experiments that compare the performance of the LCOC estimator with the alternatives. In

[1] See Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) for a discussion of the role of crossfiting in removing bias induced by overfitting in a high dimensional linear model.

Section 5, we provide empirical applications to illustrate the LCOC methods. We conclude in Section 6 with a brief summary and suggestions for further research. The Appendix contains auxiliary lemmas and proofs of theoretical propositions.

2 Setup and Estimators

We start by introducing the framework together with the accompanying notation, and then describe estimation techniques. The model assumptions and asymptotic theory are presented later in Section 3.

2.1 Framework

Consider a sample of size n composed of G non-overlapping clusters. A cluster indexed by $g \in \{1, \ldots, G\}$ contains n_g observations, with $\sum_{g=1}^{G} n_g = n$. For the i^{th} observation in the g^{th} cluster, $i \in \{1, \ldots, n_g\}$, we observe a scalar output variable $y_{g,i}$, an $r \times 1$ vector of regressors (or causal variables) $x_{g,i}$, and a $k \times 1$ vector of covariates (control variables) $w_{g,i}$. The full clustered sample is the collection $\{\{y_{g,i}, x_{g,i}, w_{g,i}\}_{i=1}^{n_g} : 1 \leq g \leq G\}$.

Suppose $\big\{\{y_{g,i}, x_{g,i}, w_{g,i}\}_{i=1}^{n_g} : 1 \leq g \leq G\big\}$ satisfy the following relation, which we refer to as the primary equation:

$$y_{g,i} = x_{g,i}'\beta + w_{g,i}'\gamma + u_{g,i} \tag{2.1}$$

for $i = 1, \ldots, n_g$ and $g = 1, \ldots, G$. The dimensionality of β is $r \times 1$, and the dimensionality of γ is $k \times 1$. The error term is assumed to satisfy conditional exogeneity, i.e.

$$\mathrm{E}(u_{g,i}|\mathcal{X}, \mathcal{W}) = 0,$$

where $\mathcal{X} = \big\{\{x_{g,i}\}_{i=1}^{n_g} : 1 \leq g \leq G\big\}$, and $\mathcal{W}$ is a collection of random variables satisfying $\mathrm{E}(w_{g,i}|\mathcal{W}) = w_{g,i}$ for all $i = 1, \ldots, n_g$ and $g = 1, \ldots, G$. Note that the number of elements in $\mathcal{W}$ can potentially be smaller than k, as $w_{g,i}$ can be a function of a smaller group of random variables (e.g., polynomial functions or interactive dummies).

The second relation, which we refer to as the auxiliary equation, reads

$$x_{g,i} = \alpha' w_{g,i} + v_{g,i} = \mathrm{E}(x_{g,i}|\mathcal{W}) + V_{g,i}. \tag{2.2}$$

Here, $\alpha = \big(\sum_{g=1}^{G}\sum_{j=1}^{n_g} \mathrm{E}(w_{g,j}w_{g,j}')\big)^{-1} \sum_{g=1}^{G}\sum_{j=1}^{n_g} \mathrm{E}(w_{g,j}x_{g,j}')$, and $v_{g,i}$, an $r \times 1$ vector, is a deviation of $x_{g,i}$ from its population linear projection.

In this paper, we assume that the linear equation is correctly specified for the conditional mean function $\mathrm{E}(y_{g,i}|\mathcal{X}, \mathcal{W})$ by assuming conditional exogeneity. The exogeneity condition is necessary for the unbiasedness of our leave-cluster-out crossfit (LCOC) estimator. Relaxing this assumption would allow an asymptotically negligible amount of misspecification bias in the primary equation (see CJN 2018b), but then the LCOC estimator would lose its conditional unbiasedness property. This assumption is crucial, as constructing test statistics using an unbiased variance estimator will often yield good size control as seen in our simulation results and other papers that use crossfit type estimators (see KSS 2018 and Anatolyev and Sølvsten 2023). However, the auxiliary equation (2.2) may be misspecified in the sense that $\alpha' w_{g,i} \neq \mathrm{E}(x_{g,i}|\mathcal{W})$. This implies that one must distinguish the population projection error v_i from the conditional mean error $V_{g,i}$, as $\mathrm{E}(v_{g,i}|\mathcal{W}) \neq \mathrm{E}(V_{g,i}|\mathcal{W}) = 0$. This two-equation representation is a common way of viewing the relationships among covariates in the high dimensional literature (see, for example, Belloni, Chernozhukov, and Hansen 2014). Since the OLS can be thought of as a two-step estimator from the Frisch-Waugh-Lovell (FWL) theorem perspective, the population least square residual $v_{g,i}$ will enter the population expression for β. While we are not interested in the reduced form coefficient α, this two-equation representation allows us to impose regularity conditions on $V_{g,i}$ and $v_{g,i}$, which in turn will regulate the distributional relationship between the regressors of interest $x_{g,i}$ and nuisance covariates $w_{g,i}$.

The dimension p of the right-side variables can potentially be comparable to the sample size n. The goal is to conduct asymptotically valid cluster-robust inference on the asymptotically finite-dimensional parameter $r \times 1$ vector β, when the dimension k of the nuisance parameter vector γ can be asymptotically proportional to the sample size n (as long as k is smaller than $n - r$), so, asymptotically, $0 \leq \overline{\lim}_{n\to\infty} k/n < 1$. Cluster dependence means that the observations can be partitioned into clusters, so that observations within a cluster are correlated while they are independent across clusters. Our inference method will be able to deal with the conditional heteroskedasticity, which is allowed in CJN (2018b), as well as with arbitrary within-cluster conditional correlatedness.

For convenience, we define the $p \times 1$ vector of all observable right-side variables, where $p = r + k$, as

$$\tilde{x}_{g,i} = \begin{pmatrix} x_{g,i} \\ w_{g,i} \end{pmatrix}$$

for $g \in \{1, \dots, G\}$ and $i \in \{1, \dots, n_g\}$, and, correspondingly, let $\tilde{\beta} = (\beta', \gamma')'$. For the sake of brevity later on, we define

$$\mu_{g,i} = x'_{g,i}\beta + w'_{g,i}\gamma = \tilde{x}'_{g,i}\tilde{\beta}$$

for $g \in \{1, \dots, G\}$ and $i \in \{1, \dots, n_g\}$. Accordingly, cluster-wise,

$$\mu_g = X_g\beta + W_g\gamma = \tilde{X}_g\tilde{\beta},$$

where $\mu_g = (\mu_{g,1}, \mu_{g,2}, \dots, \mu_{g,n_g})'$, $X_g = (x_{g,1}, x_{g,2}, \dots, x_{g,n_g})'$, $W_g = (w_{g,1}, w_{g,2}, \dots, w_{g,n_g})'$ and $\tilde{X}_g = (X_g, W_g)$ for $g \in \{1, \dots, G\}$. Last, sample-wise, let

$$\mu = X\beta + W\gamma = \tilde{X}\tilde{\beta},$$

where $\mu = (\mu'_1, \dots, \mu'_G)'$, $X = (X'_1, \dots, X'_G)'$, $W = (W'_1, \dots, W_G)'$ for each $g \in \{1, \dots, G\}$, and $\tilde{X} = (X, W)$. Similarly, we denote $y_g = (y_{g,1}, y_{g,2}, \dots, y_{g,n_g})'$ for $g \in \{1, \dots, G\}$ and

$y = (y_1', y_2', \ldots, y_G')'$. The (g,i)-specific error term is $u_{g,i} = y_{g,i} - \mu_{g,i}$. We also write $u_g = (u_{g,1}, u_{g,2}, \ldots, u_{g,n_g})'$ for $g \in \{1, \ldots, G\}$ and $u = (u_1', u_2', \ldots, u_G')'$.

As we are often interested in different cluster-indexed submatrices of projection matrices, it is useful to agree on the following convention. Let A_g denote the submatrix containing the rows of matrix A indexed by the cluster g, and let A_{gh} denote the submatrix of A containing rows given by the cluster g and columns given by the cluster h.

2.2 OLS estimator

Using the introduced notation, we can rewrite equation (2.1) in the matrix form as

$$\begin{aligned} y &= X\beta + W\gamma + u \\ &= \tilde{X}\tilde{\beta} + u. \end{aligned}$$

The OLS estimator of $\tilde{\beta}$ is given by

$$\begin{aligned} \hat{\tilde{\beta}} &= (\tilde{X}'\tilde{X})^{-1}\tilde{X}'y \\ &= \Big(\sum_{g=1}^{G} \tilde{X}_g'\tilde{X}_g\Big)^{-1} \sum_{g=1}^{G} \tilde{X}_g' y_g \\ &= \Big(\sum_{g=1}^{G}\sum_{i=1}^{n_g} \tilde{x}_{g,i}\tilde{x}_{g,i}'\Big)^{-1} \sum_{g=1}^{G}\sum_{i=1}^{n_g} \tilde{x}_{g,i} y_{g,i}. \end{aligned}$$

It is also useful to have a OLS formula for the parameters of interest β. To this end, let us define $\hat{v}_{g,i} = x_{g,i} - \hat{\alpha}'w_{g,i}$ for $g \in \{1, \ldots, G\}$ and $i \in \{1, \ldots, n_g\}$ and $\hat{v}_g = X_g - W_g\hat{\alpha}$ for $g \in \{1, \ldots, G\}$. Here, $\hat{\alpha} = \big(\sum_{g=1}^{G} W_g'W_g\big)^{-1} \sum_{g=1}^{G} W_g'X_g$ is the OLS estimate of α in equation (2.2). Let $\hat{v} = (\hat{v}_1', \hat{v}_2', \cdots, \hat{v}_G')'$. Then, by the Frisch-Waugh-Lovell theorem, the OLS estimator for β reads

$$\begin{aligned} \hat{\beta} &= (\hat{v}'\hat{v})^{-1}\hat{v}'y \\ &= \Big(\sum_{g=1}^{G} \hat{v}_g'\hat{v}_g\Big)^{-1} \sum_{g=1}^{G} \hat{v}_g' y_g \\ &= \Big(\sum_{g=1}^{G}\sum_{i=1}^{n_g} \hat{v}_{g,i}\hat{v}_{g,i}'\Big)^{-1} \sum_{g=1}^{G}\sum_{i=1}^{n_g} \hat{v}_{g,i} y_{g,i}. \end{aligned}$$

For further analysis, we define various projection matrices. Denote by $\tilde{H} = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'$ and $H = W(W'W)^{-1}W'$ the projection ("hat") matrices based on $\tilde{X}$ and W, respectively, and introduce associated annihilator matrices $\tilde{M} = I - \tilde{H}$ and $M = I - H$.

2.3 OLS variance

At the inference step, we are interested in the variance of $\hat{\beta}$ conditional on $(\mathcal{X},\mathcal{W})$. Define the conditional covariance matrix for u as

$$\mathrm{E}(uu'|\mathcal{X},\mathcal{W}) = \begin{pmatrix} \Omega_1 & 0 & \cdots & 0 \\ 0 & \Omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega_G \end{pmatrix} = \Omega,$$

whose block-diagonality follows from the assumed clustering structure. The conditional variance expression for $\hat{\beta}$ is then given by

$$\begin{aligned} \mathrm{var}\big(\hat{\beta}|\mathcal{X},\mathcal{W}\big) &= (\hat{v}'\hat{v})^{-1}\big(\hat{v}'\Omega\hat{v}\big)(\hat{v}'\hat{v})^{-1} \\ &= \Big(\sum_{g=1}^{G}\hat{v}_g'\hat{v}_g\Big)^{-1}\Big(\sum_{g=1}^{G}\hat{v}_g'\Omega_g\hat{v}_g\Big)\Big(\sum_{g=1}^{G}\hat{v}_g'\hat{v}_g\Big)^{-1} \\ &= \Big(\sum_{g=1}^{G}\sum_{i=1}^{n_g}\hat{v}_{g,i}\hat{v}_{g,i}'\Big)^{-1}\Big(\sum_{g=1}^{G}\sum_{i=1}^{n_g}\sum_{j=1}^{n_g}\hat{v}_{g,i}\hat{v}_{g,j}'(\Omega_g)_{ij}\Big)\Big(\sum_{g=1}^{G}\sum_{i=1}^{n_g}\hat{v}_{g,i}\hat{v}_{g,i}'\Big)^{-1}. \end{aligned}$$

2.4 LCOC estimator

The middle matrix

$$\Sigma = \sum_{g=1}^{G}\hat{v}_g'\Omega_g\hat{v}_g$$

in the sandwich form of $\mathrm{var}(\hat{\beta}|\mathcal{X},\mathcal{W})$ is infeasible and needs to be estimated. The estimator we propose in this paper is the following *leave-cluster-out crossfit (LCOC)* estimator.

Definition 2.1 (Leave-Cluster-Out Crossfit Estimator)**.** Define

$$\hat{\Sigma}_{\mathrm{LCOC}} = \sum_{g=1}^{G}\hat{v}_g'\,\frac{y_g\,(\hat{u}_{-g})_g' + (\hat{u}_{-g})_g\,y_g'}{2}\,\hat{v}_g, \tag{2.3}$$

where $\hat{u}_{-g} = y - \tilde{X}\hat{\hat{\beta}}_{-g}$ is the leave-cluster-out residual vector, $(\hat{u}_{-g})_g$ is its cluster g subvector, and $\hat{\hat{\beta}}_{-g} = \big(\tilde{X}_{-g}'\tilde{X}_{-g}\big)^{-1}\tilde{X}_{-g}'y_{-g}$ is the leave-cluster-out parameter estimate.

Note that the construction of $\hat{\Sigma}_{\mathrm{LCOC}}$ embeds a symmetrization device: for any square non-symmetric matrix A, the matrix $\frac{1}{2}(A + A')$ is symmetric. While all the statistical properties of $\hat{\Sigma}_{\mathrm{LCOC}}$ investigated below are the same as of its one-sided version, we prefer the estimator of the variance matrix to be symmetric.

From the practical standpoint, the LCOC estimator can be rewritten in the following vectorized from, which simplifies programming and increases computational speed:

$$\hat{\Sigma}_{\mathrm{LCOC}} = \hat{v}'\,\frac{\tilde{\Omega} + \tilde{\Omega}'}{2}\,\hat{v},$$

where

$$\tilde{\Omega} = \left(\tilde{M} \odot 1_{\mathrm{gr}}\right)^{-1}\left((\hat{u}y') \odot 1_{\mathrm{gr}}\right),$$

$\odot$ denotes Hadamard matrix product, and 1_{gr} is an adjacency matrix with elements $[1_{\mathrm{gr}}]_{ij} = 1$ if $\Omega_{ij} \neq 0$ and 0 otherwise. This uses identity $(\hat{u}_{-g})_g = \tilde{M}_{gg}^{-1}\hat{u}_g$ derived in Lemma (A.1) in the Appendix. As the matrix $\tilde{M} \odot 1_{\mathrm{gr}}$ is positive definite, it can be inverted quickly using the Cholesky decomposition.

We have developed a Stata module `vce_mcov` to compute the LCOC estimator. The module is available on the Statistical Software Components (SSC) archive. It can be installed in Stata by typing `ssc install vce_mcov`.

3 Asymptotic Theory

3.1 Assumptions

We impose the following five assumptions and discuss them in turn. The overall objective here is to generalize CJN (2018b)'s framework to data with the type of cluster dependence introduced in Hansen and Lee (2019). Naturally, the conditions generally become more restrictive when both the covariate numerosity and cluster structure are present, compared to frameworks with only one such a feature. Assumptions 1-4 are imposed for asymptotic normality of OLS estimation, and Assumption 5 is further imposed for LCOC consistency.

