Cluster-robust Standard Errors for Linear Regression Models with Many Controls*

Riccardo D'Adamo†

September 9, 2018

Abstract

The linear regression model is widely used in empirical economics to estimate the structural/treatment effect of some variable on an outcome of interest. Researchers often include a large set of regressors in order to control for observed and unobserved confounders. In this paper, we develop inference methods for linear regression models with many controls and clustering. We show that inference based on the usual cluster-robust standard errors by White (1984) is invalid in general when the number of controls is a non-vanishing fraction of the sample size. We then propose a new clustered standard errors formula that is robust to the inclusion of many controls and allows to carry out valid inference in a variety of high-dimensional linear regression models, including multi-way fixed effects panel data models and the semiparametric partially linear model. Monte Carlo evidence supports our theoretical results and shows that our proposed variance estimator performs well in finite samples.

*I am grateful to Martin Weidner for his help and support since the start of this project. I would also like to thank Ivan Canay, Andrew Chesher, Whitney Newey, Daniel Wilhelm and participants at the Bristol Econometric Study Group for helpful comments and discussions.

†Department of Economics, University College London. E-mail: riccardo.d'adamo.15@ucl.ac.uk

1 Introduction

In many important applications of the linear regression model, the object of interest is $\boldsymbol{\beta}$ in a model of the form

$$y_{i,n} = \boldsymbol{\beta}'\mathbf{x}_{i,n} + \boldsymbol{\gamma}_n'\mathbf{w}_{i,n} + u_{i,n}, \qquad i = 1, \ldots, n, \tag{1}$$

where $y_{i,n}$ is a scalar outcome variable, $\mathbf{x}_{i,n}$ is a $d \times 1$ vector of regressors of fixed dimension, $\mathbf{w}_{i,n}$ is a vector of covariates of possibly "large" dimension K_n, and $u_{i,n}$ is an unobserved scalar error term. Conducting OLS-based inference on $\boldsymbol{\beta}$ is known to be straightforward when the error $u_{i,n}$ is (conditionally) homoskedastic and/or the dimension of the nuisance covariates $\mathbf{w}_{i,n}$ is modelled as a vanishing fraction of the sample size (see, e.g., Mammen, 1993, and Anatolyev, 2012). Both of these assumptions can be problematic. First, the assumption of homoskedasticity is undesirable in most empirical applications, motivating the widespread use of inference procedures that are robust to heteroskedasticity and various forms of dependence in the errors. In particular, researchers often assume that errors are clustered at some economically relevant level, e.g. by individual or geographical location. Secondly, the assumption that $K_n/n \to 0$ is unpalatable or even violated in many applications of model (1), as discussed below.

Motivated by the above observations, this paper develops inference theory for linear regression models with many controls and clustering. In particular, we first show that the usual cluster-robust standard errors by White (1984) are inconsistent in general when $K_n/n \nrightarrow 0$. We then propose a new clustered standard error formula that allows to carry out valid inference under asymptotics in which K_n is allowed (but not required) to grow as fast as the sample size.

The findings of this paper contribute to a sizeable body of literature that deals with inference procedures in models that involve the estimation of many incidental parameters; see, e.g., Angrist and Hahn (2004), Hahn and Newey (2004), Stock and Watson (2008), Cattaneo et al. (2018), Verdier (2017), and references therein. In particular, our findings can be seen as generalising those of Cattaneo, Jansson and Newey (2018a), CJN hereafter, who establish asymptotic normality of the OLS estimator of $\boldsymbol{\beta}$ in (1) when $K_n/n \nrightarrow 0$, and provide inference methods under such asymptotics when the errors are heteroskedastic.

This paper also adds to the long-established literature initiated by White (1984) dealing with cluster-robust inference in a variety of models, a recent review of which is given by Cameron and Miller (2015); see, e.g., Arellano (1987), Bell and McCaffrey (2002), Hansen (2007), Cameron et al. (2008) and Ibragimov and Muller (2016).

The rest of this paper is organized as follows. Section 2 introduces the framework of the paper and illustrates its relevance using three leading examples. Section 3 discusses our assumptions. Section 4 presents our main theoretical results. Section 5 reports the findings of a Monte Carlo experiment. Section 6 briefly concludes. Proofs and extensions of the results are given in the Appendix.

2 Framework and Motivation

The main object of interest in our analysis is $\boldsymbol{\beta}$ in (1), on which we would like to carry out inference while treating the high-dimensional $\mathbf{w}_{i,n}$ as nuisance covariates. A natural choice of estimator for $\boldsymbol{\beta}$ is the OLS estimator, which can be written as

$$\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{n} \hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}'_{i,n})^{-1}(\sum_{i=1}^{n} \hat{\mathbf{v}}_{i,n}y_{i,n}), \qquad \hat{\mathbf{v}}_{i,n} = \sum_{j=1}^{n} M_{ij,n}\mathbf{x}_{j,n}, \tag{2}$$

where $M_{ij,n} = \mathbb{1}\{i = j\} - \mathbf{w}'_{i,n}(\sum_{k=1}^{n} \mathbf{w}_{k,n}\mathbf{w}'_{k,n})^{-1}\mathbf{w}_{j,n}$ is the (i, j) entry of the symmetric and idempotent annihilator matrix $\mathbf{M}_n$, with $\mathbb{1}\{\cdot\}$ denoting the indicator function. Defining $\hat{\boldsymbol{\Gamma}}_n = \sum_{i=1}^{n} \hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}'_{i,n}/n$ and $\boldsymbol{\Sigma}_n$ the (conditional) variance of $\sum_{i=1}^{n} \hat{\mathbf{v}}_{i,n}u_{i,n}/\sqrt{n}$, it is well-known that, when $n \to \infty$ and K_n is fixed, the asymptotic distribution of $\hat{\boldsymbol{\beta}}_n$ is

$$\boldsymbol{\Omega}_n^{-1/2}\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d), \qquad \boldsymbol{\Omega}_n = \hat{\boldsymbol{\Gamma}}_n^{-1}\boldsymbol{\Sigma}_n\hat{\boldsymbol{\Gamma}}_n^{-1}. \tag{3}$$

When the errors are assumed to be correlated only within G_n clusters of bounded size, $\boldsymbol{\Sigma}_n$ can be estimated consistently with the popular cluster-robust standard errors by White (1984):

$$\hat{\boldsymbol{\Sigma}}_n^{\mathtt{W}} = \frac{1}{n}\sum_{g=1}^{G_n}\sum_{i,j\in\mathcal{T}_{g,n}} \hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}'_{j,n}\hat{u}_{i,n}\hat{u}_{j,n}, \qquad \hat{u}_{i,n} = \sum_{j=1}^{n} M_{ij,n}(y_{j,n} - \hat{\boldsymbol{\beta}}'_n\mathbf{x}_{j,n}), \tag{4}$$

where $\mathcal{T}_{g,n}$ denotes the subset of observations contained in cluster g and $\{\mathcal{T}_{g,n} : 1 \leq g \leq G_n\}$ is a partition of the data. As a result, asymptotically valid inference can be carried out using the usual testing procedures based on the distributional approximation $\hat{\boldsymbol{\beta}}_n \overset{a}{\sim} \mathcal{N}(\boldsymbol{\beta}, \hat{\boldsymbol{\Gamma}}_n^{-1}\hat{\boldsymbol{\Sigma}}_n\hat{\boldsymbol{\Gamma}}_n^{-1}/n)$.

The objective of this paper is to establish cluster-robust inference procedures for $\boldsymbol{\beta}$ under asymptotics in which $K_n/n \nrightarrow 0$. Allowing the dimension of the nuisance covariates K_n to grow at the same rate as the sample size n enables us to cover many relevant applications of the general model in (1).

Example 1 Linear regression model with increasing dimension
This leading example takes (1) as the data generating process, in which $\mathbf{w}_{i,n}$ contains many observable individual characteristics and their nonlinear transformations, dummy variables for many categories such as age group, cohort, geographic location etc. and their interactions with the former. The inclusion of many covariates is motivated in practice by the assumption that the variable of interest $\mathbf{x}_{i,n}$ can be taken as exogenous after controlling for $\mathbf{w}_{i,n}$. Although the study of linear regression models with growing dimension has a long tradition in econometrics (see, e.g., Huber, 1973, and Mammen, 1993), until recently inference results were exiguous and limited to the case in which the number of regressors in the model is at least a vanishing fraction of the sample size. CJN (2018a) exploit the separability of model (1) to develop valid inference procedures for $\boldsymbol{\beta}$ when $K_n/n \nrightarrow 0$, but their theory only covers the case of homoskedastic and heteroskedastic errors. Li and Muller (2016) develop cluster-robust inference theory in this setting for a scalar β, i.e. $d = 1$; their results allow for $K_n \propto n$ but rely on a strong restriction on $\sum_{i=1}^{n}(\boldsymbol{\gamma}_n'\mathbf{w}_{i,n})^2$, which limits the amount of sample variation of y_i that can be induced by the high-dimensional controls $\mathbf{w}_{i,n}$. ■

Example 2 Semiparametric Partially Linear Model
Researchers often assume that data are generated by the model

$$y_i = \boldsymbol{\beta}'\mathbf{x}_i + g(\mathbf{z}_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{5}$$

where both $\mathbf{x}_i$ and $\mathbf{z}_i$ have fixed dimension, but the function $g(\cdot)$ is unknown. The partially linear model is a long-standing area of interest in econometrics (see, e.g., Heckman, 1986, and Robinson, 1988). Estimation of this semiparametric model is often carried out via series-based methods, in which the researcher assumes that the function $g(\cdot)$ can be closely approximated using the polynomial functions $\mathbf{p}_n(\mathbf{z}) = (p^1(\mathbf{z}), \ldots, p^{K_n}(\mathbf{z}))'$, so that $g(\mathbf{z}_i) \approx \boldsymbol{\gamma}_n'\mathbf{p}_n(\mathbf{z}_i)$ for some $\boldsymbol{\gamma}_n$. The series estimator for $\boldsymbol{\beta}$ is the OLS estimator as defined in (2), where $\mathbf{w}_{i,n} = \mathbf{p}_n(\mathbf{z}_i)$. When the underlying function $g(\cdot)$ is not sufficiently smooth or/and the dimension of $\mathbf{z}_i$ is relatively large, the inclusion of many polynomial terms might be required, resulting in K_n being non-negligible relative to n. CJN (2018b) are the first to consider asymptotics in which $K_n/n \nrightarrow 0$ in this setting. They establish asymptotic normality for $\hat{\boldsymbol{\beta}}_n$

and valid inference procedures under homoskedasticity and, in their subsequent paper, heteroskedasticity (CJN, 2018a). However, no results are available for the case of clustering. ■

Example 3 Multi-way fixed effects panel data models
Panel data models that use fixed effects are often used in order to control for unobserved heterogeneity, such as the one-way fixed effects panel data regression model

$$Y_{it} = \boldsymbol{\beta}'\mathbf{X}_{it} + \alpha_i + U_{it}, \qquad i = 1, \ldots, N, \qquad t = 1, \ldots, T, \tag{6}$$

where α_i is a scalar individual effect, $\mathbf{X}_{it}$ is a vector of regressors and U_{it} is a scalar error term. This model can be mapped into our baseline specification in (1) by setting $n = NT$, $y_{(i-1)T+t,n} = Y_{it}$, $\mathbf{x}_{(i-1)T+t,n} = \mathbf{X}_{it}$, $u_{(i-1)T+t,n} = U_{it}$, $\boldsymbol{\gamma}_n = (\alpha_1, \ldots, \alpha_N)$ and $\mathbf{w}_{(i-1)T+t,n}$ equal to the i-th unit vector of dimension N. It follows that $K_n = N$ and $K_n/n = 1/T$, which motivates the asymptotics of this paper under $N \to \infty$ and T fixed. For this case, Arellano (1987) shows that White's estimator (1984) is consistent when errors are clustered at the individual level. For the same setting, Stock and Watson (2008) propose a cluster-robust estimator for the variance with additional zero-restrictions on the conditional autocovariances of the errors within entities, e.g. when an MA(q) structure is imposed on U_{it}. In many empirical settings, researchers want to control for multiple terms of unobserved heterogeneity. In the analysis of student/teacher or worker/firm matched data, for example, two-way fixed effects models are commonly used, taking the form

$$Y_{it} = \boldsymbol{\beta}'\mathbf{X}_{it} + \alpha_i + e_{d_{it}} + U_{it}, \qquad i = 1, \ldots, N, \qquad t = 1, \ldots, T, \tag{7}$$

where $e_{d_{it}}$ are unobserved factors common to all observations sharing the same value of the indexing variable $d_{it} \in \{1, \ldots, N_d\}$, so that $\mathbf{w}_{(i-1)T+t,n}$ is now a $N + N_d$ vector selecting the relevant fixed effects from $\boldsymbol{\gamma}_n = (\alpha_1, \ldots, \alpha_N, e_1, \ldots, e_{N_d})$. When T is fixed and only few observations are assigned to each value of d_{it} (i.e. data are sparsely matched), then the number of fixed effects grows proportionally to the sample size and $K_n \propto n$. Verdier (2017) considers cluster-robust inference under these asymptotics, although his results are derived for a new estimation procedure for $\boldsymbol{\beta}$ that accomodates instrumental variables but is generally less efficient than OLS. ■

To simplify exposition, we present our inference theory for linear regression models with many controls for the case of strictly exogenous regressors. While all the results of this paper can be well-understood for this special case, their generalization to (potential) misspecification bias in the model is straightforward and is provided in the Appendix.

3 Assumptions

In this section we present a set of assumptions for the special case of strict exogeneity of the regressors. A more general set of assumptions that allows for misspecification bias is given in the Appendix.