Assumption 1 (Sampling and cluster asymptotics)

(i) The collections of errors $\{u_{g,i}, V_{g,i}\}_{i=1}^{n_g}$ defined in (2.1) and (2.2) are independent across $1 \leq g \leq G$ conditional on $(\mathcal{X}, \mathcal{W})$,

(ii) As $n \to \infty$, we have $G \to \infty$ and $n_g \to \infty$ for each $1 \leq g \leq G$.

Assumption 1(i) outlines the sampling properties of the observed data. It allows for the kind of one-way clustered samples that is common in modern empirical work. The asymptotic framework in Assumption 1(ii) presumes that n, n_g and G simultaneously diverge to infinity. Note that cluster sizes may be growing arbitrarily slowly.

Assumption 2 (Design)

The following holds:

(i) $\Pr\left\{\lambda_{\min}\left(\sum_{g=1}^{G} W_g' W_g\right) > 0\right\} \to 1$, $\overline{\lim}_{n\to\infty} \dfrac{k}{n} < 1$,

(ii) $\max_{1\leq g\leq G} \max_{1\leq i\leq n_g} \mathrm{E}\left(u_{g,i}^4 | \mathcal{X}, \mathcal{W}\right) + \dfrac{1}{\lambda_{\min}(\Omega)} = O_p(1)$,

(iii) $\max_{1\leq g\leq G} \max_{1\leq i\leq n_g} \mathrm{E}(\|V_{g,i}\|^4 | \mathcal{W}) + \dfrac{1}{\lambda_{\min}\left(\mathrm{E}\left(n^{-1}\sum_{g=1}^{G} \tilde{V}_g' \tilde{V}_g | \mathcal{W}\right)\right)} = O_p(1)$, where $\tilde{V}_g = \sum_{h=1}^{G} M_{gh} V_h$.

The first condition in Assumption 2(i) keeps the design matrix of covariates from being too close to singularity. This is a generalization of the uniform non-singularity assumption (White 1984, p.22), which allows the rank of $\sum_{g=1}^{G} W_g' W_g$ to grow as n grows. This assumption is not restrictive, as any linear dependent nuisance covariates can be dropped without impacting the OLS estimate. The second condition in Assumption 2(i) allows the number of parameters to be estimated to grow in line with sample size as long as we have slightly more than one observations per parameter.

Assumptions 2(ii) and 2(iii) impose moment and rank restrictions on the errors in the primary and auxiliary equations. The forth moment conditions in Assumption 2(ii,iii) are the same as some in CJN (2018b)'s assumption 2.[2]

The minimal eigenvalue restriction in Assumption 2(ii) precludes degeneracy in any of cluster-level covariance matrices; it would be violated if the errors exhibit perfect correlation within a cluster. The minimal eigenvalue bound in Assumption 2(iii) is an analog of the textbook rank condition on the design matrix, extended to possible misspecification in the auxiliary equation,[3] and precludes degeneracy in the denominator of the OLS estimator and its variance, which is important for a proper asymptotic behavior of the estimator and its studentized version. Such minimal eigenvalue conditions are typical in the many-covariate and cluster-robust literatures (e.g., CJN 2018b, Hansen and Lee 2019).

Assumption 3 (Auxiliary approximation)

Let $Q_g = \mathrm{E}(v_g|\mathcal{W})$, then

$$\frac{1}{n}\sum_{g=1}^{G} \|Q_g\|^2 = O_p(1).$$

The condition in Assumption 3 is imposed on the approximation (or specification) error in using the linear reduced form $\alpha' w_{g,i}$ to approximate the conditional mean $\mathrm{E}(X_g|\mathcal{W})$. The size of this error is given by $Q_g = \mathrm{E}(X_g|\mathcal{W}) - W_g\alpha$ with elements $Q_{g,i}$, $1 \le i \le n_g$; note that the quantity Q_g is theoretical in the sense that α is a population coefficient. As the error Q_g influences the statistical properties of the OLS estimator of β via the FWL theorem, it must be regulated. In the expectations form, the condition $n^{-1}\sum_{g=1}^{G} \mathrm{E}(\|Q_g\|^2) = O(1)$ is a cluster analog of the first restriction in CJN (2018b)'s and Jochmans (2020)'s Assumption 3. This holds if $\mathrm{E}\big(\|x_{g,i}\|^2\big) = O(1)$ for all $1 \le g \le G$, $1 \le i \le n_g$, which is true, for example, when $x_{g,i}$ is binary, an important case in models of evaluation of treatment effects, or when $x_{g,i}$

[2] As CJN (2018b) mention, these may be pretty restrictive if the regressor support or heteroskedasticity is unbounded, but hold otherwise. For example, if $u_{g,i} = \omega_{g,i}\epsilon_{g,i}$, where $\max_{1\le g\le G}\max_{1\le i\le n_g}\omega_{g,i} < \infty$, an assumption often made in the many-regressor literature, and $\epsilon_{g,i}$ is independent of the regressors and covariates and has a finite fourth moment, then $\mathrm{E}\big(u_{g,i}^4|\mathcal{W}\big)$ is uniformly bounded above by a constant. On the other hand, if $\omega_{g,i}$ is related to $x_{g,i}$ and/or $w_{g,i}$, which have infinite support, then $\omega_{g,i}$ and hence $\mathrm{E}\big(u_{g,i}^4|\mathcal{W}\big)$ may be unbounded. A similar story applies to $V_{g,i}$.

[3] If there is no misspecification, Assumption 2(iii) implies a rank condition on the design matrix of $\hat{v}$.

is distributed on a continuous but compact support. However, in some setups of practical interest the average expectation $n^{-1}\sum_{g=1}^{G}\mathrm{E}(\|Q_g\|^2)$ is even $o(1)$, such as when the elements of $w_{g,i}$ are approximating functions, or when the covariates are saturating dummy variables. Next, the variance of the average of $\|Q_g\|^2$ is typically $o(1)$ under uniform integrability of $\|Q_{g,i}\|^4$ or if the distributions of $Q_{g,i}$'s are not overly thick-tailed.[4] These restrictions on the expectation and variance of the average of $\|Q_g\|^2$ together imply Assumption 3.

Assumption 4 (Cluster size and residual negligibility)

For some $\frac{1}{4} < \psi < \frac{1}{2}$,

(i) $\dfrac{\max_{1\leq g\leq G} n_g^{\frac{4}{1-2\psi}}}{n} = O(1)$,

(ii) $\dfrac{\max_{1\leq g\leq G}\max_{1\leq i\leq n_g}\|\hat{v}_{g,i}\|}{n^{\psi}} = o_p(1)$.

Assumption 4(i) restricts the cluster growth rate, while allowing growing and heterogeneous cluster sizes.[5] This condition is analogous to one introduced in Hansen and Lee (2019) and other clustering related literature, but is more restrictive due to covariate numerosity.

Assumption 4(ii) is a negligibility condition on least square residuals that restricts the distributional relationship between the finite dimensional covariate of interest $x_{g,i}$ and the high-dimensional nuisance covariate $w_{g,i}$. Analogously to CJN (2018b), we provide Lemma (A.6) in the Appendix as a way to formulate primitive conditions for it to hold. Under additional mild conditions, such as narrowing down the region for ψ and tightening up the moment requirement for $\|Q_g\|$, it is sufficient to have one of the following: (a) $n^{-2\psi}\sum_{g=1}^{G}\mathrm{E}(\|Q_g\|^2) = o(1)$, or (b) $\max_{1\leq i\leq n}\sum_{j=1}^{n}1\big(M_{ij}\neq 0\big) = o_p\big(n^{\frac{\psi(2+\theta)-1}{1+\theta}}\big)$, or (c) $\max_{1\leq i\leq n}\sum_{j=1}^{n}|H_{ij}|^{\frac{2+\theta}{1+\theta}} = o_p\big(n^{\frac{\psi(2+\theta)-1}{1+\theta}}\big)$, for some $\theta > 0$ such that $\psi(2+\theta) > 1$. Condition (a) tightens the requirement of auxiliary approximation in Assumption 3 to negligibility of the approximation error; as discussed above, it is plausible for some setups of interest. Condition (b) is a sparsity condition placed on the annihilation matrix, and it highlights a trade-off between the maximal cluster size and sparsity in the projection matrix, while condition (c) relaxes this strict sparsity requirement to approximate sparsity. For example, the annihilation matrix implied by the within transformation in a short panel data model with individual effects is sparse and does satisfy condition (b); in contrast, the annihilation

[4] In particular, it is sufficient to impose the condition $n^{-\vartheta}\max_{1\leq g\leq G} n_g^{-1}\|Q_g\|^2 \leq O_p(1)$ for some $\vartheta \geq 0$, which is mild and realistic, as attested by the extreme value theory (e.g., Leadbetter, Lindgren and Rootzén 1983). For example, for sub-Gaussian $Q_{g,i}$ implying sub-exponentially distributed cluster-specific scaled norms $n_g^{-1}\|Q_g\|^2$, their maximal value diverges at the rate not exceeding $O(\ln G)$, and this condition holds with any positive, however tiny, ϑ.

[5] The needfulness of such restrictions has been investigated and discussed in Sasaki and Wang (2022). In particular, Sasaki and Wang (2022) show that a violation of $n^{-1}\max_{1\leq g\leq G} n_g^2 = o(1)$ may lead to an asymptotic non-normal distribution. Subsequently, two practical solutions were proposed in Chiang, Sasaki, and Wang (2025): (i) score subsampling and (ii) size-adjusted reweighing.

matrix in a two-way fixed effect panel data model is dense, but the corresponding projection matrix can be shown to satisfy the approximate sparsity condition (c).

Further, additional conditions are required for the LCOC estimator to be consistent, similarly to the LOOC estimator from KSS (2020) and the approximate LOOC estimator from Jochmans (2020). We collect these conditions into Assumption 5.

Assumption 5 (LCOC variance estimation)

The following holds:

(i) $\Pr\big\{ \min_{1 \le g \le G} \det(\tilde{M}_{gg}) > 0 \big\} \to 1$, $\dfrac{1}{\min_{1 \le g \le G} \lambda_{\min}(\tilde{M}_{gg})} = O_p(1)$,

(ii) $\dfrac{\max_{1 \le g \le G} \max_{1 \le i \le n_g} |\mu_{g,i}|}{n^{1/4-\psi/2}} = O_p(1)$ for ψ of Assumption 4,

(iii) $\lambda_{\max}(\Omega) = O_p(1)$.

The conditions of Assumption 5(i) control for no perfect or near perfect collinearity. The first one allows the estimator to exist in a large sample with high probability, and the second one prevents the variance of the estimator to blow up in a large sample.[6] Assumption 5(ii), accompanied with the maximal eigenvalue bound of Assumption 5(iii),[7] is needed because the LCOC estimator uses $y_{g,i}$, which contains its conditional mean $\mu_{g,i}$, as a proxy for the unobserved error $u_{g,i}$. This condition restricts the amount of noise in $y_{g,i}$ that comes from $\mu_{g,i}$. Primitive conditions for such an assumption in terms of moment restrictions on the covariates are discussed in Jochmans (2020), which include boundedness of $\mathrm{E}[\|x_{g,i}\|^{\zeta}]$ for some $\zeta > 2$, similar moment restrictions for the covariate part, and possibly a sparsity restriction in the auxiliary equation. Our Assumption 5(ii) is stronger than one in Jochmans (2020), so such conditions carry over to higher moments, requiring instead $\zeta \ge 4/(1-2\psi)$, and possibly sharpening the sparsity restriction.

3.2 Asymptotic normality of OLS

Our first result establishes asymptotic normality of $\hat{\beta}$.

[6] Hansen (2025) lists some examples of cluster-wise invertibility failures (e.g., non-zeroness of a covariate in only one cluster). His version of the JK variance estimator uses a generalized inverse in case of cluster-wise non-invertibility. Hansen (2025) also notes that one of “almost unbiased” modifications (so called HC_2) of the LZ variance estimator in Stata can also be implemented using a generalized inverse, with an appropriate option. We conjecture, but do not pursue further, that using a generalized inverse can fruitfully modify the LCOC estimator in cases of cluster-wise invertibility failure.

[7] It is equivalent to $\max_{1 \le g \le G} \lambda_{\max}(\Omega_g) = O_p(1)$, which restricts within-cluster covariance structure beyond the moment condition in Assumption 2(ii), as $\lambda_{\max}(\Omega_g) \ge n_g^{-1} \sum_{i=1}^{n_g} \sum_{j=1}^{n_g} (\Omega_g)_{ij}$, which may grow with growing n_g.

Theorem 3.1. Suppose Assumptions 1-4 hold. Then,

$$\operatorname{var}\big(\hat{\beta}|\mathcal{X},\mathcal{W}\big)^{-1/2}\big(\hat{\beta}-\beta\big) \xrightarrow{d} \mathcal{N}(0, I_r).$$

Portnoy (1985) and Mammen (1993) provided an asymptotic normality result for the OLS coefficients in a linear regression model, where the number of parameters p (here, $p = r + k$) grows alongside the number of observations n with a restricted rate, $\lim_{n\to\infty} p^\delta/n = 0$ for some $\delta > 1$. Later, CJN (2018b) relaxed this requirement to $\lim_{n\to\infty} k/n < 1$ by focusing on a finite subset of OLS estimated coefficients. We build on the work of CJN (2018b) to provide an asymptotic normality result for a finite subset of OLS coefficients in a linear regression model with cluster dependence and many covariates.

While CJN (2018b) have shown that the normal approximation is appropriate for $\hat{\beta}$ in a linear regression model with many covariates and heteroskedasticity, the normal approximation may not be appropriate under clustering (Sasaki and Wang 2022). This was certainly worth examining even without the presence of many covariates (see Hansen 2007, Bester, Conley, and Hansen 2011, Sasaki and Wang 2022 and Chiang, Sasaki, and Wang 2025). Nonetheless, we have shown that the normal approximation remains appropriate, provided that growth rates of cluster sizes satisfy Assumption 4.

3.3 Properties of LCOC

Conducting inference using Theorem (3.1) requires a suitable estimator for $\operatorname{var}\big(\hat{\beta}|\mathcal{X},\mathcal{W}\big)$. Recall that the proposed LCOC estimator (2.3) estimates

$$\Sigma = \sum_{g=1}^{G} \hat{v}_g' \Omega_g \hat{v}_g,$$

which is the central matrix in the sandwich form of $\operatorname{var}\big(\hat{\beta}|\mathcal{X},\mathcal{W}\big)$.

The LCOC is a cluster extension of the leave-one-out crossfit (LOOC) estimator from Kline, Saggio and Sølvsten (2020), and shares its conditional unbiasedness property.

Lemma 3.1 (Unbiasedness)**.** Suppose $\hat{\Sigma}_{\text{LCOC}}$ exists. Then,

$$\mathrm{E}\big(\hat{\Sigma}_{\text{LCOC}}|\mathcal{X},\mathcal{W}\big) = \Sigma.$$

The following Theorem provides the consistency of the LCOC estimator.

Theorem 3.2 (Consistency)**.** Suppose Assumptions 1-5 hold. Then,

$$\hat{\Sigma}_{\text{LCOC}}^{-1}\Sigma \xrightarrow{p} I_r.$$

Having established the consistency of the LCOC estimator $\hat{\Sigma}_{\text{LCOC}}$, we can combine Theorem (3.2) with Theorem (3.1) to obtain the following corollary.