Suppose then that $\{(y_{i,n}, \mathbf{x}_{i,n}', \mathbf{w}_{i,n}') : 1 \leq i \leq n\}$ is generated by (1) and set $\mathcal{X}_n = (\mathbf{x}_{1,n}, \ldots, \mathbf{x}_{n,n})$ and $\mathcal{W}_n = (\mathbf{w}_{1,n}, \ldots, \mathbf{w}_{n,n})$. We define the following quantities:

$$\begin{aligned} \chi_n &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\|\mathbf{Q}_{i,n}\|^2], \qquad \mathbf{Q}_{i,n} = \mathbb{E}[\mathbf{v}_{i,n}|\mathcal{W}_n], \\ \hat{\mathbf{\Gamma}}_n &= \sum_{i=1}^{n} \hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{i,n}'/n, \qquad \mathbf{\Sigma}_n = \mathbb{V}[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{\mathbf{v}}_{i,n}u_{i,n}|\mathcal{X}_n, \mathcal{W}_n], \end{aligned} \tag{8}$$

where $\mathbf{v}_{i,n} = \mathbf{x}_{i,n} - (\sum_{j=1}^{n} \mathbb{E}[\mathbf{x}_{j,n}\mathbf{w}_{j,n}'])(\sum_{j=1}^{n} \mathbb{E}[\mathbf{w}_{j,n}\mathbf{w}_{j,n}'])^{-1}\mathbf{w}_{i,n}$ is the population counterpart of $\hat{\mathbf{v}}_{i,n}$. Also, letting $\lambda_{\min}(\cdot)$ denote the minimum eigenvalue of its argument, define

$$\mathcal{C}_n = \max_{1\leq i\leq n} \{\mathbb{E}[u_{i,n}^4|\mathcal{X}_n, \mathcal{W}_n] + \mathbb{E}[\|\mathbf{V}_{i,n}\|^4|\mathcal{W}_n] + 1/\mathbb{E}[u_{i,n}^2|\mathcal{X}_n, \mathcal{W}_n]\} + 1/\lambda_{\min}(\mathbb{E}[\tilde{\mathbf{\Gamma}}_n|\mathcal{W}_n]) \tag{9}$$

where $\mathbf{V}_{i,n} = \mathbf{x}_{i,n} - \mathbb{E}[\mathbf{x}_{i,n}|\mathcal{W}_n]$, $\tilde{\mathbf{\Gamma}}_n = \sum_{i=1}^{n} \tilde{\mathbf{V}}_{i,n}\tilde{\mathbf{V}}_{i,n}'/n$ and $\tilde{\mathbf{V}}_{i,n} = \sum_{j=1}^{n} M_{ij,n}\mathbf{V}_{i,n}$.

We impose the following three assumptions:

Assumption 1 $\max_{1\leq g\leq G_n} \#\mathcal{T}_{g,n} = O(1)$, *where* $\#\mathcal{T}_{g,n}$ *is the cardinality of* $\mathcal{T}_{g,n}$ *and where* $\{\mathcal{T}_{g,n} : 1 \leq g \leq G_n\}$ *is a partition of* $\{1,\ldots,n\}$ *such that* $\{(u_{i,n}, \mathbf{x}_{i,n}') : i \in \mathcal{T}_{g,n}\}$ *are independent over* g *conditional on* $\mathcal{W}_n$.

Assumption 2 $\mathbb{P}[\lambda_{min}(\sum_{i=1}^{n} \mathbf{w}_{i,n}\mathbf{w}_{i,n}') > 0] \to 1$, $\limsup_{n\to\infty} K_n/n < 1$, $\mathcal{C}_n = O_p(1)$ *and* $\mathbf{\Sigma}_n^{-1} = O_p(1)$

Assumption 3 $\mathbb{E}[u_{i,n}|\mathcal{X}_n, \mathcal{W}_n] = 0 \quad \forall i, n$, $\chi_n = O(1)$, *and* $\max_{1\leq i\leq n}\|\hat{\mathbf{v}}_{i,n}\|/\sqrt{n} = o_p(1)$.

Assumption 1 defines the sampling structure, in which we allow for arbitrary dependence within clusters of finite but possibly heterogenous size for both the regressors and the errors.

In terms of clustering structure, the resulting asymptotics are the same as the typical ones of White (1984) in which $n, G_n \to \infty$ and $G_n \propto n$. We expect that the results of this paper would generalize to asymptotics where cluster sizes are allowed to diverge with n and G_n, as considered in Hansen (2007) and Hansen and Lee (2018). It is likely that such extension would require imposing more restrictive conditions on the regression design and the distributional properties of the errors, e.g. stationarity and/or mixing, and we leave it to future work.

Assumption 2 allows for asymptotics where $K_n/n \nrightarrow 0$, while imposing standard restrictions on the regression design and some bounds on the (conditional) higher-order moments of the structural residuals $u_{i,n}$ and $\mathbf{V}_{i,n}$.

The conditions on χ_n in Assumption 3 is a requirement on the quality of the linear approximation for the conditional expectation $\mathbb{E}[\mathbf{x}_{i,n}|\mathcal{W}_n]$. The high-level condition $\max_{1\leq i\leq n}\|\hat{\mathbf{v}}_{i,n}\|/\sqrt{n} = o_p(1)$ also places restrictions on the relationship between $\mathbf{x}_{i,n}$ and $\mathbf{w}_{i,n}$ and has a central importance in establishing asymptotic normality of the OLS estimator for $\boldsymbol{\beta}$ and consistency of our proposed variance estimator. CJN (2018a) show that this restriction holds under mild moment conditions when either (i) $K_n/n \to 0$, or (ii) $\chi_n = o(1)$ or (iii) $\max_{1\leq i\leq n}\sum_{i=1}^{n}\mathbb{1}\{M_{ij,n} \neq 0\} = o_p(n^{1/3})$. While condition (i) is not the case of primary interest of this paper, (ii) and (iii) accomodate $K_n/n \nrightarrow 0$ and can be used to verify the high-level condition $\max_{1\leq i\leq n}\|\hat{\mathbf{v}}_{i,n}\|/\sqrt{n} = o_p(1)$ when $\mathbf{w}_{i,n}$ can be interpreted as approximating functions, dummy/discrete variables or fixed effects. See CJN (2018a) for details.

Remark 1 In the general formulation provided in Appendix, the assumptions we consider are analogous to those in CJN (2018a) but we also allow for clustering in the errors. The set of restrictions imposed by this framework allows to cover the three leading examples presented in the previous section. A detailed discussion of conditions that satisfy the assumptions in those particular models is provided in CJN (2018a) and their Supplemental Appendix. ■

4 Main Results

This section presents our main theoretical results for inference in linear regression models with many controls and clustering under the set of simplified assumptions presented in Section 3. Proofs of the theorems and other auxiliary results are given in the Appendix for the general case that allows for misspecification bias in the model.

Our first result extends the asymptotic normality result for $\hat{\boldsymbol{\beta}}$ previously derived by CJN

(2018a) to the case of clustering.

Theorem 1 *Suppose Assumptions 1-3 hold. Then,*

$$\boldsymbol{\Omega}_n^{-1/2}\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d), \qquad \boldsymbol{\Omega}_n = \hat{\boldsymbol{\Gamma}}_n^{-1}\boldsymbol{\Sigma}_n\hat{\boldsymbol{\Gamma}}_n^{-1}, \tag{10}$$

where $\boldsymbol{\Sigma}_n = \frac{1}{n}\sum_{g=1}^{G_n}\sum_{i,j\in\mathcal{T}_{g,n}} \hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{j,n}'\mathbb{E}[u_{i,n}u_{j,n}|\mathcal{X}_n, \mathcal{W}_n]$.

Theorem 1 implies that the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ under $K_n/n \nrightarrow 0$ resembles the standard one obtainable under fixed-K_n.[1] As a result, the problem of conducting valid inference reduces to finding a consistent estimator for $\boldsymbol{\Sigma}_n$ under our asymptotics of interest.

For our discussion of variance estimation, we introduce a new class of estimators. Let $\boldsymbol{\Omega}_{u,n} = \mathbb{E}[\mathbf{u}_n\mathbf{u}_n'|\mathcal{X}_n, \mathcal{W}_n]$ be the (conditional) variance-covariance matrix of the errors $\mathbf{u}_n = (u_{1,n}, \dots, u_{n,n})'$ and $L_n = \sum_{g=1}^{G_n}(\#\mathcal{T}_{g,n})^2$ the number of non-zero elements contained in it. We define a general class of cluster-robust estimators for $\boldsymbol{\Sigma}_n$ of the form

$$\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\kappa}_n) = \frac{1}{n}\sum_{g_1=1}^{G_n}\sum_{g_2=1}^{G_n}\sum_{i_1,j_1\in\mathcal{T}_{g_1,n}}\sum_{i_2,j_2\in\mathcal{T}_{g_2,n}} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{\mathbf{v}}_{i_1,n}\hat{\mathbf{v}}_{j_1,n}'\hat{u}_{i_2,n}\hat{u}_{j_2,n}, \tag{11}$$

where $\kappa_{g_1,g_2,i_1,j_1,i_2.j_2,n}$ is an entry of the $L_n \times L_n$ symmetric matrix $\boldsymbol{\kappa}_n = \boldsymbol{\kappa}_n(\mathbf{w}_{1,n}, \dots, \mathbf{w}_{1,n})$.[2] Notice that by setting $\boldsymbol{\kappa}_n = \mathbf{I}_{L_n}$ one obtains the usual cluster-robust estimator by White (1984):

$$\hat{\boldsymbol{\Sigma}}_n^{\mathtt{W}} \equiv \hat{\boldsymbol{\Sigma}}_n(\mathbf{I}_{L_n}) = \frac{1}{n}\sum_{g=1}^{G_n}\sum_{i,j\in\mathcal{T}_{g,n}} \hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{j,n}'\hat{u}_{i,n}\hat{u}_{j,n}. \tag{12}$$

The next theorem provides an asymptotic representation for this class of estimators.

Theorem 2 *Suppose Assumptions 1-3 hold.*
If $\|\boldsymbol{\kappa}_n\|_\infty = \max_{(g_1,i_1,j_1)}\sum_{g_2=1}^{G_n}\sum_{i_2,j_2\in\mathcal{T}_{g_2,n}}|\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| = O_p(1)$, *then*

$$\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\kappa}_n) = \frac{1}{n}\sum_{g_1=1}^{G_n}\sum_{g_2=1}^{G_n}\sum_{g_3=1}^{G_n}\sum_{i_1,j_1\in\mathcal{T}_{g_1,n}}\sum_{i_2,j_2\in\mathcal{T}_{g_2,n}}\sum_{i_3,j_3\in\mathcal{T}_{g_3,n}} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{\mathbf{v}}_{i_1,n}\hat{\mathbf{v}}_{j_1,n}'M_{i_2j_3,n}M_{j_2i_3,n}\mathbb{E}[u_{i_3,n}u_{j_3,n}|\mathcal{X}_n, \mathcal{W}_n] + o_p(1). \tag{13}$$

[1]From Assumptions 1-3 it also follows that $\hat{\boldsymbol{\Omega}}_n = O_p(1)$, implying that $\hat{\boldsymbol{\beta}}_n$ is $\sqrt{n}$-consistent.

[2]In particular, $\kappa_{g_1,g_2,i_1,j_1,i_2.j_2,n}$ corresponds to the $(h(g_1,i_1,j_1), h(g_2,i_2,j_2))$ entry of $\boldsymbol{\kappa}_n$, where $h(g,i,j) = [\sum_{k=0}^{(g-1)}(\#\mathcal{T}_{k,n})^2 + (\#\mathcal{T}_{g,n})(i-1) + j]$ and we adopt the convention that $\#\mathcal{T}_{0,n} = 0$.

Heuristically, in Theorem 2 consistency of $\hat{\boldsymbol{\beta}}_n$ implies that the estimated residuals $\hat{u}_{i,n}$ asymptotically converge to $\tilde{u}_{i,n} = \sum_{j=1}^{n} M_{ij,n} u_{j,n}$, which are only affected by the estimation noise due to projecting out the high-dimensional covariates $\mathbf{w}_{i,n}$.

The result of Theorem 2 has a central importance in our analysis. First, it immediately provides an explicit characterization for the asymptotic limit of White's estimator, as shown in the following corollary.

Corollary 1 *Suppose the assumptions of Theorem 2 hold. Then,*

$$\hat{\boldsymbol{\Sigma}}_n^{\mathtt{W}} = \frac{1}{n} \sum_{g_1=1}^{G_n} \sum_{g_2=1}^{G_n} \sum_{i_1, j_1 \in \mathcal{T}_{g_1,n}} \sum_{i_2, j_2 \in \mathcal{T}_{g_2,n}} \hat{\mathbf{v}}_{i_1,n} \hat{\mathbf{v}}'_{j_1,n} M_{i_1 j_2,n} M_{j_1 i_2,n} \mathbb{E}[u_{i_2,n} u_{j_2,n} | \mathcal{X}_n, \mathcal{W}_n] + o_p(1). \quad (14)$$

Corollary 1 implies that inference based on White's clustered standard errors is invalid in general under asymptotics where $K_n/n \nrightarrow 0$. In fact, $\hat{\boldsymbol{\Sigma}}_n^{\mathtt{W}}$ does not converge to the target $\boldsymbol{\Sigma}_n$ due to elements of $\mathbf{M}_n$ arising in its asymptotic limit. Notice that the sign of the asymptotic bias of White's estimator cannot be determined in general.

In addition, Theorem 2 suggests that a particular choice of $\boldsymbol{\kappa}_n$ might set the leading term in the expansion (13) equal to the target $\boldsymbol{\Sigma}_n$. Based on this insight, we define the estimator

$$\hat{\boldsymbol{\Sigma}}_n^{\mathtt{CR}} \equiv \hat{\boldsymbol{\Sigma}}(\boldsymbol{\kappa}_n^{\mathtt{CR}}) = \frac{1}{n} \sum_{g_1=1}^{G_n} \sum_{g_2=1}^{G_n} \sum_{i_1, j_1 \in \mathcal{T}_{g_1,n}} \sum_{i_2, j_2 \in \mathcal{T}_{g_2,n}} \kappa^{\mathtt{CR}}_{g_1,g_2,i_1,j_1,i_2,j_2,n} \hat{\mathbf{v}}_{i_1,n} \hat{\mathbf{v}}'_{j_1,n} \hat{u}_{i_2,n} \hat{u}_{j_2,n}, \quad (15)$$

where $\boldsymbol{\kappa}_n^{\mathtt{CR}}$ solves the system of $L_n(L_n-1)/2$ equations

$$\sum_{g_2=1}^{G_n} \sum_{i_2, j_2 \in \mathcal{T}_{g_2,n}} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n} M_{i_2,j_3,n} M_{j_2,i_3,n} = \mathbb{1}\{(g_1, i_1, j_1) = (g_3, i_3, j_3)\}, \quad (16)$$
$$1 \leq g_1, g_3 \leq G_n, i_1, j_1 \in \mathcal{T}_{g_1,n}, i_3, j_3 \in \mathcal{T}_{g_3,n}.$$

It also turns out that $\boldsymbol{\kappa}_n^{\mathtt{CR}}$ can be characterized in closed form as

$$\boldsymbol{\kappa}_n^{\mathtt{CR}} = (\mathbf{S}'_n(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n)^{-1}, \quad (17)$$

where $\otimes$ denotes the Kronecker product and $\mathbf{S}_n$ is the $n^2 \times L_n$ selection matrix with full column rank such that $\mathbf{S}'_n \mathrm{vec}(\boldsymbol{\Omega}_{u,n})$ is the $L_n \times 1$ vector containing the non-zero elements of $\boldsymbol{\Omega}_{u,n}$.

Remark 2 When $\boldsymbol{\Omega}_{u,n}$ is assumed to be diagonal, i.e. errors are (conditionally) heroskedastic, then $G_n = n$, $\mathcal{T}_{i,n} = \{i\}$, $L_n = n$, $\mathbf{S}'_n(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n = \mathbf{M}_n \odot \mathbf{M}_n$ and our estimator reduces to the heteroskedasticity-robust estimator of Cattaneo et al. (2017). ■

The following theorem establishes consistency of our proposed estimator.

Theorem 3 *Suppose Assumptions 1-3 hold.*
If $\mathbb{P}[\lambda_{min}(\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n) > 0] \to 1$ *and* $\|\boldsymbol{\kappa}_n^{\mathtt{CR}}\|_\infty = O_p(1)$, *then*

$$\hat{\boldsymbol{\Sigma}}_n^{\mathtt{CR}} = \boldsymbol{\Sigma}_n + o_p(1) \tag{18}$$

Since $\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n$ is observable, the first high-level condition in Theorem 3 is expected to be verified whenever $\lambda_{\min}(\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n)$ is invertible. The second high-level condition could be verified using Theorem 1 of Varah (1975), which provides a bound for $\|\boldsymbol{\kappa}_n^{\mathtt{CR}}\|_\infty$ under the condition that $\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n$ is diagonally dominant. In simulations we find that diagonal dominance typically does not hold but our high-level condition is verified in a wide range of models and designs, as shown in Section 5.[3]

Remark 3 The results of this paper can be immediately extended to a more general version of the variance estimators, described in Section F of the Appendix, that allows to impose within-cluster zero-restrictions on the variace-covariance matrix of the errors. In such form, our proposed estimator reduces to the one of Stock and Watson (2008) in the case of one-way fixed effects panel data models with zero restrictions on the conditional autocovariances of U_{it} within entities. While our results allow to cover a much wider class of models, they also partly improve on Stock and Watson (2008) as we do not require $(\mathbf{X}_{i1}', \ldots, \mathbf{X}_{iT}', U_{i1}, \ldots, U_{iT})$ to be i.i.d. and $(\mathbf{X}_{it}, U_{it})$ is not required to be stationary. ■

Although consistency of $\hat{\boldsymbol{\Sigma}}_n^{\mathtt{CR}}$ is derived under asymptotic sequences that allow but do not require $K_n/n \nrightarrow 0$, it is still desirable to establish consistency of White's estimator under some sufficiently slow rate of growth for K_n. Define $\mathbf{w}_{i,n}^* = \mathbf{w}_{i,n}\hat{\boldsymbol{\Sigma}}_{\mathbf{w},n}^{-1/2}$, where $\hat{\boldsymbol{\Sigma}}_{\mathbf{w},n}^{1/2}$ is the unique symmetric positive definite $K_n \times K_n$ matrix such that $\hat{\boldsymbol{\Sigma}}_{\mathbf{w},n}^{1/2}\hat{\boldsymbol{\Sigma}}_{\mathbf{w},n}^{1/2} = \frac{1}{n}\sum_{i=1}^n \mathbf{w}_n\mathbf{w}_n'$. The following theorem provides sufficient conditions for consistency of White's cluster-robust estimator.

Theorem 4 *Suppose Assumptions 1-3 hold and that* $\max_{i,j} \mathbb{E}[w_{ij,n}^{*2}] = O(1)$. *If* $K_n^2/n \to 0$, *then*

$$\hat{\boldsymbol{\Sigma}}_n^{\mathtt{W}} = \boldsymbol{\Sigma}_n + o_p(1). \tag{19}$$

[3]CJN (2018a) instead develop their theory under the requirement that $\mathbf{M}_n \odot \mathbf{M}_n$ is diagonally dominant. It would be interesting to investigate whether this requirement could be relaxed in practice.

Moreover, if $\mathbb{E}[U_{i,n}^2|\mathcal{X}_n, \mathcal{W}_n] = \sigma_n^2\ \forall i$, *and* $\mathbb{E}[U_{i,n}U_{j,n}|\mathcal{X}_n, \mathcal{W}_n] = 0\ \forall i \neq j$, *then (19) holds under* $K_n/n \to 0$.

Although we can only prove consistency of White's (1984) estimator under $K_n^2/n \to 0$, we speculate that $K_n/n \to 0$ might suffice in general. We leave the refinement of this result for future work.[4]

5 Simulations

This section reports the findings of a simulation study that investigates the finite sample behaviour of the cluster-robust variance estimators studied in this paper. We consider two distinct designs motivated by the empirical examples covered by the theoretical framework of this paper: the linear regression models with increasing dimension and the partially linear model.

5.1 Results - Linear regression model with increasing dimension

The chosen designs for our Monte Carlo experiments closely resemble those of CJN, while also borrowing from specifications in Stock and Watson (2008) and MacKinnon (2012). The data generating process (DGP) for the linear regression model with many covariates is:

$$\begin{aligned}
&y_{gi} = \beta x_{gi} + \boldsymbol{\gamma}_n' \mathbf{w}_{gi} + U_{gi}, \\
&x_{gi}|\mathbf{w}_{gi} \sim \mathcal{N}(0, \sigma_{x,gi}^2), \quad \sigma_{x,gi}^2 = \varkappa_x(1 + (\boldsymbol{\iota}'\mathbf{w}_i)^2), \\
&U_{gi} = (\rho \mathbb{1}(x_{gi} \geq 0) - \rho(1 - \mathbb{1}(x_{gi} \geq 0))U_{g,i-1} + \varepsilon_{gi}, \quad \varepsilon_{gi} \sim \mathcal{N}(0,1), \\
&u_{g1} \sim \mathcal{N}(0, \sigma_{u1}^2), \quad \sigma_{u1}^2 = \varkappa_{u1}(1 + (t(x_{g1}) + \boldsymbol{\iota}'\mathbf{w}_{g1})^2), \\
&i = 1, \dots, n/G, \quad g = 1, \dots, G, \quad n = 700,
\end{aligned} \tag{20}$$

where $\mathbf{w}_{gi} \overset{i.i.d.}{\sim} \mathcal{U}(-1,1)$, $\boldsymbol{\iota} = (1,1,\dots,1)'$, $\beta = 1$, $\boldsymbol{\gamma} = \mathbf{0}$, $\rho = 0.3$, the constants $\varkappa_x$ and $\varkappa_{u1}$ are chosen so that $\mathbb{V}[x_{gi}] = \mathbb{V}[U_{g1}] = 1$ and $t(a) = a\mathbb{1}(-2 \leq a \leq 2) + 2\mathrm{sgn}(a)(1 - \mathbb{1}(-2 \leq a \leq 2))$.

Table 2 reports the results of our experiment for five dimensions of $\mathbf{w}_{gi}$: $K \in \{1, 71, 141, 211, 281\}$, where the first covariate is an intercept, as well as three different numbers of equal-sized clusters: $G \in \{175, 70, 35\}$. We consider three different estimators for the variance of the OLS

[4] Theorem 4 also states that $K_n/n \to 0$ is sufficient for consistency of White's estimator in the special case of homoskedastic errors.

estimator $\hat{\beta}$: the unfeasible estimator based on $\hat{\mathbf{\Sigma}}_n^{\texttt{Unf}} = \frac{1}{n}\sum_{g=1}^{G_n}\sum_{i,j\in\mathcal{T}_{g,n}}\hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{j,n}'U_{i,n}U_{j,n}$ that makes use of the true error realizations, White's cluster-robust estimator and our proposed formula, as previously defined. For each of these estimators, we report the bias (expressed in percentage), the standard deviation (denoted by Std.) and the empirical coverage probability (denoted by $\hat{p};\alpha$) of the Gaussian confidence interval of the form:

$$\mathsf{I}_\ell \doteq \left[\hat{\beta} - \Phi^{-1}(1-\alpha/2)\cdot\sqrt{\frac{\hat{\Omega}_\ell}{n}}, \hat{\beta} - \Phi^{-1}(\alpha/2)\cdot\sqrt{\frac{\hat{\Omega}_\ell}{n}}\right], \quad \hat{\Omega}_\ell = \hat{\Gamma}^{-1}\hat{\Sigma}^\ell\hat{\Gamma}^{-1}, \qquad (21)$$

where Φ^{-1} denotes the inverse of the standard normal cumulative distribution function Φ, $\hat{\Sigma}_\ell$ with $\ell \in \{\texttt{Unf, W, CR}\}$ corresponds to the variance estimators already discussed and we set $\alpha = 0.05$.

The findings from this experiment are perfectly in line with our theoretical predictions. Firstly, we find that inference based on White's clustered standard errors formula is highly inaccurate. In fact, its bias quickly increases with the dimensionality of the model, resulting in poor coverage even for $K/n = 0.101$. On the other hand, the performance of our proposed estimator is excellent, with negligible bias and close-to-correct empirical coverage even for $K/n = 0.401$. Such improvement in inference accuracy compared to White's estimator is achieved in spite of a decrease in relative precision. As expected, the performance of all estimators is adversely affected by a reduction in the number of clusters. In Tables (8)-(10) we also report on the behaviour of $\|\boldsymbol{\kappa}_n\|_\infty$ in this design; we find that $\|\boldsymbol{\kappa}_n\|_\infty$ does not only seem to be bounded but even decreasing as n grows.[5] Analogous results are obtained for a different version of this experiment that considers independent and discrete controls constructed as $\mathbb{1}\{\mathcal{N}(0,1) \geq 1\}$, as reported in Table 3.

[5]Notice that diagonal dominance of $\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n$ does not to hold in any of the simulations we carried out.

5.2 Results - Semiparametric partially linear model

The experimental design chosen for the semiparametric partially linear model takes the form:

$$\begin{aligned}
&y_{gi} = \beta x_{gi} + g(\mathbf{z}_{gi}) + U_{gi}, \\
&x_{gi} = h(\mathbf{z}_{gi}) + v_i, \quad v_{gi}|\mathbf{z}_{gi} \sim \mathcal{N}(0, \sigma^2_{v,gi}), \quad \sigma^2_{vi} = \varkappa_v(1 + (\boldsymbol{\iota}'\mathbf{z}_{gi})^2), \\
&U_{gi} = (\rho \mathbb{1}(z_{1,gi} \geq 0) - \rho(1 - \mathbb{1}(z_{1,gi} \geq 0))U_{g,i-1} + \varepsilon_{gi}, \quad \varepsilon_{gi} \sim \mathcal{N}(0,1), \\
&u_{g1} \sim \mathcal{N}(0, \sigma^2_{u1}), \quad \sigma^2_{u1} = \varkappa_{u1}(1 + (t(x_{gi}) + \boldsymbol{\iota}'\mathbf{z}_{gi})^2), \\
&i = 1, \ldots, n/G, \quad g = 1, \ldots, G, \quad n = 700,
\end{aligned} \tag{22}$$

where $\dim(\mathbf{z}_{gi}) = 6$, $\mathbf{z}_{gi} = (z_{1,gi}, \ldots, z_{6,gi})'$ with $z_{\ell,gi} \sim \mathcal{U}(-1,1)$, $\ell = 1, \ldots, 6$. The unknown regressions functions are set to $g(\mathbf{z}_{gi}) = \exp(-\|\mathbf{z}_{gi}\|^{1/2})$ and $h(\mathbf{z}_{gi}) = \exp(\|\mathbf{z}_{gi}\|^{1/2})$, and the constants $\varkappa_v$ and $\varkappa_{u1}$ are again chosen so that $\mathbb{V}[x_{gi}] = \mathbb{V}[u_{g1}] = 1$. Similarly to the previous simulation, we set $\beta = 1$ and $\rho = 0.3$.

To construct the covariates $\mathbf{w}_{gi}$ entering the estimated linear regression model $y_{gi} = \boldsymbol{\beta}'\mathbf{x}_{gi} + \boldsymbol{\gamma}_n'\mathbf{w}_{gi} + u_{gi}$, we consider power series expansions. The table below gives a summary of the expansions considered, where $\mathbf{w}_{gi} = \mathbf{p}(\mathbf{z}_{gi}; K)$ for $K \in \{1, 7, 13, 28, 34, 84, 90, 210, 216\}$ is defined as:

Table 1: Polynomial Basis Expansion: $\dim(\mathbf{z}_{gi}) = 6$ and $n = 700$

K	$\mathbf{p}(\mathbf{z}_{gi}; K)$	K/n
1	1	0.001
7	$(1, z_{1,gi}, z_{2,gi}, z_{3,gi}, z_{4,gi}, z_{5,gi}, z_{6,gi})'$	0.010
13	$(\mathbf{p}(\mathbf{z}_{gi}; 7)', z^2_{1,gi}, z^2_{2,gi}, z^2_{3,gi}, z^2_{4,gi}, z^2_{5,gi}, z^2_{6,gi})'$	0.019
28	$\mathbf{p}(\mathbf{z}_{gi}; 13)$ + first-order interactions	0.040
34	$(\mathbf{p}(\mathbf{z}_{gi}; 28)', z^3_{1,gi}, z^3_{2,gi}, z^3_{3,gi}, z^3_{4,gi}, z^3_{5,gi}, z^3_{6,gi})'$	0.049
84	$\mathbf{p}(\mathbf{z}_{gi}; 13)$ + second-order interactions	0.120
90	$(\mathbf{p}(\mathbf{z}_{gi}; 84)', z^4_{1,gi}, z^4_{2,gi}, z^4_{3,gi}, z^4_{4,gi}, z^4_{5,gi}, z^4_{6,gi})'$	0.129
210	$\mathbf{p}(\mathbf{z}_{gi}; 90)$ + third-order interactions	0.300
216	$(\mathbf{p}(\mathbf{z}_{gi}; 210)', z^5_{1,gi}, z^5_{2,gi}, z^5_{3,gi}, z^5_{4,gi}, z^5_{5,gi}, z^5_{6,gi})'$	0.309

Source: Cattaneo et al. (2017, Supplemental Appendix).

The results for this experiment are given in Table 4, in which report only $K \in \{1, 13, 34, 90, 216\}$ for reasons of parsimony. The numerical findings are perfectly coherent with those reported for the linear model with increasing dimension. The main difference between this setting and the linear model with increasing dimension considered previously is that the unfeasible estimator that uses realizations of the true structural disturbances is free not just from estimation error but also specification error, which in turn affects White's and our proposed estimator for $K = 1$; in addition, the degree of (conditional) heteroskedasticity and dependence in the errors is invariant with respect to the dimensionality of the model, since it only depends on x_{gi} and $\mathbf{z}_{gi}$ but not $\mathbf{w}_{gi}$.