Corollary 3.3. Suppose Assumptions 1-5 hold. Then,

$$\widehat{\text{var}}_{\text{LCOC}}^{-1/2}\big(\hat{\beta}-\beta\big) \xrightarrow{d} \mathcal{N}(0, I_r),$$

where

$$\widehat{\text{var}}_{\text{LCOC}} = \Big(\sum_{g=1}^{G} \hat{v}_g' \hat{v}_g\Big)^{-1} \hat{\Sigma}_{\text{LCOC}} \Big(\sum_{g=1}^{G} \hat{v}_g' \hat{v}_g\Big)^{-1}.$$

This result allows for the construction of test statistics with correct size to conduct asymptotically valid inference.

4 Simulation Evidence

In this Section, we document results on the performance of the LCOC estimator along with that of other estimators, in three contexts. In one, we have a correctly specified linear regression; in the second, we deal with a partially linear model with a misspecified covariate structure. In the last experiment, we study a setup that emulates one of our empirical examples. In the first experiment, sample sizes are relatively large, while the last two exercises mimic situations with relatively smaller samples.

4.1 Correctly specified regression

The first simulation setup is inspired by that in Carter, Schnepel, and Steigerwald (2017):

$$y_{g,i} = \beta x_{g,i} + \gamma_0 + w_{g,i}'\gamma + u_{g,i},$$

where $\dim(x_{g,i}) = 1$, $\dim(w_{g,i}) = k-1$, and the regressors, covariates and errors are generated with a cluster structure as $x_{g,i} = z_g^x + z_{g,i}^x$ for z_g^x and $z_{g,i}^x$ i.i.d. standard normal, $w_{g,i} = z_g^w + z_{g,i}^w$ for z_g^w and $z_{g,i}^w$ i.i.d. $(k-1)$-variate standard normal, and $u_{g,i} = \sigma_g \epsilon_g + \sigma_{g,i}\eta_{g,i}$ for ϵ_g and $\eta_{g,i}$ i.i.d. standard normal. The conditional variances σ_g^2 and $\sigma_{g,i}^2$ are generated by

$$\sigma_g^2 \propto (z_g^x)^2 + (z_g^w)' z_g^w, \quad \sigma_{g,i}^2 \propto (z_{g,i}^x)^2 + (z_{g,i}^w)' z_{g,i}^w$$

with a proportionality coefficient of 25, which gives fairly strong heteroskedasticity. The parameter values are set at $\beta = 1$ and $\gamma_j = 0$ for all $j = 0, 1, ..., k-1$.

We have conducted the experiment using two designs, with the sample size $n = 2500$ and number of clusters $G = 100$ in both. In Design 1, all clusters are equally sized with $n_g = 25$ for all $g = 1, ..., 100$. In Design 2, the clusters are unbalanced in the sense that there is heterogeneity in cluster sizes: $n_1 = n_2 = n_3 = 1, n_4 = n_5 = 2, \ldots, n_{96} = n_{97} = 48, n_{98} = n_{99} = n_{100} = 49$ so that $\sum_{g=1}^{100} n_g = 2500$. In both designs, we repeated the experiment with a different number of covariates in order to achieve a wide variety of the k/n ratio. Three variance estimates – LZ, LCOC and JK – are compared to the simulated variance of the OLS estimator, and also to the "True" infeasible "oracle" estimator that utilizes true errors instead of residuals. Tables 1-2 and figures 1-2 report results from at least 1,000 simulation runs.

Table 1 contains figures for the ratio of the standard deviation of the simulated OLS estimator to the average standard error implied by of each of the three estimators (i.e., their average square roots). In both designs, the LZ estimator exhibits a downward bias, which tends to become stronger, not monotonically though, as we increase k/n; in contrast, the JK estimator exhibits an upward bias, which also increases with k/n. Such biasedness takes off more slowly for JK than for LZ with an increase in k. For example, when $k = 256$, a pretty impressive number, the downward LZ bias exceeds 60% on the standard deviation scale, while the upward JK bias amounts to about 10%; but when k reaches $k = 2048$, both standard deviation/standard error ratios are around 2.4 and 2.4^{-1}, respectively. At the same time, consistent with the theory, the LCOC estimates are nearly unbiased in both designs for all degrees of covariate numerosity.

Table 2 reports figures for the actual coverage rates for nominally 95%-level confidence intervals. They show how the estimation biases translate to inferential performance. One can see that the problem of severe undercoverage rises quickly with k for LZ, though not so fast for JK; however, for the highest k/n ratio JK's coverage reaches 100%. The LCOC estimator exhibits, equally in both designs and for all k, very slight undercoverage, which, we conjecture, is due to deviations from the asymptotic normality of the OLS estimator in finite samples.

We continue examining how estimation biases translate to inferential performance by investigation of 2.5%-level two-sided power curves for the null $H_0 : \beta = 1$. They are shown in Figure 1 for the four estimates (LZ, LCOC, JK, True) and $\frac{k}{n} \in \{\frac{4}{2500}, \frac{16}{2500}, \frac{64}{2500}, \frac{256}{2500}, \frac{1024}{2500}, \frac{2048}{2500}\}$ in Design 1. As expected, when the number of parameters is very low relative to the sample size, all three tests perform extremely closely to the oracle. However, as k/n increases to $\frac{16}{2500}$, the LZ test statistic starts to over-reject, even with this still relatively low value of k/n, while JK and LCOC retain near oracle performance. As k/n increases further to $\frac{64}{2500}$, still a sufficiently low value, even the JK test statistic starts to lose power for the values of β further away from the null, and the over-rejection by LZ continues to deteriorate. At this point, only LCOC remains at a near-oracle performance. This pattern continues all the way to $\frac{k}{n} = \frac{2048}{2500}$, where LZ strongly over-rejects, while JK does not reject ever at all. The LCOC estimator stays close to the oracle counterpart across all values of k/n.

Figure 2 shows power curves for Design 2, which features unbalanced clusters. Note that here, $\max_{1 \le g \le G} n_g^2/n \approx 0.96 < 1$, which falls into the second lowest bin of Table 1 in Chiang, Sasaki, and Wang (2025), so cluster heterogeneity is not too extreme from an empirical standpoint. The pattern displayed in Figure 1 appears not to change for Design 2. The LCOC estimator continues to retain its near-oracle performance, while the other test statistics eventually fail.

Overall, the results suggest that the LCOC estimator can be an invaluable tool to carry out robust inference in overfitted models, where many irrelevant covariates are included, and the error variances depend on at least some of these irrelevant covariates. The unbiasedness property of LCOC translates well into good finite sample performance, as its power curves stay close to the oracle counterparts across all specifications.

4.2 Partially linear model

In this set of experiments, we employ the partially linear model, where the structure of the covariate part is misspecified, which induces endogeneity. The simulation setup is the following:

$$y_{g,i} = \beta x_{g,i} + \exp(-|z_{g,i}|) + \xi_{g,i}, \quad x_{g,i} = h(z_{g,i}) + V_{g,i},$$

where $z_{g,i}$ are i.i.d. $U[-1,1]$, $\xi_{g,i} = \epsilon_g^\xi + 5\sqrt{x_{g,i}^2 + z_{g,i}^2}\,\eta_{g,i}^\xi$, $V_{g,i} = \epsilon_g^V + \eta_{g,i}^V$, and $\epsilon_g^\xi, \epsilon_g^V, \eta_{g,i}^\xi, \eta_{g,i}^V$ are all i.i.d. standard normal. The true parameter is $\beta = 1$. We consider two specifications for the auxiliary equation for $x_{g,i}$:

1. $h(z_{g,i}) = \exp(|z_{g,i}|)$,
2. $h(z_{g,i}) = 1 + |z_{g,i}| + \frac{1}{2}|z_{g,i}|^2 + \frac{1}{6}|z_{g,i}|^3$.

In the first, exponential specification, the conditional mean of $x_{g,i}$ depends on $z_{g,i}$ nonlinearly, which induces "dense" dependence on its powers, while in the second specification it depends on $z_{g,i}$ in a "sparse" fashion, only through up to its third power.

We estimate the following regression, linear in powers of $|z_{g,i}|$:

$$y_{g,i} = \beta x_{g,i} + \sum_{j=0}^{k-1} \gamma_j |z_{g,i}|^j + u_{g,i},$$

where k is dependent on the sample size n, increasing to infinity with it, although more slowly. Here, the exogeneity condition is violated as

$$\mathrm{E}(u_{g,i}|x_{g,i}, z_{g,i}) = \mathrm{E}\big(\exp(-|z_{g,i}|) - \sum_{j=0}^{k-1} \gamma_j |z_{g,i}|^j \,|x_{g,i}, z_{g,i}\big) \neq 0.$$

Nonetheless, the bias vanishes asymptotically as $\exp(-|z_{g,i}|) - \lim_{n\to\infty} \sum_{j=0}^{k-1} \gamma_j |z_{g,i}|^j = 0$ for the coefficients $\{\gamma_j\}_{j=0}^{\infty}$ in a Taylor expansion of the exponential function.

We consider three different feasible inference methods for the OLS estimator – LZ, LCOC and JK. Also, we compare the OLS parameter estimates to the double debiased LASSO

(DDB-LASSO) estimates (see, e.g., Chernozhukov et al. 2018 and Chiang, Kato, Ma and Sasaki 2022) in order to study the finite sample bias caused by the model misspecification. We use a sample size of $n = 250$ and balanced $G = 25$ clusters. Note that these figures are much smaller than those in subsection 4.1, and the results are expected to be more susceptible to small-sample distortions. We set k to values in the set $\{5, 10, 20, 40, 80, 160\}$, so that the ratios k/n are comparable to those in subsection 4.1. The simulation results from 10,000 simulation runs are collected in Tables 3, 4 and 5.

Table 3 reports, for each k, for both specifications, and for the three variance estimators, figures for the ratio of the standard deviation of the simulated OLS estimator to the average standard error. Table 4 shows corresponding figures for the actual coverage rates for 95%-level confidence intervals. The LCOC and JK estimators exhibit stability in both biasedness and coverage features as the number of covariates increases, while the LZ estimator's performance slowly deteriorates with k. The LCOC estimator is almost exactly unbiased, despite, theoretically, it is not conditionally unbiased because of model misspecification. At the same time, JK exhibits an upward bias of about 5% on the standard deviation scale. The use of both these estimators, however, leads to small but persistent undercoverage, which for the unbiased LCOC estimator presumably results from the t-statistic's small-sample deviations from normality. In contrast to LCOC and JK, the LZ estimator is biased downwards and to a larger degree than JK, with the bias increasing with k from 7% to 11% on the standard deviation scale; a similar tendency is shown by the degree of undercoverage. The notable closeness of all the figures in the left and right panels of Tables 3 and 4 corresponding to the presence and absence of misspecification in the auxiliary equation shows invariance, at least in this example, to whether the auxiliary equation is correctly or misspecified.

Next, we look at average parameter estimates from the OLS and DDB-LASSO estimation documented in Table 3. The OLS estimates for all values of k are extremely close to the true value in the exponential specification. The DDB-LASSO estimates can be biased in both directions, strongly depending on λ. This evidence reveals its finite sample bias problems not only in dense setups but even in sparse structures, supporting the findings in Wüthrich and Zhu (2023). In contrast, OLS appears to be almost unbiased with the estimates extremely close to 1 across all values of k. The overall results seem to suggest that OLS is a good alternative to double debiased LASSO, at least when the underlying nonlinear covariate part can be well approximated by a polynomial function.

4.3 Emulating Donohue and Levitt (2001)

We consider a setup that emulates one of our empirical examples in Section 5, Donohue and Levitt (2001):

$$y_{it} = \beta x_{it} + w_{it}'\gamma_t + u_{it},$$

where $\beta = 1$, $\dim(x_{it}) = 1$, $\dim(w_{it}) = 9$, (x_{it}, w_{it}) are i.i.d. 10-variate standard normal, and, depending on the specification,

(i) $\gamma_t = \gamma \sim_{iid} U[-0.5, 0.5]$ for $t = 1, \ldots, T$ and $u_{it} = (0.8\epsilon_{it-1} + 0.2\epsilon_{it})|x_{it}|$,

(ii) $\gamma_t \sim_{iid} U[-0.5, 0.5]$ for $t = 1, \ldots, T$ and $u_{it} = (0.8\epsilon_{it-1} + 0.2\epsilon_{it})|x_{it}|$,

(iii) $\gamma_t = \gamma \sim_{iid} U[-0.5, 0.5]$ for $t = 1, \ldots, T$ and $u_{it} = (0.8\epsilon_{it-1} + 0.2\epsilon_{it})\sqrt{x_{it}^2 + w_{it}'w_{it}}$,

(iv) $\gamma_t \sim_{iid} U[-0.5, 0.5]$ for $t = 1, \ldots, T$ and $u_{it} = (0.8\epsilon_{it-1} + 0.2\epsilon_{it})\sqrt{x_{it}^2 + w_{it}'w_{it}}$,

where ϵ_{it} are i.i.d. standard normal. In specification (i), coefficients on the controls are constant across time, while in specification (ii), they are time varying. In both specifications (i) and (ii), the skedastic function depends on x_{it} only. Specifications (iii) and (iv) differ from (i) and (ii), respectively, only in that the skedastic function now depends on both x_{it} and w_{it}. Note that when the control parameters are time varying, the number of control parameters is $k = 180$, large compared to $k = 9$ when they are constant. As in Donohue and Levitt (2001), we set $G = 45$ clusters with $n_g = 13$ observations in each, so the sample size is $n = 624$.

The results from 10,000 simulation runs are collected in Table 6, the standard deviation/average standard error ratios on the left side, and the actual coverage rates on the right side. One can observe similar patterns of relative performance of different variance estimators to those observed in subsection 4.2. In particular, the LCOC estimator is unbiased in all specifications, the LZ estimator is biased downward, and the JK estimator is biased upward, these biases of LZ and JK increasing with the number of covariates. The use of the LCOC estimator leads to slight small-sample undercoverage, which does not vary much with the specification; coverage distortions of JK are minimal, while those for LZ are quite serious when covariates are many. Additionally, the form of heteroskedasticity does not influence the performance of LCOC, while its higher complexity (i.e., when it depends on both the main regressors and control covariates) does have some adverse impact on the biasedness of the JK estimator. Overall, and similarly to the results reported in the two previous subsections, theoretical properties of LCOC translate to its practical reliability in unbiasedness and good coverage, irrespective of how numerous control covariates are and how complex the pattern of heteroskedasticity is.

5 Empirical Illustrations

5.1 Angrist and Lavy (2009)

Angrist and Lavy (2009), AL hereafter, analyze the effect of high stakes high school achievement using a cash incentives experiment. One of their identifications of the treatment effect is given by the following linear model:

$$y_{ij} = \beta x_j + z_j'\alpha + w_i'\gamma + \sum_{q\in\{2,3,4\}} d_{qi}\delta_q + \epsilon_{ij}, \quad \mathrm{E}(\epsilon_{ij}|x_j, z_j, w_i, d_{qi}) = 0,$$

where i indexes students, j indexes schools, y_{ij} is the Bagrut status, and x_j is the treatment dummy (school level). Covariates include a vector of school-level controls z_j, a vector of individual controls w_i, and indicators d_{qi}, $q \in \{2, 3, 4\}$ corresponding to a quartile of a student's credit-unit-weighted average test score in tests taken before January 2001. Clusters are defined at the school level. The pooled sample size is $n = 3821$ with $G = 39$ clusters (cluster size varying from 9 to 248); for the Girls subsample, $n = 1861$ and $G = 34$ (cluster size varying from 12 to 146); and for the Boys subsample, $n = 1960$ and $G = 34$ (cluster size varying from 1 to 141).

AL (2009) assert that the LZ standard error is biased downward, and use Bell and Mcaffrey's (2009) jackknife (JK) estimator in an attempt to address the bias problem of LZ. Two potential concerns here are (a) that there is no guarantee that the LZ bias is downward, as clustered observations can be dependent in a complex way, and (b) that the JK standard errors are also biased in a finite sample. Thus, application of the crossfit estimator to their models could serve as a robustness check to alleviate the aforementioned concerns.

We reproduce AL (2009, Panel A of Table 2)'s results in Table 5, which contains point estimates of β, two types of their standard errors, and ratios of parameter dimensionality to sample size; the square brackets show p-values for statistical significance of β against a right-sided alternative. Two causal discoveries were made by AL (2009) from this table: (1) there is evidence that the Achievement Awards program increased Bagrut rates in 2001, and (2) the estimated treatment effect comes mainly from girls as suggested by the gender-specific regressions. However, AL (2009) do note that most of the significant results are only "marginally significant." A close examination of the implied p-value with the JK standard error seems to suggest that AL (2009) mean a significance level around or below 10%.

Inferences based on the LCOC estimator support their conclusion with some additional insights. The two key specifications in this table are SC+Q+M and SC+Q+M+P for the full sample. The first one corresponds to the model that controls for school covariates SC, quartile effect Q and micro covariates M. The second one includes all the aforementioned covariates along with the pair effect P. The coefficient p-values for these two equations are significantly lower than their jackknife counterparts: 0.0885 vs 0.1072 and 0.0274 vs 0.1055, respectively. Hence, using the LCOC inference strengthens two of their marginally significant results to significant at the 10% and 5% levels, respectively. This is consistent with what we observed in the Monte Carlo experiments. The JK estimator tends to under-reject when the true value is not near the null, which is certainly possible here, as γ may not be exactly zero. Furthermore, the ratio of squared maximum school size to sample size is $\frac{248^2}{3821} \approx 16.1$. The cluster heterogeneity in this sample is therefore even stronger than in our main simulation experiment, which can potentially lead to a lower power for the JK-based test statistics. Hence, the test statistic based on the LCOC estimator may be superior to one based on JK even when the number of parameters is relatively low as in this empirical example.

5.2 Donohue and Levitt (2001)

Donohue and Levitt (2001), DL hereafter, conclude that legalized abortion has contributed significantly to crime reduction. For three crime categories – violent crime, property crime and murder, – they run the following regression:

$$\ln(\text{crime}_{st}) = \beta \cdot \text{abort}_{st} + w_{st}'\gamma + \theta_s + \lambda_t + \epsilon_{st},$$

where $s = 1, \ldots, S$ designates state, $t = 1, \ldots, T$ designates year, the left-hand-side variable is the logged crime rate per capita; abort_{st} is the effective abortion rate for a given state and year; w_{st} is a vector of state-level controls that includes prisoners and police count per capita, a range of variables capturing state economic conditions, lagged state welfare generosity, the presence of concealed handgun laws, and per capita beer consumption; θ_s and λ_t represent state and year fixed effects. Clusters are defined at the state level. Clustering at the state level can be seen as a way to account for serial correlation in the sample. However, the cluster-robust standard errors are biased due to the presence of serial correlation. Assuming the textbook asymptotic setting, the LCOC estimator can be seen as a finite sample bias correction to the cluster-robust standard error. Furthermore, as we have shown in our emulation experiment in subsection 4.3, the LZ test statistic over-rejects when the model or heteroskedasticity pattern is complex. Therefore, the LZ estimator used in the original study is unsuitable for conducting sensitivity analysis, and the LCOC estimator may serve as an additional tool to reaffirm the robustness of the original study.

The state-specific fixed effects are not identified in the leave-cluster-out sample, so we get around this by applying a within transformation to get rid of the state-specific fixed effects. The baseline model is thus

$$\ln(\widetilde{\text{crime}}_{st}) = \beta \cdot \widetilde{\text{abort}}_{st} + \widetilde{w}_{st}'\gamma + \widetilde{\lambda}_t + \widetilde{\epsilon}_{st},$$

where $\widetilde{\text{abort}}_{st} = \text{abort}_{st} - T^{-1}\sum_{\tau=1}^{T} \text{abort}_{s\tau}$, and the other "tilded" variables are defined similarly. We estimate the model above using OLS and compute the LZ, LCOC and JK estimators. Table 6 contains point estimates of β, three types of their standard errors, and ratios of parameter dimensionality to sample size; the square brackets show p-values for statistical significance of β against left-sided alternative.

The point estimates documented in Table 6 are essentially a replication of Table IV in DL (2001). In line with their results, the coefficient estimate $\hat{\beta}$ is negative for all crime types. In terms of estimated variances, we see that the LZ estimator is the smallest across the board, while the JK estimator is the largest across the board. The LCOC estimator is exactly in the middle across the board. In terms of significance level, nothing is changed when we switch from LZ to LCOC. However, the coefficient for murder crime is no longer significant at the 1% level when we switch from LZ to JK. This suggests that the result of DL (2001) may be less robust for more serious crimes. Overall, the original results remain relatively insensitive to the choice of the variance estimator used.

Next, we consider a high dimensional specification that assumes the impact of controls is time-varying. The model is given by

$$\widetilde{\ln(\text{crime}_{st})} = \beta \cdot \widetilde{\text{abort}}_{st} + \widetilde{w}'_{st}\gamma_t + \widetilde{\lambda}_t + \widetilde{\epsilon}_{st}.$$

Note the time index of γ_t. The ratio of parameter numerosity to a number of observations is equal to $\frac{109}{624} \approx 17.5\%$, which is much larger than the ratio of $\frac{21}{624} \approx 3.4\%$ in the baseline model.

The estimate $\hat{\beta}$ of the coefficient of interest in this specification remains negative for all three crime types, while their magnitudes are reduced by a non-trivial amount. This highlights the sensitivity to the choice of controls and supports the finding of Belloni, Chernozhukov and Hansen (2014). The estimated variances are larger than the baseline ones across the board with, again, JK (LZ) being the largest (smallest) across the board. However, $\hat{\beta}$ is only highly significant for property crime across the three estimators. Violent crime is no longer significant at the 1% level with both LCOC and JK. Murder crime is no longer significant at the 10% level with JK but stays significant at the 10% level with both LZ and LCOC. This casts doubts on whether there is indeed a causal relationship between the abortion rate and more serious crimes like murder and violent crimes.

6 Conclusion

This paper established inference results for the OLS estimator of a subset of coefficients in linear regression models with many covariates, where observations are clustered with heterogeneous and growing cluster sizes. The proposed LCOC method may turn to be useful in other regression setups with many regressors and heteroskedasticity; for example, in adaptation of testing for many restrictions (Anatolyev and Sølvsten 2023), in case the regression errors are clustered. It could even potentially be applied to settings outside of clustered sampling if additional boundary conditions are met (see Leung 2023). Future researchers might also want to look at other crossfitting schemes given how powerful crossfitting technique is.

Acknowledgements

We would like to thank the Editor and Associate Editor for helpful guidance, and two anonymous referees for their insightful suggestions, which greatly improved the presentation. The first author gratefully acknowledges financial support by the grant 24-12720S from the Czech Science Foundation.

References

Anatolyev, S. (2012). Inference in regression models with many regressors. *Journal of Econometrics*, 170(2):368–382.

Anatolyev, S. (2019). Many instruments and/or regressors: A friendly guide. *Journal of Economic Surveys*, 33(2):689–726.

Anatolyev, S. and Sølvsten, M. (2023). Testing many restrictions under heteroskedasticity. *Journal of Econometrics*, 236(1):105473.

Angrist, J. and Lavy, V. (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *American Economic Review*, 99(4):1384–1414.

Arellano, M. (1987). Practitioners corner: Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, 49(4):431–434.

Bell, R. and McCaffrey, D. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, 28:169–181.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50.

Bester, C. A., Conley, T. G., and Hansen, C. B. (2011). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics*, 165(2):137–151.

Calhoun, G. (2011). Hypothesis testing in linear regression when k/n is large. *Journal of Econometrics*, 165(2):163–174.

Cameron, A. and Miller, D. (2015). A practitioners guide to cluster-robust inference. *Journal of Human Resources*, 50:317–372.

Canay, I. A., Santos, A., and Shaikh, A. M. (2021). The wild bootstrap with a “small” number of “large” clusters. *The Review of Economics and Statistics*, 103(2):346–363.

Carter, A. V., Schnepel, K. T., and Steigerwald, D. G. (2017). Asymptotic behavior of a t-test robust to cluster heterogeneity. *Review of Economics and Statistics*, 99(4):698–709.

Cattaneo, M. D., Jansson, M., and Newey, W. K. (2018a). Alternative asymptotics and the partially linear model with many regressors. *Econometric Theory*, 34:277–301.

Cattaneo, M. D., Jansson, M., and Newey, W. K. (2018b). Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association*, 113(523):1350–1361.

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

Chesher, A. and Austin, G. (1991). The finite-sample distributions of heteroskedasticity robust Wald statistics. *Journal of Econometrics*, 47(1):153–173.

Chesher, A. and Jewitt, I. (1987). The bias of a heteroskedasticity consistent covariance matrix estimator. *Econometrica*, 55(5):1217–1222.

Chiang, H. D., Kato, K., Ma, Y., and Sasaki, Y. (2022). Multiway cluster robust double/debiased machine learning. *Journal of Business & Economic Statistics*, 40(3):1046–1056.

Chiang, H. D., Sasaki, Y., and Wang, Y. (2025). Genuinely robust inference for clustered data. Technical report. arXiv 2308.10138.

Cochran, W. G. (1977). *Sampling Techniques, 3rd Edition.* John Wiley.

Currie, J. and Gruber, J. (1996). Health insurance eligibility, utilization of medical care, and child health. *Quarterly Journal of Economics*, 111(2):431–466.

D'Adamo, R. (2018). Cluster-robust standard errors for linear regression models with many controls. Technical report. arXiv 1806.07314.

Donohue, John J., I. and Levitt, S. D. (2001). The impact of legalized abortion on crime. *Quarterly Journal of Economics*, 116(2):379–420.

Gong, A. (2022). *Essays in Theoretical and Applied Econometrics.* PhD thesis, University of Michigan, Ann Arbor, MI.

Hansen, B. E. (2025). Jackknife standard errors for clustered regression. Manuscript, University of Wisconsin.

Hansen, B. E. and Lee, S. (2019). Asymptotic theory for clustered samples. *Journal of Econometrics*, 210(2):268–290.

Hansen, C. B. (2007). Asymptotic properties of a robust variance matrix estimator for panel data when T is large. *Journal of Econometrics*, 141(2):597–620.

Hanushek, E. A., Kain, J. F., Markman, J. M., and Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics*, 18(5):527–544.

Heckman, J. J. and Snyder, J. M. (1997). Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators. *Rand Journal of Economics*, 28(0):S142–S189.

Horrace, W. C. and Oaxaca, R. L. (2003). Cross-fitting and fast remainder rates for semiparametric estimation. IZA Discussion Paper, Institute of Labor Economics.

Huber, P. J. (2011). *Robust Statistics*, pages 1248–1251. Springer, Berlin, Heidelberg.

Ibragimov, R. and Müller, U. K. (2010). t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics*, 28(4):453–468.

Ibragimov, R. and Müller, U. K. (2016). Inference with few heterogeneous clusters. *Review of Economics and Statistics*, 98(1):83–96.

Imbens, G. W. and Kolesár, M. (2016). Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics*, 98(4):701–712.

Jochmans, K. (2020). Heteroscedasticity-robust inference in linear regression models with many covariates. *Journal of the American Statistical Association*, 117:1–10.

Kline, P., Saggio, R., and Sølvsten, M. (2020). Leave-out estimation of variance components. *Econometrica*, 88(5):1859–1898.

Kolesár, M. (2013). Estimation in an instrumental variables model with treatment effect heterogeneity. Manuscript, Princeton University.

Leadbetter, M., Lindgren, G., and Rootzén, H. (1983). *Extremes and related properties of random sequences and processes*. Springer Series in Statistics. Springer, Germany.

Leung, M. P. (2023). Network cluster robust inference. *Econometrica*, 91(2):641–667.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

MacKinnon, J. G., Nielsen, M. Ø., and Webb, M. D. (2023). Leverage, influence, and the jackknife in clustered regression models: Reliable inference using summclust. *Stata Journal*, 23(4):942–982.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statistics*, 21(1):255–285.

Newey, W. K. and Robins, J. M. (2017). Cross-fitting and fast remainder rates for semiparametric estimation. CeMMAP Working Paper, Institute for Fiscal Studies.

Portnoy, S. (1985). Asymptotic behavior of m estimators of p regression parameters when p^2/n is large; II. Normal approximation. *Annals of Statistics*, 13(4):1403–1417.

Postel-Vinay, F. and Robin, J.-M. (2002). Equilibrium wage dispersion with worker and employer heterogeneity. *Econometrica*, 70(6):2295–2350.

Rogers, W. (1994). Regression standard errors in clustered samples. *Stata Technical Bulletin*, 3(13).

Sasaki, Y. and Wang, Y. (2025). Non-robustness of the cluster-robust inference: with a proposal of a new robust method. Technical report. arXiv 2210.16991.

Verdier, V. (2020). Estimation and inference for linear models with two-way fixed effects and sparsely matched data. *Review of Economics and Statistics*, 102(1):1–16.

White, H. (1984). *Asymptotic Theory for Econometricians*. Academic Press.

Wooldridge, J. M. (2001). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

Wüthrich, K. and Zhu, Y. (2023). Omitted variable bias of lasso-based inference methods: A finite sample analysis. *Review of Economics and Statistics*, 105(4):982–997.

Young, A. (2018). Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *Quarterly Journal of Economics*, 134(2):557–598.

A Appendix: proofs

A.1 Useful lemmas

This section contains a few lemmas that are used in proofs of main results. Let $\dim(\beta) = 1$ to ease notation, without a loss of generality.

The first lemma relates the leave-cluster-out (LCO) residuals to the OLS residuals.