5.3 Results - Fixed effects panel data regression model

For fixed effects panel data regression model we consider the following specification:

$$y_{it} = \beta x_{it} + \alpha_i + e_{d_{it}} + u_{it}, \qquad i = 1, \ldots, N, \qquad t = 1, \ldots, T, \tag{23}$$

where α_i is time-invariant individual effect and $e_{d_{it}}$ are unobserved factors common to all observations sharing the same value of the indexing variable $d_{it} \in \{1, \ldots, N_d\}$. This model coincides with the one studied in Verdier (2018), whose theory and simulation results concern the case of two-way clustering. We instead consider the case of one-way clustering at the individual level as we postulate the following DGP:

$$\begin{aligned}
&y_{it} = \beta x_{it} + \alpha_i + e_{d_{it}} + U_{it}, \\
&x_{it}|\mathbf{z}_{it} \sim \mathcal{N}(0, \sigma^2_{x,it}), \quad \sigma^2_{x,gi} = \varkappa_x(1 + (\boldsymbol{\iota}'\mathbf{z}_{it})^2), \\
&U_{it} = (\rho \mathbb{1}(x_{it} \geq 0) - \rho(1 - \mathbb{1}(x_{it} \geq 0))U_{i,t-1} + \varepsilon_{it}, \quad \varepsilon_{it} \sim \mathcal{N}(0, 1), \\
&u_{i1} \sim \mathcal{N}(0, \sigma^2_{u1}), \quad \sigma^2_{u1} = \varkappa_{u1}(1 + (t(x_{i1}) + \boldsymbol{\iota}'\mathbf{z}_{i1})^2), \\
&i = 1, \ldots, N, \quad t = 1, \ldots, T,
\end{aligned} \tag{24}$$

where $\dim(\mathbf{z}_{it}) = 6$, $\mathbf{z}_{it} = (z_{1,it}, \ldots, z_{6,it})'$ with $z_{\ell,gi} \sim \mathcal{U}(-1, 1)$, $\ell = 1, \ldots, 6$, the constants $\varkappa_x$ and $\varkappa_{u1}$ are chosen so that $\mathbb{V}[x_{it}] = \mathbb{V}[U_{i1}] = 1$, the function $t(\cdot)$ is as previously defined and we set $\beta = 1$ and $\alpha_i = e_{d_{it}} = 0$. For estimation we transform (23) by partialling out the individual fixed effects α_i, so that the estimated model $\tilde{y}_{it} = \beta\tilde{x}_{it} + \tilde{e}_{d_{it}} + \tilde{u}_{it}$ has $\dim(\mathbf{w}_i) = N_d$.[6] We consider $T \in \{4, 10, 20\}$ and $G = N = [700/T]$ a well as $N_d = 700/r$ for $r \in \{700, 10, 5, 4, 3\}$, so that the total sample size is always roughly $n = 700$. Table 5

[6]The motivation for this transformation is that the correction matrix $\boldsymbol{\kappa}_n$ is not invertible when the controls $\mathbf{w}_{i,n}$ include indicators for the clusters (see, e.g., Stock and Watson, 2008). Notice that partialling out of the fixed effects does not affect the correlation structure of the errors.

reports the numerical findings of this experiment, which are consistent with our theoretical predictions and in line with the results obtained for the other simulation models.

Table 2: Monte Carlo simulations for linear regression model with increasing dimension (continuous controls), $n = 700$

	$\hat{\beta}$		Unfeasible			Classical			Robust		
	Mean	Variance	Bias (%)	Std.	$\hat{p}$; .05	Bias (%)	Std.	$\hat{p}$; .05	Bias (%)	Std.	$\hat{p}$; .05
$\boldsymbol{G = 175}$											
$K/n = 0.001$	1.00	.0042	1.29	.0011	.047	.101	.0011	.051	.360	.0011	.051
$K/n = 0.101$	1.00	.752	-1.42	.492	.044	-19.3	.375	.080	-4.45	.460	.055
$K/n = 0.201$	.981	2.89	-.578	2.09	.043	19.6	1.21	.106	-4.41	1.87	.053
$K/n = 0.301$	1.00	6.56	2.04	4.18	.042	38.0	2.02	.143	-1.71	3.93	.054
$K/n = 0.401$	1.02	11.7	-1.42	7.28	.044	-53.2	2.76	.179	-6.17	6.76	.057
$\boldsymbol{G = 70}$											
$K/n = 0.001$	1.00	.0027	-2.00	7.93×10^{-4}	.053	-4.02	7.73×10^{-4}	.064	-3.84	7.75×10^{-4}	.064
$K/n = 0.101$	1.01	.310	-3.16	.317	.039	-21.4	.235	.072	-7.15	.289	.052
$K/n = 0.201$	1.02	1.10	2.45	1.15	.036	-30.0	.702	.096	-1.80	1.07	.050
$K/n = 0.301$	1.04	2.60	1.77	2.67	.034	-42.5	1.273	.133	-3.32	2.48	.052
$K/n = 0.401$	1.01	4.82	-4.27	4.50	.039	-54.7	1.690	.181	-10.3	4.12	.058
$\boldsymbol{G = 35}$											
$K/n = 0.001$	1.00	.0021	.51	7.00×10^{-4}	.047	-2.86	6.80×10^{-4}	.062	-2.70	6.82×10^{-4}	.062
$K/n = 0.101$	1.00	.156	1.71	.027	.037	-18.3	.194	.065	-4.3	.238	.045
$K/n = 0.201$	1.01	.558	1.73	.858	.027	-31.6	.510	.089	-5.12	.770	.048
$K/n = 0.301$	1.00	1.33	-.16	1.81	.033	-45.6	.840	.140	-9.10	1.61	.070
$K/n = 0.401$	.973	2.41	.70	3.05	.037	-53.7	1.12	.186	-9.50	2.72	.083

Notes: Simulation results based on 5,000 replications. DGP as described in Equation 20.

Table 3: Monte Carlo simulations for linear regression model with increasing dimension (discrete controls), $n = 700$

	$\hat{\beta}$		Unfeasible			Classical			Robust		
	Mean	Variance	Bias (%)	Std.	$\hat{p}$; .05	Bias (%)	Std.	$\hat{p}$; .05	Bias (%)	Std.	$\hat{p}$; .05
$\boldsymbol{G = 175}$											
$K/n = 0.001$	1.00	.0043	-.46	.0011	.048	-1.70	.0011	.051	-1.50	.0011	.051
$K/n = 0.101$	1.00	.0019	-1.84	3.01×10^{-4}	.052	-18.4	2.60×10^{-4}	.072	-3.03	3.21×10^{-4}	.057
$K/n = 0.201$	1.00	.0020	1.63	3.10×10^{-4}	.048	-20.9	2.39×10^{-4}	.086	-.410	3.60×10^{-4}	.053
$K/n = 0.301$	1.00	.0023	-3.40	3.52×10^{-4}	.055	-34.2	2.41×10^{-4}	.114	-5.20	4.53×10^{-4}	.064
$K/n = 0.401$	1.00	.0026	-1.00	4.05×10^{-4}	.050	-41.5	2.44×10^{-4}	.140	-2.48	5.88×10^{-4}	.061
$\boldsymbol{G = 70}$											
$K/n = 0.001$	1.00	.0026	-.161	8.07×10^{-4}	.046	-2.17	7.88×10^{-4}	.056	-1.99	7.91×10^{-4}	.056
$K/n = 0.101$	1.00	.0018	-1.48	3.72×10^{-4}	.049	-14.4	3.27×10^{-4}	.075	-3.88	4.07×10^{-4}	.059
$K/n = 0.201$	1.00	.0020	-.773	4.00×10^{-4}	.052	-22.9	3.16×10^{-4}	.091	-3.26	4.87×10^{-4}	.062
$K/n = 0.301$	1.00	.0024	-4.56	4.60×10^{-4}	.054	-35.3	3.13×10^{-4}	.124	-8.16	6.11×10^{-4}	.071
$K/n = 0.401$	1.00	.0027	-1.16	5.20×10^{-4}	.053	-42.5	3.13×10^{-4}	.146	-5.96	8.00×10^{-4}	.076
$\boldsymbol{G = 35}$											
$K/n = 0.001$	1.00	.0021	.233	7.26×10^{-4}	.046	-2.97	7.06×10^{-4}	.063	-2.97	7.07×10^{-4}	.063
$K/n = 0.101$	1.00	.0018	1.29	4.85×10^{-4}	.046	-13.0	4.19×10^{-4}	.075	-2.77	5.21×10^{-4}	.062
$K/n = 0.201$	1.00	.0020	.570	5.32×10^{-4}	.046	-23.1	4.11×10^{-4}	.099	-4.24	6.37×10^{-4}	.069
$K/n = 0.301$	1.00	.0023	-.892	6.08×10^{-4}	.052	-34.2	4.07×10^{-4}	.122	-8.16	8.03×10^{-4}	.080
$K/n = 0.401$	1.00	.0028	-4.11	6.97×10^{-4}	.055	-45.4	4.06×10^{-4}	.160	-13.9	11.0×10^{-4}	.102

Notes: Simulation results based on 5,000 replications. DGP as described in Equation 20, with $w_{\ell,gi} = \mathbb{1}\{\mathcal{N}(0,1,) \geq 1\}, \quad \forall \ell, g, i.$

Table 4: Monte Carlo simulations for semiparametric partially linear model, $n = 700$

	$\hat{\beta}$		Unfeasible			Classical			Robust		
	Mean	Variance	Bias (%)	Std.	$\hat{p}$; .05	Bias (%)	Std.	$\hat{p}$; .05	Bias (%)	Std	$\hat{p}$; .05
$\boldsymbol{G = 175}$											
$K/n = 0.001$	.984	.0436	-1.42	.0192	.048	-2.84	.0182	.053	-2.70	.0183	.052
$K/n = 0.019$	1.00	.0519	-1.80	.0231	.046	-5.33	.0210	.052	-3.23	.0220	.049
$K/n = 0.049$	.994	.0515	2.22	.0225	.044	-5.14	.0195	.056	.680	.0219	.049
$K/n = 0.129$	.993	.0570	.922	.0231	.050	-15.0	.0174	.078	-.760	.0230	.056
$K/n = 0.309$	.993	.0733	.824	.0287	.049	-31.0	.0170	.103	-1.94	.0304	.056
$\boldsymbol{G = 70}$											
$K/n = 0.001$	.986	.0433	-.89	.0196	.050	-2.31	.0189	.054	-2.20	.0183	.054
$K/n = 0.019$	1.00	.0209	.28	.0132	.042	-4.13	.0120	.052	-2.08	.0125	.050
$K/n = 0.049$	1.00	.0218	-.59	.0136	.041	-8.36	.0117	.0548	-3.07	.0130	.049
$K/n = 0.129$	1.00	.0241	-.04	.0153	.046	-16.5	.0113	.074	-2.87	.0148	.059
$K/n = 0.309$	1.00	.0315	-1.44	.0172	.043	-33.0	.0100	.115	-5.68	.0174	.067
$\boldsymbol{G = 35}$											
$K/n = 0.001$	.988	.0098	-1.37	.0083	.042	-5.01	.0079	.059	-4.88	.0079	.059
$K/n = 0.019$	1.00	.0121	-3.54	.0108	.050	-9.40	.0095	.066	-3.41	.0099	.061
$K/n = 0.049$	1.00	.0116	-.41	.0092	.038	-9.70	.0079	.064	-4.64	.0087	.060
$K/n = 0.129$	1.00	.0127	1.70	.0100	.033	-16.0	.0073	.070	-3.02	.0096	.058
$K/n = 0.309$	1.00	.0169	1.29	.0134	.040	-31.9	.0074	.115	-6.50	.0118	.065

Notes: Simulation results based on 5,000 replications. DGP as described in Equation 22.

Table 5: Monte Carlo simulations for two-way fixed effects panel data regression model, $n = 700$

	$\hat{\beta}$		Unfeasible			Classical			Robust		
	Mean	Variance	Bias (%)	Std.	$\hat{p}$; .05	Bias (%)	Std.	$\hat{p}$; .05	Bias (%)	Std.	$\hat{p}$; .05
$\boldsymbol{G = 175}$											
$K/n = 0.001$	1.00	.0025	.435	7.02×10^{-4}	.045	-1.28	6.72×10^{-4}	.053	-1.28	6.72×10^{-4}	.053
$K/n = 0.100$	1.00	.0029	-1.26	7.74×10^{-4}	.048	-17.6	6.07×10^{-4}	.076	-3.08	7.82×10^{-4}	.056
$K/n = 0.200$	1.00	.0033	.350	8.30×10^{-4}	.046	-30.4	5.18×10^{-4}	.105	-2.02	8.80×10^{-4}	.056
$K/n = 0.250$	1.00	.0036	-1.22	8.71×10^{-4}	.052	-38.0	5.04×10^{-4}	.127	-3.86	9.95×10^{-4}	.062
$K/n = 0.333$	1.00	.0042	-1.51	9.61×10^{-4}	.051	-48.8	4.63×10^{-4}	.167	-5.18	.0012	.066
$\boldsymbol{G = 70}$											
$K/n = 0.001$	1.00	.0020	-2..30	5.15×10^{-4}	.049	-4.60	5.04×10^{-4}	.059	-4.50	5.05×10^{-4}	.059
$K/n = 0.100$	1.00	.0022	-1.58	5.83×10^{-4}	.049	-15.4	4.89×10^{-4}	.075	-3.99	6.04×10^{-4}	.061
$K/n = 0.200$	1.00	.0024	-.482	6.20×10^{-4}	.046	-25.9	4.41×10^{-4}	.100	-4.13	6.92×10^{-4}	.062
$K/n = 0.250$	1.00	.0026	.498	6.53×10^{-4}	.045	-30.8	4.32×10^{-4}	.111	-3.63	7.66×10^{-4}	.063
$K/n = 0.333$	1.00	.0029	-1.84	7.13×10^{-4}	.051	-41.2	4.11×10^{-4}	.136	-6.55	9.14×10^{-4}	.071
$\boldsymbol{G = 35}$											
$K/n = 0.001$	1.00	.0018	-3.66	5.26×10^{-4}	.049	-7.32	5.07×10^{-4}	.068	-7.19	5.08×10^{-4}	.068
$K/n = 0.100$	1.00	.0020	-1.98	5.87×10^{-4}	.046	-16.1	4.98×10^{-4}	.080	-6.30	4.98×10^{-4}	.066
$K/n = 0.200$	1.00	.0022	-1.21	6.22×10^{-4}	.044	-26.0	4.65×10^{-4}	.099	-6.70	7.20×10^{-4}	.068
$K/n = 0.250$	1.00	.0023	-.057	6.83×10^{-4}	.045	-29.7	4.83×10^{-4}	.109	-5.45	8.41×10^{-4}	.073
$K/n = 0.333$	1.00	.0027	-2.90	7.21×10^{-4}	.050	-40.8	4.43×10^{-4}	.145	-10.7	9.72×10^{-4}	.093

Notes: Simulation results based on 5,000 replications. DGP as described in Equation 24.

6 Empirical Illustration

This section provides an illustration of the inference methods discussed in this paper using an application of the gravity model for trade to a real data set. Since its original formulation by Tinbergen (1962), the gravity model has been widely used in international economics to study how different factors such as geographical distance, trade agreements and other trade frictions influence trade flows between two countries. In particular, we use a light version of the CEPII 'Gravity' Dataset constructed by Head et al. (2010), which contains data about trade flows and geographical distance for all world pairs of countries, along with other relevant variables, for the period 1984-2006.[7] We consider an application of the linear regression model to the following specification of the gravity equation for trade:

$$y_{ijt} = \beta d_{ij} + \alpha_{it} + \theta_{jt} + \varepsilon_{ijt} \tag{25}$$

where y_{ijt} is log-trade volume from country i to country j at time t, d_{ij} is log-distance, α_{it} is a country/time-specific exporter fixed effect and θ_{jt} is a country/time-specific importer fixed effect. The gravity equation in (25) can be easily recasted into the linear regression model (1) by collecting the fixed effects into $\boldsymbol{\gamma}_n$, with $\mathbf{w}_{i,n}$ being the high-dimensional vector that selects the relevant fixed effects. For our illustration, we restrict the sample to observations in which both countries i and j are members of the Eurozone in the period 1999-2006, giving a total of $K_n = 188$ and $n = 840$ ($K_n/n = 0.224$), and we consider clustering at the country-pair/year level, resulting in $G_n = 420$.

The empirical findings are reported in Table 6. We find that the estimated elasticity of trade with respect to distance is close to minus one ($\hat{\beta} = -1.092$), a result which is line with estimates previously found in the literature (see, e.g., Disdier and Head, 2008). Although the statistical significance of the elasticity coefficient does not depend on the inference method employed in this case, we find that standard errors based on our proposed variance estimator are 16 per cent bigger than those based on White's estimator. Such considerable difference highlights the relevance of our proposed inference method for applications of the linear regression model in the presence of possibly many controls; in this application, the inclusion of many controls arises naturally as a way to control for unobserved time-varying heterogeneity in country characteristics. The next subsection provides a calibrated simulation experiment that provides further evidence on the performance of the inference methods studied in this

[7] The dataset is available at `http://www.cepii.fr/CEPII/fr/bdd_modele/download.asp?id=8`.

paper in this type of application.

Table 6: Empirical illustration - Gravity model

Outcome: Log-trade		
$\frac{K}{n}$	0.224	
$\Vert\boldsymbol{\kappa}\Vert_\infty$	2.067	
$\hat{\beta}$	-1.092	
	Std. Err.	**p-value**
(1) White's	0.0504	< 0.001
(2) Robust	0.0594	< 0.001
	Ratio (2)/(1)	
	1.18	

Data source: CEPII.

6.1 Calibrated Monte Carlo simulations

Here we report the results a Monte Carlo experiment calibrated to the CEPII dataset used in the empirical illustration of this paper. In particular, the DGP for this experiment is:

$$y_{ijt} = \beta d_{ij} + \alpha_{it} + \theta_{jt} + \varepsilon_{ijt}, \tag{26}$$
$$\begin{pmatrix} \varepsilon_{ijt} \\ \varepsilon_{jit} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho \\ \rho & \sigma^2 \end{pmatrix}\right],$$

where the value of the regressor d_{ij} is taken directly from the dataset, the parameters $\rho = 0.575$, $\sigma = 0.288$, α_{it}, θ_{jt} are calibrated to the estimates obtained for model (26) in the same sample, and we set $\beta = 0$. As in the empirical application previously considered in this section, the resulting linear regression model has $K_n = 188$, $n = 840$ ($K_n/n = 0.224$) and $G_n = 420$.

Table 7 reports the results of this experiment, which show how the variance estimator based on White's estimator is affected by considerable bias, as reflected by the poor coverage of the corresponding confidence interval. On the other hand, the variance estimator based on our proposed estimator displays negligible bias and close-to-correct coverage of the corresponding confidence interval. These results are in line with those obtained in Section 5, and motivate the use of our proposed inference method in the type of application considered in this section.

Table 7: Calibrated Monte Carlo simulations - Gravity model

	Mean	**Variance**	
$\hat{\beta}$	0.0014	0.0020	
	Bias (%)	**Std.**	$\hat{\mathbf{p}}$**; 05**
Unfeasible	.006	3.252×10^{-4}	.0480
White's	-22	2.600×10^{-4}	.0844
Robust	2	3.958×10^{-4}	.0540

Notes: Simulation results based on 5,000 replications. DGP as described in Equation 26.
Data source: CEPII.

7 Conclusion

This paper provides inference results for the OLS estimator of a subset of coefficients in linear regression models with many controls and clustering. We show that the usual White's cluster-robust variance estimator does not deliver consistent standard errors when the number of controls is a non-vanishing fraction of the sample size and we propose a new clustered standard error formula that is robust to the inclusion of many controls. Monte Carlo evidence supports our theoretical results and shows that our proposed variance estimator performs well in finite samples.

While our results are presented for the case of one-way clustering, we expect that they can be easily adapted to multi-way clustering structures. It would also be of interest to investigate whether the analysis of this paper could be extended to cases where variance estimation does not rely on zero-restrictions on the covariance matrix of the errors, e.g. when time series or spacial dependence in the errors is assumed.

References

ANATOLYEV, S. (2012), "Inference in Regression Models with Many Regressors", *Journal of Econometrics*, 170(2), 368-382.

ANGRIST, J., and HAHN, J. (2004), "When to Control for Covariates? Panel Asymptotics for Estimates of Treatment Effects," *The Review of Economics and Statistics*, 86(1), 58-72.

ARELLANO, M. (1987), "Computing Robust Standard Errors for Within-Group Estimators", *Oxford Bulletin of Economics and Statistic*, 49(4), 431-434.

BELL, R. M. and McCAFFREY, D. F. (2002), "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples," *Survey Methodology*, 28, 169-181.

CAMERON, A. C. and MILLER, D. L. (2015), "A Practitioner's Guide to Cluster-Robust Inference," *Journal of Human Resources*, 50(2), 317-373.

CAMERON, A. C., GELBACH, J.B. and MILLER, D. L. (2008), "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples," *The Review of Economics and Statistics*, 98(4), 701-712.

CATTANEO, M. D., JANSSON, M. and NEWEY, W. K. (2018a), "Inference in Linear Regression Models with Many Covariates and Heteroskedasticity," *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2017.1328360.

CATTANEO, M. D., JANSSON, M. and NEWEY, W. K. (2018b), "Alternative Asymptotics and the Partially Linear Model with Many Regressors," *Econometric Theory*, 34(2), 277-301.

HEAD, K., MAYER, T. and RIES, J. (2010), "The Erosion of Colonial Trade Linkages after Independence", *Journal of International Economics*, 81(1), 1-14.

HUBER, P. J . (1973), "Robust Regression: Asymptotics, Conjectures and Monte Carlo," *Annals of Statistics*, 1(5), 799-821.

IBRAGIMOV, R. and MÜLLER, U. K. (2016), "Inference with Few Heterogeneous Clusters," *The Review of Economics and Statistics*, 98(1), 83-96.

KOENKER, R. (1988), "Asymptotic Theory and Econometric Practice," *Journal of Applied Econometrics*, 3(2), 139-147.

LI, C. and MÜLLER, U. K. (2017), "Linear Regression with Many Controls of Limited Expanatory Power," working paper, Princeton University.

MAMMEN, E. (1993), "Bootstrap and Wild Bootstrap for High Dimensional Linear Models," *Annals of Statistics*, 21(1), 255-285.

MACKINNON, J. G. (2012), "Thirty Year of Heteroskedasticity-Robust Inference," in *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, ed. by X. Chen, and N. R. Swanson. Springer.

STOCK, J. H. and WATSON, M. W. (2008), "Heteroskedasticity-Robust Standard Errors for Fixed Effects Panel Data Regression," *Econometrica*, 76(1), 155-174.

VERDIER, V. (2017), "Estimation and Inference for Linear Models with Two-Way Fixed Effects and Sparsely Matched Data," Working paper, University of North Carolina.

WHITE, H. (1984) *Asymptotic Theory for Econometricians*, San Diego: Academic Press.

WOOLDRIDGE, J. M. (2003), "Cluster-Sample Methods in Applied Econometrics," *American Economic Review*, 93(2), 133-138.

Table 8: Absolute row sum of $\boldsymbol{\kappa}_n$ for $K_n/n = 0.200$

	$\boldsymbol{G = 140}$		
$n = 250$	$n = 500$	$n = 750$	$n = 1000$
4.93	4.23	3..95	3.80
(0.26)	(0.14)	(0.11)	(0.090)
	$\boldsymbol{G = 70}$		
$n = 250$	$n = 500$	$n = 750$	$n = 1000$
8.41	6.71	6.11	5.81
(0.56)	(0.24)	(0.18)	(0.14)
	$\boldsymbol{G = 35}$		
$n = 240$	$n = 500$	$n = 740$	$n = 1000$
18.1	12.4	11.1	10.1
(1.35)	(0.50)	(0.38)	(0.25)

Notes: 250 repetitions. Standard deviations in parenthesis. DGP as described in Equation 20.

Table 9: Absolute row sum of $\boldsymbol{\kappa}_n$ for $K_n/n = 0.300$

	$\boldsymbol{G = 140}$		
$n = 250$	$n = 500$	$n = 750$	$n = 1000$
9.36	7.66	7.12	6.77
(0.65)	(0.32)	(0.21)	(0.17)
	$\boldsymbol{G = 70}$		
$n = 250$	$n = 500$	$n = 750$	$n = 1000$
17.5	13.2	11.9	11.2
(1.10)	(0.67)	(0.37)	(0.29)
	$\boldsymbol{G = 35}$		
$n = 240$	$n = 500$	$n = 740$	$n = 1000$
41.2	26.12	22.6	20.7
(2.93)	(1.13)	(0.77)	(0.55)

Notes: 250 repetitions. Standard deviations in parenthesis. DGP as described in Equation 20.

Table 10: Absolute row sum of $\boldsymbol{\kappa}_n$ for $K_n/n = 0.400$

	$G = 140$		
$n = 250$	$n = 500$	$n = 750$	$n = 1000$
18.36 (1.42)	14.5 (0.64)	13.3 (0.50)	12.6 (0.36)
	$G = 70$		
$n = 250$	$n = 500$	$n = 750$	$n = 1000$
37.3 (3.26)	26.6 (1.15)	23.8 (0.84)	22.0 (0.59)
	$G = 35$		
$n = 240$	$n = 500$	$n = 740$	$n = 1000$
100 (8.30)	59.9 (2.64)	47.6 (1.77)	43.0 (1.20)

Notes: 250 repetitions. Standard deviations in parenthesis. DGP as described in Equation 20.

Table 11: Absolute row sum of $\boldsymbol{\kappa}_n$ for $K_n/n = 0.200$

	$G = 140$		
$n = 250$	$n = 500$	$n = 750$	$n = 1000$
4.93 (0.26)	4.23 (0.14)	3..95 (0.11)	3.80 (0.090)
	$G = 70$		
$n = 250$	$n = 500$	$n = 750$	$n = 1000$
8.41 (0.56)	6.71 (0.24)	6.11 (0.18)	5.81 (0.14)
	$G = 35$		
$n = 240$	$n = 500$	$n = 740$	$n = 1000$
18.1 (1.35)	12.4 (0.50)	11.1 (0.38)	10.1 (0.25)

Notes: 250 repetitions. Standard deviations in parenthesis. DGP as described in Equation 20.

Appendix

This appendix is organized as follows. Section A presents the assumptions and the variance estimators studied in this paper for the fully general case that allows for misspecification bias in the model. Section B presents the main results of the paper under the setup described in Section A. Section C presents the technical lemmas needed to establish the main results of the paper. Section D provides the proofs for the main results of the paper. Section E provides the proofs for the technical lemmas. Section F presents an extension of the variance estimators studied in the paper which allows to impose within-cluster zero-restrictions on the variance-covariance matrix of the errors.

A Setup - general case

A.1 Assumptions

Suppose that $\{(y_{i,n}, \mathbf{x}'_{i,n}, \mathbf{w}'_{i,n}) : 1 \leq i \leq n\}$ is generated by

$$y_{i,n} = \boldsymbol{\beta}'\mathbf{x}_{i,n} + \boldsymbol{\gamma}'_n\mathbf{w}_{i,n} + u_{i,n}, \qquad i = 1, \dots, n, \tag{1}$$

for which $\mathcal{W}_n$ is a collection of random variables such that $\mathbb{E}[\mathbf{w}_{i,n}|\mathcal{W}_n] = \mathbf{w}_{i,n}$ and we set $\mathcal{X}_n = (\mathbf{x}_{1,n}, \dots, \mathbf{x}_{n,n})$. We define the following quantities:

$$\begin{aligned}
\varrho_n &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[R_{i,n}^2], \qquad R_{i,n} = \mathbb{E}[u_{i,n}|\mathcal{X}_n, \mathcal{W}_n], \\
\rho_n &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[r_{i,n}^2], \qquad r_{i,n} = \mathbb{E}[u_{i,n}|\mathcal{W}_n], \\
\chi_n &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\mathbf{Q}_{i,n}\|^2], \qquad \mathbf{Q}_{i,n} = \mathbb{E}[\mathbf{v}_{i,n}|\mathcal{W}_n], \\
\hat{\mathbf{\Gamma}}_n &= \sum_{i=1}^{n}\hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}'_{i,n}/n, \qquad \mathbf{\Sigma}_n = \mathbb{V}[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\mathbf{v}}_{i,n}U_{i,n}|\mathcal{X}_n, \mathcal{W}_n],
\end{aligned} \tag{2}$$

where $\mathbf{v}_{i,n} = \mathbf{x}_{i,n} - (\sum_{j=1}^{n}\mathbb{E}[\mathbf{x}_{j,n}\mathbf{w}'_{j,n}])(\sum_{j=1}^{n}\mathbb{E}[\mathbf{w}_{j,n}\mathbf{w}'_{j,n}])^{-1}\mathbf{w}_{i,n}$ is the population counterpart of $\hat{\mathbf{v}}_{i,n}$. Also, letting $\lambda_{\min}(\cdot)$ denote the minimum eigenvalue of its argument, define

$$\mathcal{C}_n = \max_{1\leq i\leq n}\{\mathbb{E}[U_{i,n}^4|\mathcal{X}_n, \mathcal{W}_n] + \mathbb{E}[\|\mathbf{V}_{i,n}\|^4|\mathcal{W}_n] + 1/\mathbb{E}[U_{i,n}^2|\mathcal{X}_n, \mathcal{W}_n]\} + 1/\lambda_{\min}(\mathbb{E}[\tilde{\mathbf{\Gamma}}_n|\mathcal{W}_n]) \tag{3}$$

where $U_{i,n} = y_{i.n} - \mathbb{E}[y_{i,n}|\mathcal{X}_n, \mathcal{W}_n]$, $\mathbf{V}_{i,n} = \mathbf{x}_{i,n} - \mathbb{E}[\mathbf{x}_{i,n}|\mathcal{W}_n]$, $\tilde{\mathbf{\Gamma}}_n = \sum_{i=1}^{n}\tilde{\mathbf{V}}_{i,n}\tilde{\mathbf{V}}'_{i,n}/n$ and $\tilde{\mathbf{V}}_{i,n} = \sum_{j=1}^{n}M_{ij,n}\mathbf{V}_{i,n}$.