Lemma A.1. Provided that $\tilde{M}_{gg}^{-1}$ exists,

$$(\hat{u}_{-g})_g = \tilde{M}_{gg}^{-1}\hat{u}_g.$$

Proof. Let us look at $(\hat{u}_{-g})_g = y_g - \tilde{X}_g\big(\tilde{X}'_{-g}\tilde{X}_{-g}\big)^{-1}\tilde{X}'_{-g}y_{-g}$. Using the Woodbury matrix identity,

$$\big(\tilde{X}'_{-g}\tilde{X}_{-g}\big)^{-1} = \big(\tilde{X}'\tilde{X} - \tilde{X}'_g\tilde{X}_g\big)^{-1} = \big(\tilde{X}'\tilde{X}\big)^{-1} + \big(\tilde{X}'\tilde{X}\big)^{-1}\tilde{X}'_g\tilde{M}_{gg}^{-1}\tilde{X}_g\big(\tilde{X}'\tilde{X}\big)^{-1},$$

hence

$$\tilde{X}_g\big(\tilde{X}'_{-g}\tilde{X}_{-g}\big)^{-1} = \tilde{X}_g\big(\tilde{X}'\tilde{X}\big)^{-1} + \big(I_{n_g} - \tilde{M}_{gg}\big)\tilde{M}_{gg}^{-1}\tilde{X}_g\big(\tilde{X}'\tilde{X}\big)^{-1} = \tilde{M}_{gg}^{-1}\tilde{X}_g\big(\tilde{X}'\tilde{X}\big)^{-1},$$

and so

$$\begin{aligned}(\hat{u}_{-g})_g &= y_g - \tilde{M}_{gg}^{-1}\tilde{X}_g\big(\tilde{X}'\tilde{X}\big)^{-1}\big(\tilde{X}'y - \tilde{X}'_g y_g\big)\\ &= y_g - \tilde{M}_{gg}^{-1}\tilde{X}_g\big(\tilde{X}'\tilde{X}\big)^{-1}\tilde{X}'y + \tilde{M}_{gg}^{-1}\big(I_{n_g} - \tilde{M}_{gg}\big)y_g\\ &= \tilde{M}_{gg}^{-1}\big(y_g - \tilde{X}_g\big(\tilde{X}'\tilde{X}\big)^{-1}\tilde{X}'y\big) = \tilde{M}_{gg}^{-1}\hat{u}_g.\end{aligned}$$

□

The second lemma relates the LCO residuals to the true errors.

Lemma A.2.

$$(\hat{u}_{-g})_g = \tilde{M}_{gg}^{-1}\tilde{M}_{g,-g}u_{-g} + u_g,$$

where $\tilde{M}_{g,-g}$ is the submatrix of $\tilde{M}_g$ after omitting the columns related to cluster g.

Proof. Using Lemma (A.1),

$$\begin{aligned}(\hat{u}_{-g})_g &= \tilde{M}_{gg}^{-1}\hat{u}_g\\ &= \tilde{M}_{gg}^{-1}\tilde{M}_g u\\ &= \tilde{M}_{gg}^{-1}(\tilde{M}_{g,-g}u_{-g} + \tilde{M}_{gg}u_g)\\ &= \tilde{M}_{gg}^{-1}\tilde{M}_{g,-g}u_{-g} + u_g.\end{aligned}$$

□

The third lemma shows that the sample variance of the OLS residuals in the auxiliary model is never too close to zero in probability.

Lemma A.3.

$$\Big(n^{-1}\sum_{g=1}^{G}\|\hat{v}_g\|^2\Big)^{-1} = O_p(1).$$

Proof. This is SA-1 from CJN (2018b)'s appendix. The condition $\max_{1\le g\le G} n_g^2/n = o(1)$ implied by our Assumption 4(i) is equivalent to assuming $\mathcal{C}^2_{\mathcal{T},n}/n = o(1)$ in CJN (2018b). Plugging this result back into their proof allows us to adopt their result under the setting of growing cluster sizes. □

The fourth lemma shows that the sample variance of the OLS residuals in the auxiliary model is bounded in probability.

Lemma A.4.

$$\frac{1}{n}\sum_{g=1}^{G}\|\hat{v}_g\|^2 = O_p(1).$$

Proof. Observe that

$$\begin{aligned}\frac{1}{n}\sum_{g=1}^{G}\|\hat{v}_g\|^2 &\le \frac{1}{n}\sum_{g=1}^{G}\|v_g\|^2 = \frac{1}{n}\sum_{g=1}^{G}\|Q_g + V_g\|^2\\ &\le \frac{2}{n}\sum_{g=1}^{G}\|Q_g\|^2 + \frac{2}{n}\sum_{g=1}^{G}\|V_g\|^2 = O_p(1).\end{aligned}$$

where the first inequality follows from the optimality of least squares. The second inequality follows from $(a+b)^2 \le 2(a^2+b^2)$, and Assumptions 2(iii), 3 and 4(i). Indeed, Assumption 3 bounds the first term, while the bound of the second term comes from the combination of the Cauchy-Schwarz inequality and Assumptions 2(iii) and 4(i), as

$$\mathrm{E}\Big(\frac{1}{n}\sum_{g=1}^{G}\|V_g\|^2\Big) \le \max_{1\le g\le G}\max_{1\le i\le n_g}\mathrm{E}\big(V_{g,i}^4|\mathcal{W}\big)^{1/2}\,\frac{1}{n}\sum_{g=1}^{G}n_g = O_p(1)\times 1 = O_p(1)$$

and

$$\begin{aligned}var\Big(\frac{1}{n}\sum_{g=1}^{G}\|V_g\|^2\Big) &= \frac{1}{n^2}\sum_{g=1}^{G}var\big(\|V_g\|^2\big) < \frac{1}{n^2}\sum_{g=1}^{G}\mathrm{E}\big(\|V_g\|^4\big) = \frac{1}{n^2}\sum_{g=1}^{G}\mathrm{E}\Big(\big(\sum_{i=1}^{n_g}V_{g,i}^2\big)^2\Big)\\ &\le \frac{1}{n^2}\sum_{g=1}^{G}\mathrm{E}\Big(n_g\sum_{i=1}^{n_g}V_{g,i}^4\Big) \le \max_{1\le g\le G}\max_{1\le i\le n_g}\mathrm{E}\big(V_{g,i}^4|\mathcal{W}\big)\,\frac{\max_{1\le g\le G}n_g}{n}\,\frac{1}{n}\sum_{g=1}^{G}n_g\\ &= O_p(1)o_p(1)\times 1 = o_p(1).\end{aligned}$$

□

The fifth lemma describes the block structure of matrix $\tilde{M}\Omega\tilde{M}$. This lemma will be useful in establishing consistency of the LCOC estimator.

Lemma A.5.

$$\tilde{M}\Omega\tilde{M} = \begin{pmatrix} \sum_{i=1}^{n_g} \tilde{M}_{g_1 g_i}\Omega_{g_i}\tilde{M}_{g_i g_1} & \sum_{i=1}^{n_g} \tilde{M}_{g_1 g_i}\Omega_{g_i}\tilde{M}_{g_i g_2} & \cdots & \sum_{i=1}^{n_g} \tilde{M}_{g_1 g_i}\Omega_{g_i}M_{g_i g_{n_g}} \\ \sum_{i=1}^{n_g} \tilde{M}_{g_2 g_i}\Omega_{g_i}\tilde{M}_{g_i g_1} & \ddots & \cdots & \vdots \\ \vdots & \ddots & \cdots & \vdots \\ \sum_{i=1}^{n_g} \tilde{M}_{g_{n_g} g_i}\Omega_{g_i}\tilde{M}_{g_i g_1} & \sum_{i=1}^{n_g} \tilde{M}_{g_{n_g} g_i}\Omega_{g_i}\tilde{M}_{g_i g_2} & \cdots & \sum_{i=1}^{n_g} \tilde{M}_{g_{n_g} g_i}\Omega_{g_i}\tilde{M}_{g_{n_g} g_{n_g}} \end{pmatrix}$$

Proof. Straightforward. □

Analogous to CJN (2018b)'s SA7, the sixth lemma establishes the primitive conditions of Assumption 4(ii). Recall that $H = I - M$.

Lemma A.6. Suppose Assumption 1 to Assumption 4 hold, ψ of Assumption 4 belongs to the region $\frac{3}{10} < \psi < \frac{1}{2}$, and additionally

$$\frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g} \mathrm{E}(|Q_{g,i}|^{2+\theta}) = O(1)$$

for $\theta > 0$ such that $\psi(2+\theta) > 1$. Suppose one of the following conditions holds:

(a) $n^{-2\psi}\sum_{g=1}^{G}\mathrm{E}(|Q_g|^2) = o(1)$, or

(b) $\max_{1\le i\le n}\sum_{j=1}^{n} 1\big(M_{ij} \neq 0\big) = o_p\big(n^{\frac{\psi(2+\theta)-1}{1+\theta}}\big)$, or

(c) $\max_{1\le i\le n}\sum_{j=1}^{n} |H_{ij}|^{\frac{2+\theta}{1+\theta}} = o_p\big(n^{\frac{\psi(2+\theta)-1}{1+\theta}}\big)$.

Then,

$$\frac{\max_{1\le g\le G}\max_{1\le i\le n_g}|\hat{v}_{g,i}|}{n^{\psi}} = o_p(1).$$

In Lemma (A.6), condition (b) can be interpreted as a sparsity condition on the projection matrix M, in addition to the restriction on higher moments of Q_i, while condition (c) relaxes sparsity to approximate sparsity. In contrast to CJN (2018b), we allow $\max_{1\le g\le G}\max_{1\le i\le n_g}|\hat{v}_{g,i}|$ to diverge at a slower rate. A smaller ψ ultimately implies a stronger sparsity or approximate sparsity condition, and stronger moment conditions on Q_i are required.

Proof. Recall $\hat{v}_{g,i} = \tilde{V}_{g,i} + \tilde{Q}_{g,i}$ and

$$\frac{\max_{1\le g\le G}\max_{1\le i\le n_g}|\hat{v}_{g,i}|}{n^{\psi}} \le \frac{\max_{1\le g\le G}\max_{1\le i\le n_g}|\tilde{V}_{g,i}|}{n^{\psi}} + \frac{\max_{1\le g\le G}\max_{1\le i\le n_g}|\tilde{Q}_{g,i}|}{n^{\psi}}.$$

First note that

$$\begin{aligned} \Pr\Big\{\frac{\max_{1\le g\le G}\max_{1\le i\le n_g}|\tilde{V}_{g,i}|}{n^{\psi}} > \epsilon|\mathcal{W}\Big\} &\le \sum_{i=1}^{n}\Pr\Big\{\big|\sum_{g=1}^{G} M_{ig}V_g\big| > \epsilon n^{\psi}|\mathcal{W}\Big\} \\ &\le \frac{1}{\epsilon^4 n^{4\psi}}\sum_{i=1}^{n}\mathrm{E}\Big(\big(\sum_{g=1}^{G} M_{ig}V_g\big)^4|\mathcal{W}\Big), \qquad \text{(A.1)} \end{aligned}$$

where the first inequality follows from the Boole inequality, and the second inequality follows from the Markov inequality. Here, M_{ig} denotes the i^{th} row of cluster g's columns of M.

The fourth moment term can be simplified due independence across clusters:

$$\begin{aligned}\mathrm{E}\Big(\big(\sum_{g=1}^{G} M_{ig}V_g\big)^4|\mathcal{W}\Big) =& \sum_{g=1}^{G}\sum_{\tilde{g}=1}^{G}\sum_{h=1}^{G}\sum_{\tilde{h}=1}^{G}\mathrm{E}\big((M_{ig}V_g)(M_{i\tilde{g}}V_{\tilde{g}})(M_{ih}V_h)(M_{i\tilde{h}}V_{\tilde{h}})|\mathcal{W}\big)\\ =& \sum_{g=1}^{G}\mathrm{E}\big((M_{ig}V_g)^4|\mathcal{W}\big) + 3\sum_{g=1}^{G}\sum_{h=1,h\neq g}^{G}\mathrm{E}\big((M_{ig}V_g)^2(M_{ih}V_h)^2|\mathcal{W}\big),\end{aligned}$$

because

$$\mathrm{E}\big((M_{ig}V_g)(M_{i\tilde{g}}V_{\tilde{g}})(M_{ih}V_h)(M_{i\tilde{h}}V_{\tilde{h}})|\mathcal{W}\big)$$

is only non-zero when either (i) all four terms belong to the same cluster or (ii) when two of four terms belong to one cluster while the two other belong to another cluster. We can enumerate the number of cases where (ii) happens. Fix $1 \leq g \leq G$, then (ii) happens when:

1. $\tilde{g} = g$, $h \neq g$, $\tilde{h} = h$, or
2. $\tilde{g} \neq g$, $h = g$, $\tilde{h} = \tilde{g}$, or
3. $\tilde{g} \neq g$, $h = \tilde{g}$, $\tilde{h} = h$,

so there are three possible cases. We will show that both summands are bounded by $O_p\big(\max_{1\leq g\leq G} n_g^2\big)$. First, the left summand is, by $\sum_{j=1}^{n} M_{ij}^2 = M_{ii}$ and $\sum_{j=1}^{n_g}(M_{ig})_j^2 \leq M_{ii}$,

$$\begin{aligned}\sum_{g=1}^{G}\mathrm{E}\big((M_{ig}V_g)^4|\mathcal{W}\big) \leq& \sum_{g=1}^{G}\big(\sum_{j=1}^{n_g}(M_{ig})_j^2\big)^2\,\mathrm{E}\Big(\big(\sum_{j=1}^{n_g}V_{g,j}^2\big)^2|\mathcal{W}\Big)\\ \leq& M_{ii}\sum_{g=1}^{G}\sum_{j=1}^{n_g}(M_{ig})_j^2\,\mathrm{E}\Big(n_g\sum_{j=1}^{n_g}V_{g,j}^4|\mathcal{W}\Big)\\ \leq& M_{ii}^2\max_{1\leq g\leq G}\max_{1\leq i\leq n_g}\mathrm{E}\big(V_{g,i}^4|\mathcal{W}\big)\max_{1\leq g\leq G} n_g^2 \leq\ O_p\big(\max_{1\leq g\leq G} n_g^2\big),\end{aligned}$$

where first and second inequalities follow from the Cauchy-Schwarz inequality, and the last inequality from Assumption 2(iii). The right summand does not exceed 3 times

$$\begin{aligned}&\sum_{g=1}^{G}\sum_{h=1,h\neq g}^{G}\sum_{j=1}^{n_g}\sum_{\tilde{j}=1}^{n_h}(M_{ig})_j^2(M_{ig})_{\tilde{j}}^2\mathrm{E}\Big(\sum_{j=1}^{n_g}\sum_{\tilde{j}=1}^{n_h}V_j^2V_{\tilde{j}}^2|\mathcal{W}\Big)\\ \leq& \frac{1}{2}\sum_{g=1}^{G}\sum_{j=1}^{n_g}(M_{ig})_j^2\sum_{h=1,h\neq g}^{G}\sum_{\tilde{j}=1}^{n_h}(M_{ig})_{\tilde{j}}^2\mathrm{E}\Big(n_g\sum_{j=1}^{n_g}V_j^4 + n_h\sum_{\tilde{j}=1}^{n_h}V_{\tilde{j}}^4|\mathcal{W}\Big)\\ \leq& M_{ii}^2\max_{1\leq g\leq G}\max_{1\leq i\leq n_g}\mathrm{E}\big(V_{g,i}^4|\mathcal{W}\big)\max_{1\leq g\leq G} n_g^2 \leq\ O_p\big(\max_{1\leq g\leq G} n_g^2\big),\end{aligned}$$

where the first inequality follows from applying the Cauchy-Schwarz inequality twice and the inequality $|ab| \leq \frac{1}{2}(a^2 + b^2)$, and the last inequality from Assumption 2(iii).

We now go back to equation (A.1) and continue the argument:

$$\frac{1}{\epsilon^4 n^{4\psi}}\sum_{i=1}^{n}\mathrm{E}\Big(\big(\sum_{g=1}^{G} M_{ig}V_g\big)^4|\mathcal{W}\Big) \leq \frac{n\max_{1\leq g\leq G}\max_{1\leq i\leq n_g}\mathrm{E}\Big(\big(\sum_{g=1}^{G} M_{ig}V_g\big)^4|\mathcal{W}\Big)}{\epsilon^4 n^{4\psi}}$$

$$= \frac{O_p\big(\max_{1\leq g\leq G} n_g^2\big)}{\epsilon^4 n^{4\psi-1}} = o_p(1),$$

as $n^{-1}\max_{1\leq g\leq G} n_g^{\frac{2}{4\psi-1}} = o(1)$ by Assumption 4(i) with the imposed region for ψ.