We impose the following three assumptions:

Assumption 1* *$\max_{1\leq g\leq G_n}\#\mathcal{T}_{g,n} = O(1)$, where $\#\mathcal{T}_{g,n}$ is the cardinality of $\mathcal{T}_{g,n}$ and where $\{\mathcal{T}_{g,n} : 1 \leq g \leq G_n\}$ is a partition of $\{1,\ldots,n\}$ such that $\{(U_{i,n}, \mathbf{x}'_{i,n}) : i \in \mathcal{T}_{g,n}\}$ are independent over g conditional on $\mathcal{W}_n$.*

Assumption 2* *$\mathbb{P}[\lambda_{min}(\sum_{i=1}^{n}\mathbf{w}_{i,n}\mathbf{w}'_{i,n}) > 0] \to 1$, $\limsup_{n\to\infty} K_n/n < 1$, $\mathcal{C}_n = O_p(1)$ and $\mathbf{\Sigma}_n^{-1} = O_p(1)$*

Assumption 3* *$\chi_n = O(1)$, $\varrho_n + n(\varrho_n - \rho_n) + n\chi_n\varrho_n = o(1)$, and $\max_{1\leq i\leq n}\|\hat{\mathbf{v}}_{i,n}\|/\sqrt{n} = o_p(1)$.*

The only difference with the simplified set of assumptions presented in Section 3 of this paper is that we now allow for misspecification bias, i.e. $\mathbb{E}[u_i|\mathcal{X}_n, \mathcal{W}_n] \neq 0$. In particular, Assumption 3* now also includes conditions on ϱ_n and ρ_n, which are requirements on the quality of the linear approximation for the conditional expectations $\mathbb{E}[y_{i,n}|\mathcal{X}_n, \mathcal{W}_n]$ and $\mathbb{E}[y_{i,n}|\mathcal{W}_n]$, respectively. The misspecification bias is required to vanish asymptotically, thus ruling out the presence of lagged outcomes in the model. Notice that when no misspecification bias is present one gets $\varrho_n = \rho_n = 0$ and this set of assumptions reduces to the one presented in Section 3 of the paper.

A.2 Variance estimators

Let $\mathbf{\Omega}_{U,n} = \mathbb{E}[\mathbf{U}_n\mathbf{U}'_n|\mathcal{X}_n, \mathcal{W}_n]$ be the (conditional) variance-covariance matrix of the errors $\mathbf{U}_n = (U_{1,n}, \ldots, U_{n,n})'$ and $L_n = \sum_{g=1}^{G_n}(\#\mathcal{T}_{g,n})^2$ the number of non-zero elements contained in it. We define a general class of cluster-robust estimators for $\mathbf{\Sigma}_n$ of the form

$$\hat{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n) = \frac{1}{n}\sum_{g_1=1}^{G_n}\sum_{g_2=1}^{G_n}\sum_{i_1,j_1\in\mathcal{T}_{g_1,n}}\sum_{i_2,j_2\in\mathcal{T}_{g_2,n}}\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{\mathbf{v}}_{i_1,n}\hat{\mathbf{v}}'_{j_1,n}\hat{u}_{i_2,n}\hat{u}_{j_2,n}, \tag{4}$$

where $\kappa_{g_1,g_2,i_1,j_1,i_2.j_2,n}$ is an entry of the $L_n \times L_n$ symmetric matrix $\boldsymbol{\kappa}_n = \boldsymbol{\kappa}_n(\mathbf{w}_{1,n}, \ldots, \mathbf{w}_{1,n})$.[8]

Furthermore, define

$$\boldsymbol{\kappa}_n^{\mathtt{CR}} = (\mathbf{S}'_n(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n)^{-1}, \tag{5}$$

where $\otimes$ denotes the Kronecker product and $\mathbf{S}_n$ is the $n^2 \times L_n$ selection matrix with full column rank such that $\mathbf{S}'_n\text{vec}(\mathbf{\Omega}_{U,n})$ is the $L_n \times 1$ vector containing the non-zero elements of $\mathbf{\Omega}_{U,n}$. Our proposed cluster-robust estimator is then defined as

$$\hat{\mathbf{\Sigma}}_n^{\mathtt{CR}} \equiv \hat{\mathbf{\Sigma}}(\boldsymbol{\kappa}_n^{\mathtt{CR}}) = \frac{1}{n}\sum_{g_1=1}^{G_n}\sum_{g_2=1}^{G_n}\sum_{i_1,j_1\in\mathcal{T}_{g_1,n}}\sum_{i_2,j_2\in\mathcal{T}_{g_2,n}}\kappa^{\mathtt{CR}}_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{\mathbf{v}}_{i_1,n}\hat{\mathbf{v}}'_{j_1,n}\hat{u}_{i_2,n}\hat{u}_{j_2,n}. \tag{6}$$

[8]In particular, $\kappa_{g_1,g_2,i_1,j_1,i_2.j_2,n}$ corresponds to the $(h(g_1, i_1, j_1), h(g_2, i_2, j_2))$ entry of $\boldsymbol{\kappa}_n$, where $h(g, i, j) = [\sum_{k=0}^{(g-1)}(\#\mathcal{T}_{k,n})^2 + (\#\mathcal{T}_{g,n})(i-1) + j]$ and we adopt the convention that $\#\mathcal{T}_{0,n} = 0$.

B Main results - general

In this section we present the generalization of the main results of the paper to the case of potential misspecification bias in the model.

The first theorem establishes asymptotic normality of the OLS estimator for $\boldsymbol{\beta}_n$.

Theorem 1* *Suppose Assumptions 1*-3* hold. Then,*

$$\boldsymbol{\Omega}_n^{-1/2}\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d), \qquad \boldsymbol{\Omega}_n = \hat{\boldsymbol{\Gamma}}_n^{-1}\boldsymbol{\Sigma}_n\hat{\boldsymbol{\Gamma}}_n^{-1}, \tag{7}$$

where $\boldsymbol{\Sigma}_n = \frac{1}{n}\sum_{g=1}^{G_n}\sum_{i,j\in\mathcal{T}_{g,n}} \hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}'_{j,n}\mathbb{E}[U_{i,n}U_{j,n}|\mathcal{X}_n, \mathcal{W}_n]$.

The second theorem provides an asymptotic representation for the general class of variance estimators defined in (4).

Theorem 2* *Suppose Assumptions 1*-3* hold.*
If $\|\boldsymbol{\kappa}_n\|_\infty = \max_{(g_1,i_1,j_1)}\sum_{g_2=1}^{G_n}\sum_{i_2,j_2\in\mathcal{T}_{g_2,n}}|\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| = O_p(1)$, *then*

$$\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\kappa}_n) = \frac{1}{n}\sum_{g_1=1}^{G_n}\sum_{g_2=1}^{G_n}\sum_{g_3=1}^{G_n}\sum_{i_1,j_1\in\mathcal{T}_{g_1,n}}\sum_{i_2,j_2\in\mathcal{T}_{g_2,n}}\sum_{i_3,j_3\in\mathcal{T}_{g_3,n}} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{\mathbf{v}}_{i_1,n}\hat{\mathbf{v}}'_{j_1,n}M_{i_2j_3,n}M_{j_2i_3,n}\mathbb{E}[U_{i_3,n}U_{j_3,n}|\mathcal{X}_n, \mathcal{W}_n] + o_p(1). \tag{8}$$

Corollary 1* characterizes the asymptotic limit of White's estimator.

Corollary 1* *Suppose the assumptions of Theorem 2* hold. Then,*

$$\hat{\boldsymbol{\Sigma}}_n^{\mathtt{W}} = \frac{1}{n}\sum_{g_1=1}^{G_n}\sum_{g_2=1}^{G_n}\sum_{i_1,j_1\in\mathcal{T}_{g_1,n}}\sum_{i_2,j_2\in\mathcal{T}_{g_2,n}} \hat{\mathbf{v}}_{i_1,n}\hat{\mathbf{v}}'_{j_1,n}M_{i_1j_2,n}M_{j_1i_2,n}\mathbb{E}[U_{i_2,n}U_{j_2,n}|\mathcal{X}_n, \mathcal{W}_n] + o_p(1).$$

The third theorem establishes consistency of our proposed estimator.

Theorem 3* *Suppose Assumptions 1*-3* hold.*
If $\mathbb{P}[\lambda_{min}(\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n) > 0] \to 1$ *and* $\left\|\boldsymbol{\kappa}_n^{\mathtt{CR}}\right\|_\infty = O_p(1)$, *then*

$$\hat{\boldsymbol{\Sigma}}_n^{\mathtt{CR}} = \boldsymbol{\Sigma}_n + o_p(1) \tag{9}$$

Finally, the fourth theorem provides sufficient conditions for consistency of White's estimator.

Theorem 4* *Suppose Assumptions 1-3 hold and that* $\max_{i,j} \mathbb{E}[w^{*2}_{ij,n}] = O(1)$*. If* $K_n^2/n \to 0$*, then*

$$\hat{\mathbf{\Sigma}}_n^{\mathtt{W}} = \mathbf{\Sigma}_n + o_p(1). \tag{10}$$

Moreover, if $\mathbb{E}[U_{i,n}^2|\mathcal{X}_n, \mathcal{W}_n] = \sigma_n^2\ \forall i$*, and* $\mathbb{E}[U_{i,n}U_{j,n}|\mathcal{X}_n, \mathcal{W}_n] = 0\ \forall i \neq j$*, then (10) holds under* $K_n/n \to 0$*.*

C Technical Lemmas

Here we present the technical lemmas needed to establish the main results of the paper.[9]

The first lemma can be used to approximate $\hat{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n)$ by means of $\tilde{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n)$, where

$$\hat{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n) = \frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n} \hat{\mathbf{v}}_{i_1,n} \hat{\mathbf{v}}'_{j_1,n} \hat{u}_{i_2,n} \hat{u}_{j_2,n}$$

$$\tilde{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n) = \frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n} \hat{\mathbf{v}}_{i_1,n} \hat{\mathbf{v}}'_{j_1,n} \tilde{U}_{i_2,n} \tilde{U}_{j_2,n}, \qquad \tilde{U}_{i,n} = \sum_{j=1}^{n} M_{ij,n} U_{j,n}$$

Lemma 1 *Suppose Assumptions 1-3 hold. If* $\|\boldsymbol{\kappa}_n\|_\infty = O_p(1)$*, then*

$$\hat{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n) = \mathbb{E}[\tilde{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n)|\mathcal{X}_n, \mathcal{W}_n] + o_p(1).$$

The second lemma can be combined with Lemma 1 to show consistency of $\hat{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n)$ under a high-level condition.

Lemma 2 *Suppose Assumption 2 holds. If*

$$\max_{(g_1,i_1,j_1)} \Big\{ \Big| \sum_{(g_2,i_2,j_2)} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n} M_{i_1 j_2,n} M_{j_1 i_2,n} - 1 \Big| + \sum_{(g_3,i_3,j_3) \neq (g_1,i_1,j_1)} \Big| \sum_{(g_2,i_2,j_2)} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n} M_{i_3 j_2,n} M_{j_3 i_2,n} \Big| \Big\} = o_p(1),$$

then $\mathbb{E}[\tilde{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n)|\mathcal{X}_n, \mathcal{W}_n] = \mathbf{\Sigma}_n + o_p(1)$*.*

The third lemma gives sufficient conditions for the condition of Lemma 2 for our proposed estimator $\hat{\mathbf{\Sigma}}_n^{\mathtt{CR}}$.

Lemma 3 *Suppose Assumption 2 holds. If* $\mathbb{P}[\lambda_{min}(\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n) > 0] \to 1$*, then*

$$\mathbb{E}[\tilde{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n^{\mathtt{CR}})|\mathcal{X}_n, \mathcal{W}_n] = \mathbf{\Sigma}_n + o_p(1).$$

[9]Throughout the Technical Lemmas we adopt the notational convention $\sum_{(g,i,j)} \equiv \sum_{g=1}^{G_n} \sum_{i,j \in \mathcal{T}_{g,n}}$

with $\boldsymbol{\kappa}_n^{\texttt{CR}} = (\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n)^{-1}$.

The fourth lemma finds sufficient conditions for the condition of Lemma 2 for White's estimator $\hat{\boldsymbol{\Sigma}}_n^{\texttt{W}}$.

Lemma 4 *Suppose Assumption 2 holds and* $\boldsymbol{\kappa}_n = \mathbf{I}_{L_n}$. *Also define* $\mathbf{w}_{i,n}^* = \mathbf{w}_{i,n}\hat{\boldsymbol{\Sigma}}_{\mathbf{w},n}^{-1/2}$, *where* $\hat{\boldsymbol{\Sigma}}_{\mathbf{w},n}^{1/2}$ *is the unique symmetric positive definite* $K_n \times K_n$ *matrix such that* $\hat{\boldsymbol{\Sigma}}_{\mathbf{w},n}^{1/2}\hat{\boldsymbol{\Sigma}}_{\mathbf{w},n}^{1/2} = \frac{1}{n}\sum_{i=1}^n \mathbf{w}_n\mathbf{w}_n'$. *If* $\max_{i,j}\mathbb{E}[w_{ij,n}^{*2}] = O(1)$ *and* $K_n = o(n^{1/2})$, *then*

$$\mathbb{E}[\tilde{\boldsymbol{\Sigma}}_n(\boldsymbol{\kappa}_n)|\mathcal{X}_n, \mathcal{W}_n] = \boldsymbol{\Sigma}_n + o_p(1).$$

Finally, the fifth lemma establishes sufficient conditions for the condition of Lemma 2 for White's estimator for the special case of homoskedastic errors.

Lemma 5 *Suppose Assumption 2 holds and* $\boldsymbol{\kappa}_n = \mathbf{I}_{L_n}$. *If* $\max_{i,j}\mathbb{E}[w_{ij,n}^{*2}] = O(1)$, $K_n = o(n)$, $\mathbb{E}[U_{i,n}^2|\mathcal{X}_n, \mathcal{W}_n] = \sigma_n^2 \quad \forall i$ *and* $\mathbb{E}[U_{i,n}U_{j,n}|\mathcal{X}_n, \mathcal{W}_n] = 0 \quad \forall i \neq j$, *then*

$$\mathbb{E}[\tilde{\boldsymbol{\Sigma}}_n(\boldsymbol{\kappa}_n)|\mathcal{X}_n, \mathcal{W}_n] = \boldsymbol{\Sigma}_n + o_p(1).$$

D Proof of Main Results

Theorem 1* follows from Lemma SA-1 and Lemma SA-2 in CJN (2018a), combined with the fact that $\boldsymbol{\Sigma}_n^{-1} = O_p(1)$ in Assumption 2. Theorem 2* follows from Theorem 1* combined with Lemma 1. Theorem 3* follows from Theorem 2* combined with Lemma 2 and 3. Theorem 4* follows from Theorem 2* combined with Lemma 2, 4 and 5.