Second, recall $M_{ij} = 1(i=j) - H_{ij}$, hence

$$\frac{\max_{1\leq g\leq G}\max_{1\leq i\leq n_g}|\tilde{Q}_{g,i}|}{n^\psi} \leq \frac{\max_{1\leq g\leq G}\max_{1\leq i\leq n_g}|Q_{g,i}|}{n^\psi} + \frac{1}{n^\psi}\max_{1\leq g\leq G}\max_{1\leq i\leq n_g}\Big|\sum_{j=1}^{n}(H_{ig})_j Q_{g,j}\Big|,$$

where, similar to M_{ig}, H_{ig} denotes the i^{th} row of cluster g's columns of H.

First, $\max_{1\leq g\leq G}\max_{1\leq i\leq n_g}|Q_{g,i}|/n^\psi = o_p(1)$ because

$$P\Big\{\frac{\max_{1\leq g\leq G}\max_{1\leq i\leq n_g}|Q_{g,i}|}{n^\psi} > \epsilon|\mathcal{W}\Big\} \leq \sum_{g=1}^{G}\sum_{i=1}^{n_g} P\Big\{|Q_{g,i}| > \epsilon n^\psi|\mathcal{W}\Big\}$$

$$\leq \frac{1}{\epsilon^{2+\theta}n^{(2+\theta)\psi}}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\mathrm{E}(|Q_{g,i}|^{2+\theta}|\mathcal{W})$$

$$= o_p\Big(\frac{1}{n^{(2+\theta)\psi-1}}\Big).$$

It remains to show that $n^{-\psi}\max_{1\leq g\leq G}\max_{1\leq i\leq n_g}\big|\sum_{j=1}^{n}(H_{ig})_j Q_{g,j}\big| = o_p(1)$. Note that

$$\Big(\frac{\max_{1\leq g\leq G}\max_{1\leq i\leq n_g}|\sum_{j=1}^{n}(H_{ig})_j Q_{g,j}|}{n^\psi}\Big)^2 \leq \Big(\max_{1\leq i\leq n}\sum_{j=1}^{n} H_{ij}^2\Big)\Big(\frac{1}{n^{2\psi}}\sum_{g=1}^{G}\sum_{i=1}^{n_g} Q_{g,i}^2\Big)$$

$$= \max_{1\leq i\leq n} H_{ii}\frac{1}{n^{2\psi}}\sum_{g=1}^{G}\|Q_g\|^2.$$

Under condition (a),

$$\max_{1\leq i\leq n} H_{ii}\frac{1}{n^{2\psi}}\sum_{g=1}^{G}\|Q_g\|^2 \leq O_p(1)o_p(1) = o_p(1).$$

Under condition (b), setting $q = 2+\theta$ and $p = 1/(1-1/q) = (2+\theta)/(1+\theta)$,

$$\Big(\frac{\max_{1\leq g\leq G}\max_{1\leq i\leq n_g}|\sum_{j=1}^{n}(H_{ig})_j Q_{g,j}|}{n^\psi}\Big)^2$$

$$\leq \Big(\max_{1\leq i\leq n}\frac{1}{n^{\frac{\psi(2+\theta)-1}{1+\theta}}}\sum_{j=1}^{n}|H_{ij}|^{\frac{2+\theta}{1+\theta}}\Big)^{1+\theta}\Big(\frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}|Q_{g,i}|^{2+\theta}\Big)$$

$$\leq \Big(\max_{1\leq i\leq n}\frac{1}{n^{\frac{\psi(2+\theta)-1}{1+\theta}}}\big[1+\sum_{j=1}^{n}1(M_{ij}\neq 0)\big]\Big)^{1+\theta}O_p(1)$$

$$= o_p(1),$$

where the first inequality comes from the Holder's inequality and last equality follows from condition (b). Finally, under condition (c), from the second line of the previous chain of (in)equalities,

$$\Big(\frac{\max_{1\leq g\leq G}\max_{1\leq i\leq n_g}|\sum_{j=1}^{n}(H_{ig})_j Q_{g,j}|}{n^{\psi}}\Big)^2 \leq o_p(1)O_p(1) = o_p(1).$$

□

A.2 Proof of main propositions

Proof of Theorem 3.1. We closely follow the roadmap of the proof of Theorem 1 in CJN (2018b). Let $S_n = n^{-1/2}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\hat{v}_{g,i}u_{g,i}$ and $\mathbb{V}_n = \text{var}(S_n|\mathcal{X},\mathcal{W})$. It suffices to prove that $\mathbb{V}_n^{-1} = O_p(1)$ and $\mathbb{V}_n^{-1/2}S_n \xrightarrow{d} \mathcal{N}(0,1)$. We first show that $\mathbb{V}_n^{-1} = O_p(1)$:

$$\begin{aligned}\mathbb{V}_n &= \frac{1}{n}\sum_{g=1}^{G}\hat{v}_g'\Omega_g\hat{v}_g \\ &\geq \lambda_{\min}(\Omega)\frac{1}{n}\sum_{g=1}^{G}\|\hat{v}_g\|^2.\end{aligned}$$

Note that $\big(n^{-1}\sum_{g=1}^{G}\|\hat{v}_g\|^2\big)^{-1} = O_p(1)$ by Lemma (A.3) and $\lambda_{\min}(\Omega)^{-1} = O_p(1)$ by Assumption 2(ii). Hence,

$$\begin{aligned}\mathbb{V}_n^{-1} &\leq \frac{1}{\lambda_{\min}(\Omega)}\frac{1}{n^{-1}\sum_{g=1}^{G}\|\hat{v}_g\|^2} \\ &= O_p(1).\end{aligned}$$

We now show that $\mathbb{V}_n^{-1/2}S_n \xrightarrow{d} \mathcal{N}(0,1)$. Because

$$\mathbb{V}_n^{-1/2}S_n = \frac{1}{\sqrt{n}}\sum_{g=1}^{G}\mathbb{V}_n^{-1/2}\sum_{i=1}^{n_g}\hat{v}_{g,i}u_{g,i},$$

this will hold if

$$\begin{aligned}&\sup_{z\in\mathbb{R}}\big|\Pr\{\mathbb{V}_n^{-1/2}S_n \leq z|\mathcal{X},\mathcal{W}\} - \Phi(z)\big| \\ &\leq \min\Big\{1, \frac{1}{n^{3/2}}\sum_{g=1}^{G}\mathrm{E}\Big(\big|\mathbb{V}_n^{-1/2}\sum_{i=1}^{n_g}\hat{v}_{g,i}u_{g,i}\big|^3|\mathcal{X},\mathcal{W}\Big)\Big\} \\ &= o_p(1),\end{aligned}$$

which follows from the conditional Berry-Esseen inequality. Indeed,

$$
\begin{aligned}
&\frac{1}{n^{3/2}}\sum_{g=1}^{G}\mathrm{E}\Big(\big|\mathbb{V}_n^{-1/2}\sum_{i=1}^{n_g}\hat v_{g,i}u_{g,i}\big|^3|\mathcal{X},\mathcal{W}\Big)\\
&\leq \mathbb{V}_n^{-3/2}\frac{1}{n^{3/2}}\sum_{g=1}^{G}\mathrm{E}\Big(\sum_{i=1}^{n_g}\sum_{j=1}^{n_g}\sum_{j_1=1}^{n_g}|\hat v_{g,i}u_{g,i}||\hat v_{g,j}u_{g,j}||\hat v_{g,j_1}u_{g,j_1}|\ |\mathcal{X},\mathcal{W}\Big)\\
&\leq \max_{1\leq g\leq G}\max_{1\leq i\leq n_g}\mathrm{E}\big(|u_{g,i}|^3|\mathcal{X},\mathcal{W}\big)\mathbb{V}_n^{-3/2}\frac{1}{n^{3/2}}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\sum_{j=1}^{n_g}\sum_{j_1=1}^{n_g}|\hat v_{g,i}\hat v_{g,j}\hat v_{g,j_1}|\\
&\leq O_p(1)\frac{\max_{1\leq g\leq G}\max_{1\leq i\leq n_g}|\hat v_{g,i}|}{n^{\psi}}\frac{1}{n^{3/2-\psi}}\sum_{g=1}^{G}n_g^2\|\hat v_g\|^2\\
&\leq O_p(1)\frac{\max_{1\leq g\leq G}\max_{1\leq i\leq n_g}|\hat v_{g,i}|}{n^{\psi}}\Big(\frac{\max_{1\leq g\leq G}n_g^{\frac{4}{1-2\psi}}}{n}\Big)^{\frac{1}{2}-\psi}\frac{1}{n}\sum_{g=1}^{G}\|\hat v_g\|^2\\
&= O_p(1)o_p(1)O_p(1)O_p(1)=o_p(1),
\end{aligned}
$$

using the Cauchy-Schwarz inequality, Assumption 4(i,ii), and Lemma (A.4). □

Proof of Lemma 3.1. Note that

$$
\begin{aligned}
y_g\hat u_{-g} &= y_g(y_g-\tilde X_g(\tilde X'_{-g}\tilde X_{-g})^{-1}\tilde X'_{-g}y_{-g})'\\
&= y_gy_g'-(\tilde X_g\beta+u_g)(\tilde X_{-g}\beta+u_{-g})'\tilde X_{-g}(\tilde X'_{-g}\tilde X_{-g})^{-1}\tilde X_g'\\
&= y_gy_g'-\tilde X_g\beta\beta'\tilde X'_{-g}\tilde X_{-g}(\tilde X'_{-g}\tilde X_{-g})^{-1}\tilde X_g'\\
&\quad +u_g\beta'\tilde X'_{-g}\tilde X_{-g}(\tilde X'_{-g}\tilde X_{-g})^{-1}\tilde X_g'+X_g\beta u'_{-g}\tilde X_{-g}(\tilde X'_{-g}\tilde X_{-g})^{-1}\tilde X_g',
\end{aligned}
$$

and so,

$$
\mathrm{E}(y_g\hat u_{-g}|\mathcal{X},\mathcal{W})=\mathrm{E}(y_gy_g'|\mathcal{X},\mathcal{W})-\tilde X_g\beta\beta'\tilde X_g'=\mathrm{E}(u_gu_g'|\mathcal{X},\mathcal{W})=\Omega_g.
$$

□

Proof of Theorem 3.3. It is sufficient to prove the consistency of the one-sided version of the LCOC estimator, i.e. that

$$
\frac{1}{n}\sum_{g=1}^{G}\hat v_g'(\mu_g+u_g)(\hat u_{-g})_g'\hat v_g-\frac{1}{n}\sum_{g=1}^{G}\hat v_g'\Omega_g\hat v_g=o_p(1).
$$

By Lemma (A.1) and triangular inequality, this will follow from the following two statements:

$$
\frac{1}{n}\sum_{g=1}^{G}\hat v_g'\mu_g\hat u_g'\tilde M_{gg}^{-1}\hat v_g=o_p(1), \tag{A.2}
$$

$$
\frac{1}{n}\sum_{g=1}^{G}\hat v_g'\big(u_g\hat u_g'\tilde M_{gg}^{-1}-\Omega_g\big)\hat v_g=o_p(1). \tag{A.3}
$$

To prove statement (A.2), note that the expression on the left of (A.2) has mean zero and, by the conditional Markov inequality, it is sufficient to show that

$$\mathrm{E}\Big(\big(\frac{1}{n}\sum_{g=1}^{G}\hat{v}_g'\mu_g\hat{u}_g'\tilde{M}_{gg}^{-1}\hat{v}_g\big)^2 \,|\, \mathcal{X},\mathcal{W}\Big) = o_p(1).$$

Reshuffling various scalar products and using Lemma (A.5), we find that this equals

$$\begin{aligned}
&\frac{1}{n^2}\sum_{g=1}^{G}\sum_{h=1}^{G}\mathrm{E}\Big(\mu_g'\hat{v}_g\hat{v}_g\tilde{M}_{gg}^{-1}\hat{u}_g\hat{u}_h'\tilde{M}_{hh}^{-1}\hat{v}_h\hat{v}_h'\mu_h \,|\, \mathcal{X},\mathcal{W}\Big)\\
&= \frac{1}{n^2}\sum_{g=1}^{G}\sum_{h=1}^{G}\mathrm{E}\big(\mu_g'\hat{v}_g\hat{v}_g'\tilde{M}_{gg}^{-1}\tilde{M}_g uu'\tilde{M}_h'\tilde{M}_{hh}^{-1}\hat{v}_h\hat{v}_h'\mu_h \,|\, \mathcal{X},\mathcal{W}\big)\\
&= \frac{1}{n^2}\sum_{g=1}^{G}\sum_{h=1}^{G}\mu_g'\hat{v}_g\hat{v}_g'\tilde{M}_{gg}^{-1}(\tilde{M}\Omega\tilde{M})_{gh}\tilde{M}_{hh}^{-1}\hat{v}_h\hat{v}_h'\mu_h\\
&= \frac{1}{n^2}\sum_{g=1}^{G}\sum_{h=1}^{G}a_g'(\tilde{M}\Omega\tilde{M})_{gh}a_h\\
&= \frac{1}{n^2}a_n'(\tilde{M}\Omega\tilde{M})a_n,
\end{aligned}$$

where $a_g = \mu_g'\hat{v}_g\hat{v}_g'\tilde{M}_{gg}^{-1}$ and $a_n' = (a_{g_1}', a_{g_2}', \ldots, a_{g_{n_g}}')$. Noting that $\|\tilde{M}\|^2 \leq 1$ and continuing the argument,

$$\begin{aligned}
\frac{1}{n^2}a_n'(\tilde{M}\Omega\tilde{M})a_n &\leq \frac{1}{n^2}\|a_n\|^2\lambda_{\max}(\Omega)\\
&\leq \frac{1}{n^2}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}|\mu_{g,i}||\hat{v}_{g,i}|\Big)^2\|\hat{v}_g\|^2\|\tilde{M}_{gg}^{-1}\|^2\lambda_{\max}(\Omega)\\
&\leq \frac{\max_{1\leq g\leq G}\max_{1\leq i\leq n_g}\mu_{g,i}^2}{n^{1/2-\psi}}\,\frac{\max_{1\leq g\leq G}\max_{1\leq i\leq n_g}|\hat{v}_{g,i}|^2}{n^{2\psi}}\\
&\qquad\times\Big(\frac{\max_{1\leq g\leq G}n_g^{\frac{4}{1-2\psi}}}{n}\Big)^{\frac{1}{2}-\psi}\frac{1}{n}\sum_{g=1}^{G}\|\hat{v}_g\|^2\cdot O_p(1)\\
&\leq O_p(1)o_p(1)O(1)O_p(1)O_p(1) = o_p(1),
\end{aligned}$$

where we use Assumption 4(i,ii), Lemma (A.4), that $\|a_g\| \leq |\mu_g'\hat{v}_g|\|\hat{v}_g\|\|\tilde{M}_{gg}^{-1}\|$, that $\|\tilde{M}_{gg}^{-1}\| = O_p(1)$ by Assumption 5(i), and that $\lambda_{\max}(\Omega) = O_p(1)$ by Assumption 5(iii).

To prove statement (A.3), we use Lemma (A.2):

$$\frac{1}{n}\sum_{g=1}^{G}\hat{v}_g'\big(u_g\hat{u}_g'\tilde{M}_{gg}^{-1} - \Omega_g\big)\hat{v}_g = \frac{1}{n}\sum_{g=1}^{G}\hat{v}_g'\big(u_gu_g' - \Omega_g\big)\hat{v}_g + \frac{1}{n}\sum_{g=1}^{G}\hat{v}_g'\big(u_gu_{-g}'\tilde{M}_{g,-g}'\tilde{M}_{gg}^{-1}\big)\hat{v}_g,$$

where $\mathrm{E}\left(u_gu_g' - \Omega_g \,|\, \mathcal{X},\mathcal{W}\right) = 0$ and $\mathrm{E}\left(u_gu_{-g}' \,|\, \mathcal{X},\mathcal{W}\right) = 0$, so that both terms on the right side have mean zero, and so, by the conditional Markov inequality, it is sufficient to show that

$$\mathrm{E}\Big(\big(\frac{1}{n}\sum_{g=1}^{G}\hat{v}_g'(u_gu_g'-\Omega_g)\hat{v}_g \mid \mathcal{X},\mathcal{W}\big)^2\Big)=o_p(1), \tag{A.4}$$

$$\mathrm{E}\Big(\big(\frac{1}{n}\sum_{g=1}^{G}\hat{v}_g'(u_gu_{-g}'\tilde{M}_{g,-g}'\tilde{M}_{gg}^{-1})\hat{v}_g\big)^2 \mid \mathcal{X},\mathcal{W}\Big)=o_p(1). \tag{A.5}$$

The left side of (A.4) is

$$\begin{aligned}&\frac{1}{n^2}\sum_{g=1}^{G}\mathrm{E}\big(\hat{v}_g'(u_gu_g'-\Omega_g)\hat{v}_g\hat{v}_g'(u_gu_g'-\Omega_g)\hat{v}_g \,|\, \mathcal{X},\mathcal{W}\big)\\ &=\frac{1}{n^2}\sum_{g=1}^{G}\mathrm{E}\big(\hat{v}_g'(u_gu_g')\hat{v}_g\hat{v}_g'(u_gu_g')\hat{v}_g \,|\, \mathcal{X},\mathcal{W}\big)-\frac{1}{n^2}\sum_{g=1}^{G}\hat{v}_g'\Omega_g\hat{v}_g\hat{v}_g'\Omega_g\hat{v}_g,\end{aligned}$$

by conditional unbiasedness and independence across g. Continuing,

$$\begin{aligned}&\le\frac{1}{n^2}\sum_{g=1}^{G}\mathrm{E}\big((\hat{v}_g'(u_gu_g')\hat{v}_g)^2 \,|\, \mathcal{X},\mathcal{W}\big)\\ &\le\frac{1}{n^2}\sum_{g=1}^{G}\|\hat{v}_g\|^4\mathrm{E}\big((\lambda_{\max}(u_gu_g'))^2 \,|\, \mathcal{X},\mathcal{W}\big)=\frac{1}{n^2}\sum_{g=1}^{G}\|\hat{v}_g\|^4\mathrm{E}\big(\|u_g\|^4 \,|\, \mathcal{X},\mathcal{W}\big)\\ &\le\max_{1\le g\le G}\max_{1\le i\le n_g}\mathrm{E}(|u_{g,i}|^4 \,|\, \mathcal{X},\mathcal{W})\frac{1}{n^2}\sum_{g=1}^{G}n_g^2\|\hat{v}_g\|^4\\ &\le\max_{1\le g\le G}\max_{1\le i\le n_g}\mathrm{E}(|u_{g,i}|^4 \,|\, \mathcal{X},\mathcal{W})\ \frac{\max_{1\le g\le G}\max_{1\le i\le n_g}|\hat{v}_{g,i}|^2}{n^{2\psi}}\\ &\qquad\times n^{-\frac{1}{4}(1-2\psi)}\ \Big(\frac{\max_{1\le g\le G}n_g^{\frac{4}{1-2\psi}}}{n}\Big)^{\frac{3}{4}(1-2\psi)}\ \frac{1}{n}\sum_{g=1}^{G}\|\hat{v}_g\|^2\\ &=O_p(1)o_p(1)o(1)O(1)O_p(1)=o_p(1),\end{aligned}$$

where we use Assumption 2(ii), Assumption 4(i,ii) and Lemma (A.4). The left side of (A.5) is

$$\begin{aligned}&\frac{1}{n^2}\mathrm{E}\Big(\sum_{g=1}^{G}\sum_{h=1}^{G}\hat{v}_g'u_gu_{-g}'\tilde{M}_{g,-g}'\tilde{M}_{gg}^{-1}\hat{v}_g\hat{v}_h'u_hu_{-h}'\tilde{M}_{h,-h}'\tilde{M}_{hh}^{-1}\hat{v}_h \,|\, \mathcal{X},\mathcal{W}\Big)\\ &=\frac{1}{n^2}\sum_{g=1}^{G}\sum_{h=1}^{G}\hat{v}_g'\tilde{M}_{gg}^{-1}\tilde{M}_{g,-g}\mathrm{E}\big(u_{-g}u_g'\hat{v}_g\hat{v}_h'u_hu_{-h}' \,|\, \mathcal{X},\mathcal{W}\big)\tilde{M}_{h,-h}'\tilde{M}_{hh}^{-1}\hat{v}_h\\ &=\frac{1}{n^2}\sum_{g=1}^{G}\sum_{h=1}^{G}\hat{v}_g'\tilde{M}_{gg}^{-1}\tilde{M}_{g,-g}\mathrm{E}\Big(\big(\sum_{j=1}^{n_h}\hat{v}_{h,j}u_{h,j}\big)u_{-g}u_{-h}'\big(\sum_{i=1}^{n_g}\hat{v}_{g,i}u_{g,i}\big)|\ \mathcal{X},\mathcal{W}\Big)\tilde{M}_{h,-h}'\tilde{M}_{hh}^{-1}\hat{v}_h.\end{aligned} \tag{A.6}$$

Note that the expectation $\mathrm{E}\Big(\big(\sum_{j=1}^{n_h}\hat{v}_{h,j}u_{h,j}\big)u_{-g}u_{-h}'\big(\sum_{i=1}^{n_g}\hat{v}_{g,i}u_{g,i}\big)|\ \mathcal{X},\mathcal{W}\Big)$ for $h\neq g$ is equal to

$$\begin{pmatrix} 0 \\ \vdots \\ \mathrm{E}\Big(u_h\big(\sum_{j=1}^{n_h}\hat{v}_{h,j}u_{h,j}\big)|\ \mathcal{X},\mathcal{W}\Big) \\ \vdots \\ 0 \end{pmatrix}\Big(0 \quad \dots \quad \mathrm{E}\Big(\big(\sum_{i=1}^{n_g}\hat{v}_{g,i}u_{g,i}\big)u_g'|\ \mathcal{X},\mathcal{W}\Big) \quad \dots \quad 0\Big).$$

This is because

$$\mathrm{E}\Big(\big(\sum_{j=1}^{n_h}\hat{v}_{h,j}u_{h,j}\big)u_l u_{\tilde{l}}'\big(\sum_{i=1}^{n_g}\hat{v}_{g,i}u_{g,i}\big)|\ \mathcal{X},\mathcal{W}\Big) = 0,$$

whenever $\tilde{l} \neq h, g$ or $l \neq h, g$, due to independence of u_g's across clusters. For $h = g$, the expectation term becomes

$$\mathrm{E}\Big(\big(\sum_{i=1}^{n_g}\hat{v}_{g,i}u_{g,i}\big)u_{-g}u_{-g}'\big(\sum_{i=1}^{n_g}\hat{v}_{g,i}u_{g,i}\big)|\,\mathcal{X},\mathcal{W}\Big) = \mathrm{E}\Big(\big(\sum_{i=1}^{n_g}\hat{v}_{g,i}u_{g,i}\big)^2|\,\mathcal{X},\mathcal{W}\Big)\Omega_{-g}.$$

Plugging these back into (A.6) and continuing,

$$= \frac{1}{n^2}\sum_{g=1}^{G}\hat{v}_g'\tilde{M}_{gg}^{-1}\tilde{M}_{g,-g}\mathrm{E}\Big(\big(\sum_{i=1}^{n_g}\hat{v}_{g,i}u_{g,i}\big)^2|\,\mathcal{X},\mathcal{W}\Big)\Omega_{-g}\tilde{M}_{g,-g}'\tilde{M}_{gg}^{-1}\hat{v}_g \tag{A.7}$$

$$+\ \frac{1}{n^2}\sum_{g=1}^{G}\sum_{h=1,h\neq g}^{G}\hat{v}_g'\tilde{M}_{gg}^{-1}\tilde{M}_{gh}\mathrm{E}\Big(u_h\big(\sum_{i=1}^{n_h}\hat{v}_{h,j}u_{h,j}\big)|\mathcal{X},\mathcal{W}\Big)$$

$$\times\,\mathrm{E}\Big(\big(\sum_{i=1}^{n_g}\hat{v}_{g,i}u_{g,i}\big)u_g'|\mathcal{X},\mathcal{W}\Big)\tilde{M}_{hg}'\tilde{M}_{hh}^{-1}\hat{v}_h. \tag{A.8}$$

The summand (A.7) equals

$$\begin{aligned}
&\frac{1}{n^2}\sum_{g=1}^{G}\hat{v}_g'\tilde{M}_{gg}^{-1}\tilde{M}_{g,-g}\mathrm{E}\big(\hat{v}_g'u_gu_g'\hat{v}_g|\ \mathcal{X},\mathcal{W}\big)\Omega_{-g}\tilde{M}_{g,-g}'\tilde{M}_{gg}^{-1}\hat{v}_g \\
&= \frac{1}{n^2}\sum_{g=1}^{G}\hat{v}_g'\tilde{M}_{gg}^{-1}\tilde{M}_{g,-g}\hat{v}_g'\Omega_g\hat{v}_g\Omega_{-g}\tilde{M}_{g,-g}'\tilde{M}_{gg}^{-1}\hat{v}_g \\
&\leq \frac{1}{n^2}\sum_{g=1}^{G}\|\hat{v}_g'\tilde{M}_{gg}^{-1}\tilde{M}_{g,-g}\|\|\hat{v}_g'\Omega_g\|\|\hat{v}_g\Omega_{-g}\|\|\tilde{M}_{g,-g}'\tilde{M}_{gg}^{-1}\hat{v}_g\| \\
&\leq \|\tilde{M}_{gg}^{-1}\|^2\|\tilde{M}_{g,-g}\|^2\|\Omega_g\|\|\Omega_{-g}\|\frac{\sum_{g=1}^{G}\|\hat{v}_g\|^4}{n^2} \\
&\leq O_p(1)\frac{\max_{1\leq g\leq G}\max_{1\leq i\leq n_g}|\hat{v}_{g,i}|^2}{n^{2\psi}}n^{-\frac{3}{4}(1-2\psi)}\ \Big(\frac{\max_{1\leq g\leq G}n_g^{\frac{4}{1-2\psi}}}{n}\Big)^{\frac{1}{4}(1-2\psi)}\ \frac{1}{n}\sum_{g=1}^{G}\|\hat{v}_g\|^2 \\
&= O_p(1)o_p(1)o(1)O(1)O_p(1) = o_p(1)
\end{aligned}$$

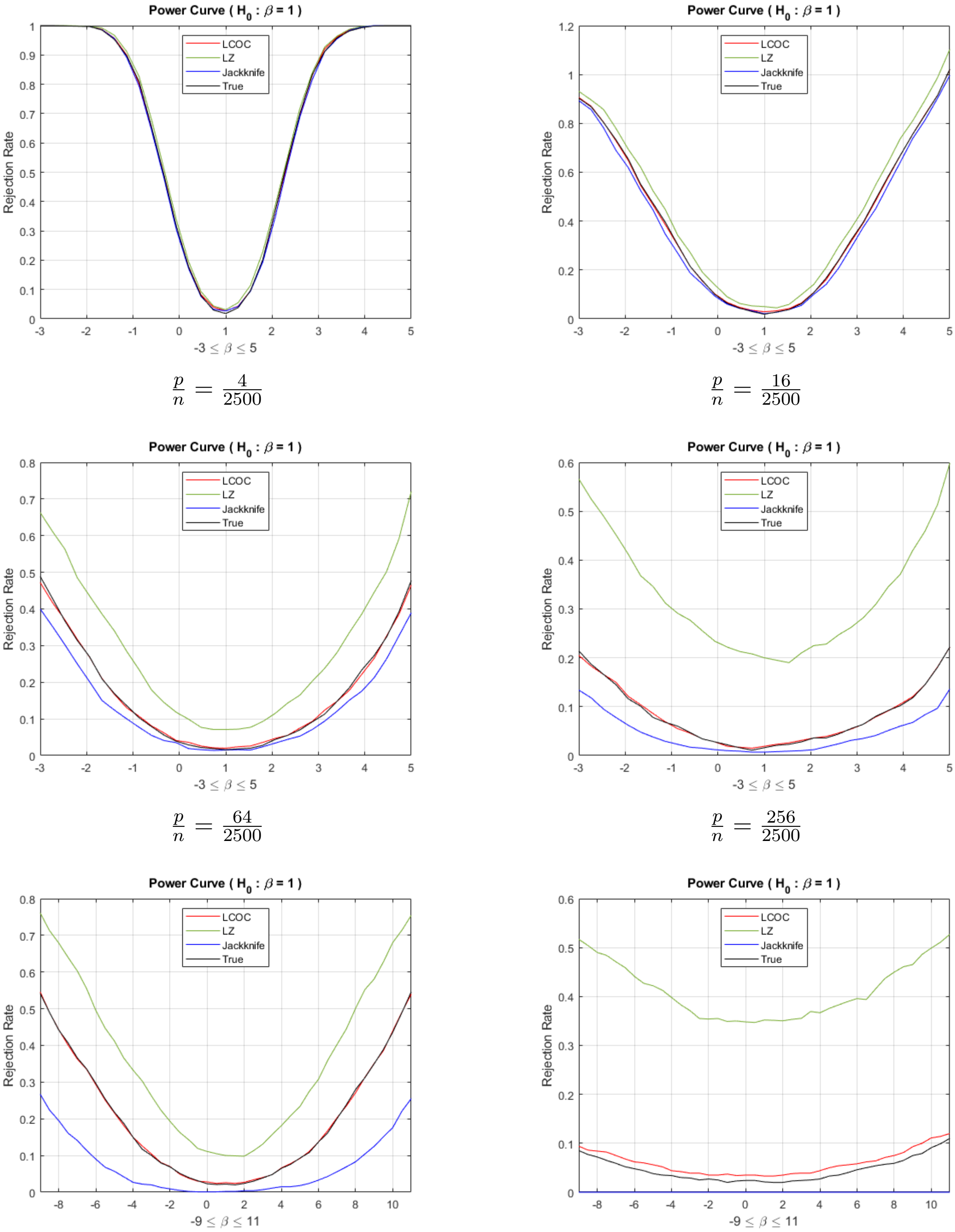
by Assumption 4(i,ii), Assumption 5(i,iii), Lemma (A.4), and because, using the eigenvalue interlacing theorem and Weyl's inequality,

$$\begin{aligned}
\|\tilde{M}_{g,-g}\|^2 &= \lambda_{\max}(\tilde{M}_{g,-g}\tilde{M}_{g,-g}') \\
&= \lambda_{\max}\Big(\sum_{h=1,h\neq g}^{G} \tilde{M}_{gh}\tilde{M}_{hg}\Big) \\
&= \lambda_{\max}(\tilde{M}_{gg} - \tilde{M}_{gg}^2) \\
&\leq \lambda_{\max}(\tilde{M}_{gg}) - \lambda_{\min}(\tilde{M}_{gg}^2) \leq 1.
\end{aligned}$$