E Proofs of Technical Lemmas

Here we provide the proofs for the technical lemmas. To simplify notation, throughout the proofs we assume $d = 1$ without loss of generality.

E.1 Proof of Lemma 1

It suffices to show that $\hat{\Sigma}_n(\boldsymbol{\kappa}_n) = \tilde{\Sigma}_n(\boldsymbol{\kappa}_n) + o_p(1)$ and that $\tilde{\Sigma}_n(\boldsymbol{\kappa}_n) = \mathbb{E}[\tilde{\Sigma}_n(\boldsymbol{\kappa}_n)|\mathcal{X}_n, \mathcal{W}_n] + o_p(1)$. First,

$$\tilde{\Sigma}_n(\boldsymbol{\kappa}_n) = \frac{1}{n}\sum_{1\leq i\leq G_n} c_{ii,n} + \frac{2}{n}\sum_{1\leq i,j\leq G_n, i<j} c_{ij,n},$$

$$c_{ij,n} = \sum_{s\in\mathcal{T}_i, t\in\mathcal{T}_j}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{v}_{i_1,n}\hat{v}_{j_1,n}M_{i_2s,n}M_{j_2t,n}U_{s,n}U_{t,n}$$

where $\sum_{1\leq i,j\leq G_n}\mathbb{V}[c_{ij,n}|\mathcal{X}_n\mathcal{W}_n]=o_p(n^2)$ because

$$\begin{aligned}
\mathbb{V}[c_{ij,n}|\mathcal{X}_n,\mathcal{W}_n] &\leq (\#\mathcal{T}_{i,n})(\#\mathcal{T}_{j,n})\sum_{s\in\mathcal{T}_{i,n},t\in\mathcal{T}_{j,n}}(\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)}\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{v}_{i_1,n}\hat{v}_{j_1,n}M_{i_2s,n}M_{j_2t,n})^2\mathbb{V}[U_{s,n}U_{t,n}|\mathcal{X}_n\mathcal{W}\\
&\leq \mathcal{C}_{\mathcal{T},n}^2\mathcal{C}_{U,n}\sum_{s\in\mathcal{T}_{i,n},t\in\mathcal{T}_{j,n}}(\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)}\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{v}_{i_1,n}\hat{v}_{j_1,n}M_{i_2s,n}M_{j_2t,n})^2\\
&\leq \mathcal{C}_{\mathcal{T},n}^2\mathcal{C}_{U,n}\sum_{1\leq s,t\leq n}(\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)}\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{v}_{i_1,n}\hat{v}_{j_1,n}M_{i_2s,n}M_{j_2t,n})^2\\
&= \mathcal{C}_{\mathcal{T},n}^2\mathcal{C}_{U,n}\sum_{1\leq s,t\leq n}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)}\sum_{(g_3,i_3,j_3)}\sum_{(g_4,i_4,j_4)}\\
&\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\kappa_{g_3,g_4,i_3,j_3,i_4,j_4,n}\hat{v}_{i_1,n}\hat{v}_{j_1,n}\hat{v}_{i_3,n}\hat{v}_{j_3,n}M_{i_2s,n}M_{j_2t,n}M_{i_4s,n}M_{j_4t,n}\\
&= \mathcal{C}_{\mathcal{T},n}^2\mathcal{C}_{U,n}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)}\sum_{(g_3,i_3,j_3)}\sum_{(g_4,i_4,j_4)}\\
&\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\kappa_{g_3,g_4,i_3,j_3,i_4,j_4,n}\hat{v}_{i_1,n}\hat{v}_{j_1,n}\hat{v}_{i_3,n}\hat{v}_{j_3,n}M_{i_2i_4,n}M_{j_2j_4,n}\\
&\leq \mathcal{C}_{\mathcal{T},n}^2\mathcal{C}_{U,n}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)}\sum_{(g_3,i_3,j_3)}\sum_{(g_4,i_4,j_4)}\\
&|\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}||\kappa_{g_3,g_4,i_3,j_3,i_4,j_4,n}||\hat{v}_{i_1,n}||\hat{v}_{j_1,n}||\hat{v}_{i_3,n}||\hat{v}_{j_3,n}||M_{i_2i_4,n}||M_{j_2j_4,n}|,
\end{aligned}$$

where $\mathcal{C}_{\mathcal{T},n}=\max_{1\leq i\leq G_n}\#(\mathcal{T}_{i,n})$, $\mathcal{C}_{U,n}=1+\max_{1\leq i\leq n}\mathbb{E}[U_{i,n}^4|\mathcal{X}_n,\mathcal{W}_n]$, and

$$\begin{aligned}
&\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)}\sum_{(g_3,i_3,j_3)}\sum_{(g_4,i_4,j_4)}|\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}||\kappa_{g_3,g_4,i_3,j_3,i_4,j_4,n}||\hat{v}_{i_1,n}||\hat{v}_{j_1,n}||\hat{v}_{i_3,n}||\hat{v}_{j_3,n}||M_{i_2i_4,n}||M_{j_2j_4,n}|\\
&\leq (\max_{1\leq i\leq n}|\hat{v}_{i,n}|)^2\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)}\sum_{(g_3,i_3,j_3)}\sum_{(g_4,i_4,j_4)}|\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}||\kappa_{g_3,g_4,i_3,j_3,i_4,j_4,n}||\hat{v}_{i_3,n}||\hat{v}_{j_3,n}||M_{i_2i_4,n}||M_{j_2j_4,n}|\\
&\leq (\max_{1\leq i\leq n}|\hat{v}_{i,n}|)^2\,\|\boldsymbol{\kappa}_n\|_\infty\sum_{(g_2,i_2,j_2)}\sum_{(g_3,i_3,j_3)}\sum_{(g_4,i_4,j_4)}|\kappa_{g_3,g_4,i_3,j_3,i_4,j_4,n}||\hat{v}_{i_3,n}||\hat{v}_{j_3,n}||M_{i_2i_4,n}||M_{j_2j_4,n}|\\
&\leq (\max_{1\leq i\leq n}|\hat{v}_{i,n}|)^2\,\|\boldsymbol{\kappa}_n\|_\infty\,\mathcal{C}_{\mathcal{T},n}\sum_{(g_3,i_3,j_3)}\sum_{(g_4,i_4,j_4)}|\kappa_{g_3,g_4,i_3,j_3,i_4,j_4,n}||\hat{v}_{i_3,n}||\hat{v}_{j_3,n}|\\
&\leq (\max_{1\leq i\leq n}|\hat{v}_{i,n}|)^2\,\|\boldsymbol{\kappa}_n\|_\infty^2\,\mathcal{C}_{\mathcal{T},n}\sum_{(g_3,i_3,j_3)}|\hat{v}_{i_3,n}||\hat{v}_{j_3,n}|\\
&\leq n^2(\frac{\max_{1\leq i\leq n}|\hat{v}_i|}{\sqrt{n}})^2\,\|\boldsymbol{\kappa}_n\|_\infty^2\,\mathcal{C}_{\mathcal{T},n}^2(\frac{1}{n}\sum_{1\leq i\leq n}\hat{v}_{i,n}^2)=o_p(n^2),
\end{aligned}$$

where the third inequality uses

$$\sum_{g_2=1}^{G_n}\sum_{i_2,j_2\in\mathcal{T}_{g_2,n}}|M_{i_2i_4,n}||M_{j_2j_4,n}| \leq \sqrt{(\sum_{g_2=1}^{G_n}\sum_{i_2,j_2)\in\mathcal{T}_{g_2,n}}M^2_{i_2i_4,n})(\sum_{g_2=1}^{G_n}\sum_{(i_2,j_2)\in\mathcal{V}_{g_2,n}}M^2_{j_2j_4,n})}$$

$$\leq \sqrt{(\mathcal{C}_{\mathcal{T},n}\sum_{k=1}^{n}M^2_{ki_4,n})(\mathcal{C}_{\mathcal{T}_n}\sum_{l=1}^{n}M^2_{lj_4,n})}$$

$$= \mathcal{C}_{\mathcal{T},n}\sqrt{M_{i_4i_4,n}M_{j_4j_4,n}}$$

$$\leq \mathcal{C}_{\mathcal{T},n},$$

and the last inequality similarly uses[10]

$$\sum_{g_3=1}^{G}\sum_{i_3,j_3\in\mathcal{T}_{g_2,n}}|\hat{v}_{i_3,n}||\hat{v}_{j_3,n}| \leq \sqrt{(\sum_{g_3=1}^{G}\sum_{i_3,j_3\in\mathcal{T}_{g_3,n}}\hat{v}^2_{i_3,n})(\sum_{g_3=1}^{G}\sum_{i_3,j_3\in\mathcal{T}_{g_3,n}}\hat{v}^2_{j_3,n})}$$

$$\leq \sqrt{(\mathcal{C}_{\mathcal{T},n}\sum_{k=1}^{n}\hat{v}^2_{k,n})(\mathcal{C}_{\mathcal{T},n}\sum_{l=1}^{n}\hat{v}^2_{l,n})}$$

$$= \mathcal{C}_{\mathcal{T},n}(\sum_{i=1}^{n}\hat{v}^2_{i,n}).$$

As a consequence,

$$\mathbb{V}[\frac{1}{n}\sum_{1\leq i\leq G_n}c_{ii,n}|\mathcal{X}_n,\mathcal{W}_n] = \frac{1}{n^2}\sum_{1\leq i\leq G_n}\mathbb{V}[c_{ii,n}|\mathcal{X}_n,\mathcal{W}_n] \leq \frac{1}{n^2}\sum_{1\leq i,j\leq G_n}\mathbb{V}[c_{ij,n}|\mathcal{X}_n,\mathcal{W}_n] = o_p(1)$$

and

$$\mathbb{V}[\frac{1}{n}\sum_{1\leq i,j\leq G_n,i<j}c_{ij,n}|\mathcal{X}_n,\mathcal{W}_n] = \frac{1}{n^2}\sum_{1\leq i,j\leq G_n,i<j}\mathbb{V}[c_{ij,n}|\mathcal{X}_n,\mathcal{W}_n] \leq \frac{1}{n^2}\sum_{1\leq i,j\leq G_n}\mathbb{V}[c_{ij,n}|\mathcal{X}_n,\mathcal{W}_n] = o_p(1).$$

[10]We also make use of the bound $\frac{1}{n}\sum_{i=1}^{n}\hat{v}_i^2 = O_p(1)$, as shown in Lemma SA-1 of CJN (2017a, Supplemental Appendix).

In particular, $\tilde{\Sigma}_n(\boldsymbol{\kappa}_n) = \mathbb{E}[\tilde{\Sigma}_n(\boldsymbol{\kappa}_n)|\mathcal{X}_n, \mathcal{W}_n] + o_p(1)$, where

$$
\begin{aligned}
|\mathbb{E}[\tilde{\Sigma}_n(\boldsymbol{\kappa}_n)|\mathcal{X}_n, \mathcal{W}_n]| &\leq \frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} \sum_{(g_3,i_3,j_3)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}||\hat{v}_{i_1,n}||\hat{v}_{j_1,n}||M_{i_2j_3,n}||M_{j_2i_3,n}||\mathbb{E}[U_{i_3,n}U_{j_3,n}|\mathcal{X}_n, \mathcal{W}_n]| \\
&\leq \mathcal{C}_{U,n}\frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} \sum_{(g_3,i_3,j_3)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}||\hat{v}_{i_1,n}||\hat{v}_{j_1,n}||M_{i_2j_3,n}||M_{j_2i_3,n}| \\
&\leq \mathcal{C}_{U,n}\mathcal{C}_{\mathcal{T},n}\frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}||\hat{v}_{i_1,n}||\hat{v}_{j_1,n}| \\
&\leq \mathcal{C}_{U,n}\mathcal{C}_{\mathcal{T},n} \left\|\kappa_n\right\|_\infty \frac{1}{n} \sum_{(g_1,i_1,j_1)} |\hat{v}_{i_1,n}||\hat{v}_{j_1,n}| \\
&\leq \mathcal{C}_{U,n}\mathcal{C}^2_{\mathcal{T},n} \left\|\kappa_n\right\|_\infty (\frac{1}{n} \sum_{1\leq i\leq n} \hat{v}^2_{i,n}) = O_p(1)
\end{aligned}
$$

We have therefore established that $\tilde{\Sigma}_n(\boldsymbol{\kappa}_n) = \mathbb{E}[\tilde{\Sigma}_n(\boldsymbol{\kappa}_n)|\mathcal{X}_n, \mathcal{W}_n] + o_p(1)$. It remains to show that $\hat{\Sigma}_n(\boldsymbol{\kappa}_n) = \tilde{\Sigma}_n(\boldsymbol{\kappa}_n) + o_p(1)$.
By using that $\hat{u}_{i,n} - \tilde{U}_{i,n} = \tilde{R}_{i,n} - \hat{v}_{i,n}(\hat{\beta}_n - \beta)$, where $\tilde{R}_{i,n} = \sum_{j=1}^n M_{ij,n}R_{j,n}$, we obtain

$$
\begin{aligned}
&\hat{\Sigma}_n(\boldsymbol{\kappa}_n) - \tilde{\Sigma}_n(\boldsymbol{\kappa}_n) = \frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{v}_{i_1,n}\hat{v}_{j_1,n}(\hat{u}_{i_2,n}\hat{u}_{j_2,n} - \tilde{U}_{i_2,n}\tilde{U}_{j_2,n}) \\
&= \frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} \\
&\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{v}_{i_1,n}\hat{v}_{j_1,n}[(\tilde{R}_{i_2,n} - \hat{v}_{i_2,n}(\hat{\beta}_n - \beta) + \tilde{U}_{i_2,n})(\tilde{R}_{j_2,n} - \hat{v}_{j_2,n}(\hat{\beta}_n - \beta) + \tilde{U}_{j_2,n}) - \tilde{U}_{i_2,n}\tilde{U}_{j_2,n}].
\end{aligned}
$$

By the Cauchy-Schwarz inequality, it suffices to show that

$$
\begin{aligned}
&\frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{v}^2_{i_1,n}(\tilde{R}_{i_2,n} - \hat{v}_{i_2,n}(\hat{\beta}_n - \beta))^2 = o_p(1), \\
&\frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{v}^2_{i_1,n}\tilde{U}^2_{j_2,n} = O_p(1).
\end{aligned}
$$

The latter can be straightforwardly shown by means of the arguments previously used to show $\tilde{\Sigma}_n(\boldsymbol{\kappa}_n) = O_p(1)$. For the former, since $\hat{v}_{j,n} = \tilde{V}_{j,n} + \tilde{Q}_{j,n}$, where $\tilde{Q}_{i,n} = \sum_{j=1}^n M_{ij,n}Q_{j,n}$, and