The summand (A.8) equals

$$\begin{aligned}
&\frac{1}{n^2}\sum_{g=1}^{G}\sum_{h=1,h\neq g}^{G} \hat{v}_g'\tilde{M}_{gg}^{-1}\tilde{M}_{gh}\Omega_h\hat{v}_h\hat{v}_g'\Omega_g\tilde{M}_{gh}\tilde{M}_{hh}^{-1}\hat{v}_h \\
&\leq \frac{1}{n^2}\sum_{g=1}^{G}\sum_{h=1}^{G} \left|\hat{v}_g'\tilde{M}_{gg}^{-1}\tilde{M}_{gh}\Omega_h\hat{v}_h\hat{v}_g'\Omega_g\tilde{M}_{gh}\tilde{M}_{hh}^{-1}\hat{v}_h\right| \\
&\leq \frac{1}{n^2}\sum_{g=1}^{G}\sum_{h=1}^{G} \|\hat{v}_g'\tilde{M}_{gg}^{-1}\tilde{M}_{gh}\|\|\Omega_h\hat{v}_h\|\|\hat{v}_g'\Omega_g\tilde{M}_{gh}\|\|\tilde{M}_{hh}^{-1}\hat{v}_h\| \\
&\leq \max\Big\{\max_{1\leq h\leq G}\|\tilde{M}_{hh}^{-1}\|^2, \max_{1\leq h\leq G}\|\Omega_h\|^2\Big\}\max_{1\leq h\leq G}\|\hat{v}_h\|^2\frac{1}{n^2}\sum_{g=1}^{G}\sum_{h=1}^{G}\|\hat{v}_g'\tilde{M}_{gg}^{-1}\tilde{M}_{gh}\|\|\hat{v}_g'\Omega_g\tilde{M}_{gh}\| \\
&\leq O_p(1)\max_{1\leq h\leq G}\|\hat{v}_h\|^2\frac{1}{n^2}\sum_{g=1}^{G}\sqrt{\Big(\sum_{h=1}^{G}\|\hat{v}_g'\tilde{M}_{gg}^{-1}\tilde{M}_{gh}\|^2\Big)\Big(\sum_{h=1}^{G}\|\hat{v}_g'\Omega_g\tilde{M}_{gh}\|^2\Big)} \\
&\leq O_p(1)\max_{1\leq h\leq G}\|\hat{v}_h\|^2\frac{1}{n^2}\sum_{g=1}^{G}\sqrt{\big(\hat{v}_g'\tilde{M}_{gg}^{-1}\tilde{M}_{gg}\tilde{M}_{gg}^{-1}\hat{v}_g\big)\big(\hat{v}_g'\Omega_g\tilde{M}_{gg}\Omega_g\hat{v}_g\big)} \\
&\leq O_p(1)\frac{\max_{1\leq h\leq G}\max_{1\leq i\leq n_h}|\hat{v}_{h,i}|^2}{n^{2\psi}}n^{-\frac{3}{4}(1-2\psi)}\left(\frac{\max_{1\leq g\leq G} n_g^{\frac{4}{1-2\psi}}}{n}\right)^{\frac{1}{4}(1-2\psi)}\frac{1}{n}\sum_{g=1}^{G}\|\hat{v}_g\|^2 \\
&= O_p(1)o_p(1)o(1)O(1)O_p(1) = o_p(1)
\end{aligned}$$

by Assumption 4(i,ii), Assumption 5(i,iii) and Lemma (A.4). Hence, the consistency is proved. □



$\frac{p}{n} = \frac{4}{2500}$

$\frac{p}{n} = \frac{16}{2500}$

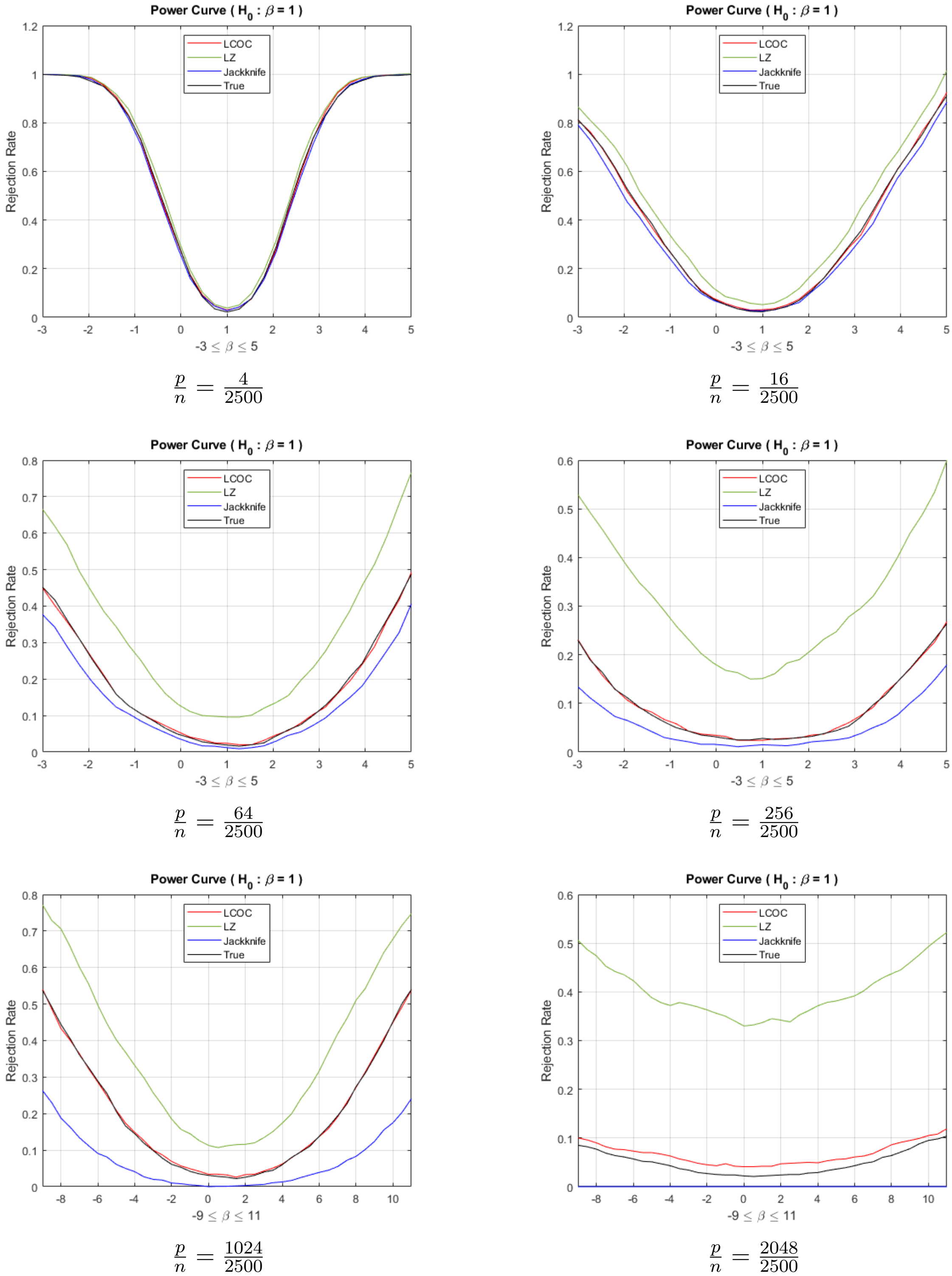
$\frac{p}{n} = \frac{64}{2500}$

$\frac{p}{n} = \frac{256}{2500}$

$\frac{p}{n} = \frac{1024}{2500}$

$\frac{p}{n} = \frac{2048}{2500}$

Figure 1: Power curves in correctly specified regression setup with balanced design. Sample size $n = 2500$, number of clusters $G = 100$.



$\frac{p}{n} = \frac{4}{2500}$

$\frac{p}{n} = \frac{16}{2500}$

$\frac{p}{n} = \frac{64}{2500}$

$\frac{p}{n} = \frac{256}{2500}$

$\frac{p}{n} = \frac{1024}{2500}$

$\frac{p}{n} = \frac{2048}{2500}$

Figure 2: Power curves in correctly specified regression setup with unbalanced design. Sample size $n = 2500$, number of clusters $G = 100$.

k	LZ	LCOC	JK	LZ	LCOC	JK
	Design 1 (balanced clusters)			Design 2 (unbalanced clusters)		
4	1.03	0.99	0.97	1.05	0.99	0.97
16	1.08	1.00	0.97	1.13	1.00	0.97
64	1.27	0.99	0.91	1.38	0.99	0.91
256	1.66	1.00	0.87	1.67	1.03	0.90
1024	1.39	1.01	0.75	1.38	1.01	0.75
2048	2.38	1.02	0.42	2.36	1.00	0.41

Table 1: Ratios of simulated OLS estimator's standard deviation to average standard errors in simulation experiment 1. Sample size $n = 2500$, number of clusters $G = 100$.

k	LZ	LCOC	JK	LZ	LCOC	JK
	Design 1 (balanced clusters)			Design 2 (unbalanced clusters)		
4	93.5%	94.7%	94.8%	92.8%	94.5%	94.7%
16	92.9%	95.0%	95.6%	91.0%	94.4%	94.9%
64	86.2%	94.7%	96.3%	83.0%	94.8%	96.6%
256	75.7%	94.4%	97.4%	76.5%	93.8%	96.5%
1024	84.1%	95.1%	98.7%	84.8%	94.9%	99.1%
2048	56.5%	93.9%	100.0%	42.0%	94.1%	100.0%

Table 2: Actual coverage rates for 95%-level confidence intervals in simulation experiment 1. Sample size $n = 2500$, number of clusters $G = 100$.

k	LZ	LCOC	JK	LZ	LCOC	JK
	Exponential specification			Sparse specification		
5	1.07	1.00	0.95	1.07	1.00	0.95
10	1.09	1.00	0.95	1.09	1.00	0.96
20	1.09	1.00	0.95	1.09	1.00	0.95
40	1.10	1.00	0.95	1.09	1.00	0.95
80	1.11	1.01	0.95	1.10	1.01	0.95
160	1.11	1.01	0.95	1.11	1.01	0.96

Table 3: Ratios of simulated OLS estimator's standard deviation to average standard errors in simulation experiment 2 (partially linear model). Sample size $n = 250$, number of clusters $G = 25$, cluster size $n_g = 10$.

k	LZ	LCOC	JK	LZ	LCOC	JK
	Exponential specification			Sparse specification		
5	91.0%	92.6%	93.5%	91.0%	92.1%	93.5%
10	90.6%	92.5%	93.9%	90.6%	92.0%	93.8%
20	90.7%	92.9%	94.2%	90.7%	92.7%	94.1%
40	90.1%	92.5%	93.8%	90.2%	92.3%	93.7%
80	89.8%	92.3%	93.8%	89.9%	92.0%	93.6%
160	89.5%	92.4%	93.9%	89.6%	92.2%	93.8%

Table 4: Actual coverage rates for 95%-level confidence intervals in simulation experiment 2 (partially linear model). Sample size $n = 250$, number of clusters $G = 25$, cluster size $n_g = 10$.

Estimator	Bandwidth	Exponential specification	Sparse specification
OLS	$k = 5$	0.99	0.99
	$k = 10$	0.99	0.99
	$k = 20$	0.99	0.99
	$k = 40$	1.01	1.01
	$k = 80$	0.99	0.99
	$k = 160$	1.01	1.01
DDB-LASSO	$\lambda = 0.001$	0.89	0.96
	$\lambda = 0.01$	1.02	1.10
	$\lambda = 0.1$	0.98	1.05
	$\lambda = 0.2$	0.83	0.90

Table 5: Averages of parameter estimates in simulation experiment 2 (partially linear model). Sample size $n = 250$, number of clusters $G = 25$, cluster size $n_g = 10$.

	k	Skedastic function	LZ	LCOC	JK	LZ	LCOC	JK
			Standard error ratios			Actual coverage rates		
(i)	9	depends on x's only	1.02	1.00	0.98	93.7%	94.1%	94.6%
(ii)	180	depends on x's only	1.19	1.00	0.95	89.1%	93.5%	95.5%
(iii)	9	depends on x's and w's	1.02	1.00	0.97	93.8%	94.2%	94.8%
(iv)	180	depends on x's and w's	1.15	1.00	0.89	90.2%	93.9%	96.6%

Table 6: Ratios of simulated OLS estimator's standard deviation to average standard errors and actual coverage rates for 95%-level confidence intervals in simulation experiment 3 (emulating Donohue and Levitt 2001). Sample size: $n = 624$, number of clusters $G = 48$, cluster size $n_g = 13$.

Sample	$\hat{\beta}$	LCOC	JK	$\frac{p}{N}$
		Girls, $n = 1861$, $G = 34$		
SC	0.1046	0.0688	0.0649	0.0034
		[0.064]	[0.054]	
SC + Q + M	0.1047	0.0491	0.0514	0.0059
		[0.017]	[0.021]	
		Boys, $n = 1960$, $G = 34$		
SC	−0.0104	0.0518	0.0564	0.0021
		[0.579]	[0.573]	
SC + Q + M	−0.0222	0.0428	0.0475	0.0056
		[0.698]	[0.680]	
		Full, $n = 3821$, $G = 39$		
SC	0.0561	0.0507	0.0521	0.0010
		[0.134]	[0.141]	
SC + P	0.0523	0.0452	0.0724	0.0058
		[0.124]	[0.235]	
SC + Q + M	0.0524	0.0388	0.0422	0.0029
		[0.089]	[0.107]	
SC + Q + M + P	0.0675	0.0351	0.0539	0.0076
		[0.027]	[0.106]	

Table 7: Replication of Angrist and Lavy (2009)'s Table 2, panel A. Covariate notation: SC for school covariates, Q for quartile dummies, M for micro covariates, P for pair effects. The square brackets contain p-values for statistical significance of β against right-sided alternative.

$N = 48$, $T = 13$	$\hat{\beta}$	LZ	LCOC	JK	$\dfrac{k}{NT}$
		Violent crime			
Baseline	-0.1304	0.0420 [0.001]	0.0441 [0.002]	0.0500 [0.005]	0.034
Time-varying controls	-0.1005	0.0413 [0.008]	0.0495 [0.022]	0.0533 [0.030]	0.175
		Property crime			
Baseline	-0.0910	0.0145 [0.000]	0.0163 [0.000]	0.0166 [0.000]	0.034
Time-varying controls	-0.0817	0.0190 [0.000]	0.0221 [0.000]	0.0244 [0.000]	0.175
		Murder			
Baseline	-0.1305	0.0534 [0.007]	0.0552 [0.009]	0.0619 [0.018]	0.034
Time-varying controls	-0.1118	0.0695 [0.054]	0.0722 [0.061]	0.0876 [0.101]	0.175

Table 8: Replication of Donohue and Levitt (2001)'s Table IV. The square brackets contain p-values for statistical significance of β against left-sided alternative.