$\tilde{R}_{i,n} = \tilde{r}_{i,n} + (\tilde{R}_{i,n} - \tilde{r}_{i,n})$, where $\tilde{r}_{i,n} = \sum_{j=1}^{n} M_{ij,n} r_{j,n}$ it suffices to show that

$$\frac{1}{n}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| \tilde{Q}^2_{i_1,n}\tilde{R}^2_{i_2,n} = o_p(1),$$
$$\frac{1}{n}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| \tilde{V}^2_{i_1,n}\tilde{r}^2_{i_2,n} = o_p(1),$$
$$\frac{1}{n}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| \tilde{V}^2_{i_1,n}|\tilde{R}_{i_2,n} - \tilde{r}_{i_2,n}|^2 = o_p(1),$$
$$(\hat{\beta}_n - \beta)^2 \frac{1}{n}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| \hat{v}^2_{i_1,n}\hat{v}^2_{i_2,n} = o_p(1).$$

First, $n^{-1}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| \tilde{V}^2_{i_1,n}\tilde{r}^2_{i_2,n} = o_p(1)$ because

$$\begin{aligned}
\mathbb{E}[\frac{1}{n}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| \tilde{V}^2_{i_1,n}\tilde{r}^2_{i_2,n}|\mathcal{W}_n] &= \frac{1}{n}\sum_{(g_2,i_2,j_2)} \tilde{r}^2_{i_2,n} \sum_{(g_1,i_1,j_1)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| \mathbb{E}[\tilde{V}^2_{i_1,n}|\mathcal{W}_n] \\
&\leq \mathcal{C}_{V,n}\mathcal{C}_{\mathcal{T},n}\frac{1}{n}\sum_{(g_2,i_2,j_2)} \tilde{r}^2_{i_2,n} \sum_{(g_1,i_1,j_1)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| \\
&\leq \mathcal{C}_{V,n}\mathcal{C}_{\mathcal{T},n} \left\|\boldsymbol{\kappa}_n\right\|_\infty (\frac{1}{n}\sum_{(g_2,i_2,j_2)} \tilde{r}^2_{i_2,n}) \\
&\leq \mathcal{C}_{V,n}\mathcal{C}^2_{\mathcal{T},n} \left\|\boldsymbol{\kappa}_n\right\|_\infty (\frac{1}{n}\sum_{i=1}^{n} \tilde{r}^2_{i,n}) = O_p(\rho_n) = o_p(1),
\end{aligned}$$

where the first inequality uses the fact that $\mathbb{E}[\tilde{V}_{i,n}|\mathcal{W}_n] \leq \mathcal{C}_{\mathcal{T},n}\mathcal{C}_{V,n}$, with $\mathcal{C}_{V,n} = 1 + \max_{1\leq i\leq n}, \mathbb{E}[\left\|V_{i,n}\right\|^4 |\mathcal{W}_n]$ as shown in CJN (2018a, Supplemental Appendix).

Next,

$$\begin{aligned}
\frac{1}{n}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| \tilde{V}^2_{i_1,n}|\tilde{R}_{i_2,n} - \tilde{r}_{i_2,n}|^2 &\leq n\left\|\boldsymbol{\kappa}_n\right\|_\infty (\frac{1}{n}\sum_{(g_1,i_1,j_1)} \tilde{V}^2_{i_1,n})(\frac{1}{n}\sum_{(g_2,i_2,j_2)} |\tilde{R}_{i_2,n} - \tilde{r}_{i_2,n}|^2) \\
&\leq n\left\|\boldsymbol{\kappa}_n\right\|_\infty \mathcal{C}^2_{\mathcal{T},n}(\frac{1}{n}\sum_{i=1}^{n} \tilde{V}^2_{i,n})(\frac{1}{n}\sum_{i=1}^{n}|\tilde{R}_{i,n} - \tilde{r}_{i,n}|^2) \\
&= O_p[n(\varrho_n - \rho_n)] = o_p(1)
\end{aligned}$$

and

$$\begin{aligned}
\frac{1}{n}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| \tilde{Q}^2_{i_1,n}\tilde{R}^2_{i_2,n} &\leq n\left\|\boldsymbol{\kappa}_n\right\|_\infty (\frac{1}{n}\sum_{(g_1,i_1,j_1)} \tilde{Q}^2_{i_1,n})(\frac{1}{n}\sum_{(g_2,i_2,j_2)} \tilde{R}^2_{i_2,n}) \\
&\leq n\left\|\boldsymbol{\kappa}_n\right\|_\infty \mathcal{C}^2_{\mathcal{T},n}(\frac{1}{n}\sum_{i=1}^{n} \tilde{Q}^2_{i,n})(\frac{1}{n}\sum_{i=1}^{n} \tilde{R}^2_{i,n}) \\
&= O_p(n\chi_n\varrho_n)] = o_p(1)
\end{aligned}$$

Finally,

$$(\hat{\beta}_n - \beta)^2 \frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| \hat{v}_{i_1,n}^2 \hat{v}_{i_2,n}^2 = o_p(1)$$

because $\sqrt{n}(\hat{\beta}_n - \beta) = O_p(1)$ and

$$\begin{aligned}
\frac{1}{n^2} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| \hat{v}_{i_1,n}^2 \hat{v}_{i_2,n}^2 &\le (\max_{1 \le i \le n} |\hat{v}_{i,n}|)^2 \frac{1}{n^2} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| \hat{v}_{i_2,n}^2 \\
&\le (\frac{\max_{1 \le i \le n} |\hat{v}_{i,n}|}{\sqrt{n}})^2 \, \|\boldsymbol{\kappa}_n\|_\infty \, (\frac{1}{n} \sum_{(g_2,i_2,j_2)} \hat{v}_{i_2,n}^2) \\
&\le (\frac{\max_{1 \le i \le n} |\hat{v}_{i,n}|}{\sqrt{n}})^2 \, \|\boldsymbol{\kappa}_n\|_\infty \, \mathcal{C}_{\mathcal{T},n} (\frac{1}{n} \sum_{i=1}^{n} \hat{v}_{i,n}^2) \\
&= o_p(1),
\end{aligned}$$

which concludes the proof.

E.2 Proof of Lemma 2

Let us define $d_{i_1 j_1, i_3 j_3, n} = \sum_{(g_2,i_2,j_2)} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n} M_{i_3,j_2,n} M_{j_3,i_2,n} - \mathbb{1}\{(i_1, j_1) = (i_3, j_3)\}$. We hence have

$$\mathbb{E}[\tilde{\Sigma}_n(\boldsymbol{\kappa}_n) | \mathcal{X}_n, \mathcal{W}_n] - \Sigma_n = \frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_3,i_3,j_3)} d_{i_1 j_1, i_3 j_3, n} \hat{v}_{i_1,n} \hat{v}_{j_1,n} \mathbb{E}[U_{i_3,n} U_{j_3,n} | \mathcal{X}_n, \mathcal{W}_n],$$

so if $\max_{(g_1,i_1,j_1)} \sum_{(g_3,i_3,j_3)} |d_{i_1 j_1, i_3 j_3, n}| = o_p(1)$, then

$$\begin{aligned}
|\mathbb{E}[\tilde{\Sigma}_n(\boldsymbol{\kappa}_n) | \mathcal{X}_n, \mathcal{W}_n] - \Sigma_n| &\le \frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_3,i_3,j_3)} |d_{i_1 j_1, i_3 j_3, n}| |\hat{v}_{i_1,n}| |\hat{v}_{j_1,n}| |\mathbb{E}[U_{i_3,n} U_{j_3,n} | \mathcal{X}_n, \mathcal{W}_n]| \\
&\le \mathcal{C}_{U,n} \frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_3,i_3,j_3)} |d_{i_1 j_1, i_3 j_3, n}| |\hat{v}_{i_1,n}| |\hat{v}_{j_1,n}| \\
&\le \mathcal{C}_{U,n} (\frac{1}{n} \sum_{(g_1,i_1,j_1)} |\hat{v}_{i_1,n}| |\hat{v}_{j_1,n}|) (\max_{(g_1,i_1,j_1)} \sum_{(g_3,i_3,j_3)} |d_{i_1 j_1, i_3 j_3, n}|) \\
&\le \mathcal{C}_{U,n} \mathcal{C}_{\mathcal{T},n} (\frac{1}{n} \sum_{i}^{n} \hat{v}_{i,n}^2) (\max_{(g_1,i_1,j_1)} \sum_{(g_3,i_3,j_3)} |d_{i_1 j_1, i_3 j_3, n}|) = o_p(1).
\end{aligned}$$

E.3 Proof of Lemma 3

If $\lambda_{\min}(\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n) > 0$, then

$$\Big|\sum_{(g_2,i_2,j_2)} \kappa^{\mathtt{CR}}_{g_1,g_2,i_1,j_1,i_2,j_2,n} M_{i_1j_2,n}M_{j_1i_2,n} - 1\Big| + \sum_{(g_3,i_3,j_3)\neq(g_1,i_1,j_1)} \Big|\sum_{(g_2,i_2,j_2)} \kappa^{\mathtt{CR}}_{g_1,g_2,i_1,j_1,i_2,j_2} M_{i_3j_2,n}M_{j_3i_2,n}\Big| = 0$$

which combined with Lemma 2 gives $\mathbb{E}[\tilde{\boldsymbol{\Sigma}}_n(\boldsymbol{\kappa}^{\mathtt{CR}}_n)|\mathcal{X}_n,\mathcal{W}_n] = \boldsymbol{\Sigma}_n + o_p(1)$.

E.4 Proof of Lemma 4

Recall that for White's estimator we have

$$\begin{aligned}\sum_{g_2,i_2,j_2}|d_{i_1,j_1,i_2,j_2,n}| &= |M_{i_1i_1,n}M_{j_1j_1,n} - 1| + \sum_{(g_2,i_2,j_2)\neq(g_1,i_1,j_1)} |M_{i_1j_2,n}||M_{j_1i_2,n}| \\ &= (1 - M_{i_1i_1,n}M_{j_1j_1,n}) + M_{i_1i_1,n}\Big(\sum_{\substack{i_2\in g_1\\ i_2\neq j_1}}|M_{j_1i_2,n}|\Big) + M_{j_1j_1,n}\Big(\sum_{\substack{j_2\in g_1\\ j_2\neq i_1}}|M_{i_1j_2,n}|\Big) + \sum_{\substack{(g_2,i_2,j_2)\\ i_2\neq j_1, j_2\neq i_1}} |M_{i_1,j_2,n}||M_{j_1,i_2,n}|,\end{aligned}$$

Defining $\mathcal{M}_n = 1 - \min_{1\leq i\leq n} M_{ii,n}$, we have that $\mathcal{M}_n = O_p(\frac{K_n}{n})$ (see Cattaneo et al., 2017) and

$$\max_{\substack{i,j\\ i\neq j}}|M_{ij,n}| \leq \max_{\substack{i,j\\ i\neq j}} \frac{1}{n}\sum_{l=1}^{K_n}|w^*_{il,n}||w^*_{jl,n}| \leq \max_i \frac{1}{n}\sum_{l=1}^{K_n} w^{*2}_{il,n} = O_p(\frac{K_n}{n}).$$

As a result, we have

$$\begin{aligned}\max_{(g_1,i_1,j_1)}\sum_{g_2,i_2,j_2}|d_{i_1,j_1,i_2,j_2,n}| &\leq 2\mathcal{M}_n + 2(\mathcal{C}_{\mathcal{T},n} - 1)O_p(\frac{K_n}{n}) + \mathcal{C}^2_{\mathcal{T},n}G_nO_p(\frac{K_n^2}{n^2}) \\ &\leq O_p(\frac{K_n}{n}) + 2(\mathcal{C}_{\mathcal{T},n} - 1)O_p(\frac{K_n}{n}) + \mathcal{C}^2_{\mathcal{T},n}O(n)O_p(\frac{K_n^2}{n^2}) = O_p(\frac{K_n^2}{n}),\end{aligned}$$

which combined with Lemma 2 gives $\mathbb{E}[\tilde{\boldsymbol{\Sigma}}_n(\boldsymbol{\kappa}_n)|\mathcal{X}_n,\mathcal{W}_n] = \boldsymbol{\Sigma}_n + o_p(1)$.

E.5 Proof of Lemma 5

Under homoskedasticity one has

$$\begin{aligned}\mathbb{E}[\tilde{\Sigma}_n(\mathbf{I}_{L_n})|\mathcal{X}_n,\mathcal{W}_n] &= \frac{\sigma_n^2}{n}\sum_{(g_1,i_1,j_1)}\sum_{k=1}^{n}\hat{v}_{i_1,n}\hat{v}_{j_1,n}M_{i_1k,n}M_{j_1k,n} \\ &= \frac{\sigma_n^2}{n}\sum_{(g_1,i_1,j_1)}\hat{v}_{i_1,n}\hat{v}_{j_1,n}M_{i_1j_1,n},\end{aligned}$$

and

$$\Sigma_n = \frac{\sigma_n^2}{n}\sum_{i=1}^n \hat{v}_{i,n}^2.$$

As a result, we have

$$\begin{aligned}
|\mathbb{E}[\tilde{\Sigma}_n(\mathbf{I}_{L_n})|\mathcal{X}_n,\mathcal{W}_n] - \Sigma_n| &\leq \frac{\sigma_n^2}{n}\sum_{i=1}^n \hat{v}_{i,n}^2|M_{ii.n}-1| + \frac{\sigma_n^2}{n}\sum_{\substack{(g_1,i_1,j_1)\\ i_1\neq j_1}} |\hat{v}_{i_1,n}||\hat{v}_{j_1,n}||M_{i_1j_1,n}| \\
&\leq \mathcal{C}_{U,n}\mathcal{M}_n(\frac{1}{n}\sum_{i=1}^n \hat{v}_{i,n}^2) + \mathcal{C}_{U,n}\mathcal{C}_{\mathcal{T},n}(\max_{\substack{i,j\\ i\neq j}}|M_{ij,n}|)(\frac{1}{n}\sum_{i=1}^n \hat{v}_{i,n}^2) \\
&\leq \mathcal{C}_{U,n}O_p(\frac{K_n}{n})(\frac{1}{n}\sum_{i=1}^n \hat{v}_{i,n}^2) + \mathcal{C}_{U,n}\mathcal{C}_{\mathcal{T},n}O_p(\frac{K_n}{n})(\frac{1}{n}\sum_{i=1}^n \hat{v}_{i,n}^2) = O_p(\frac{K_n}{n}),
\end{aligned}$$

and therefore $\mathbb{E}[\tilde{\Sigma}_n(\mathbf{I}_{L_n})|\mathcal{X}_n,\mathcal{W}_n] = \Sigma_n + o_p(1)$.

F Extension to within-cluster restrictions

In this section, we present an extension of the class of estimators studied in this paper that allows to impose zero-restrictions on the variance-covariance matrix of the errors within clusters.

Define the sets

$$\begin{aligned}
\mathcal{V}_{g,n} &= \{(i,j)\in\mathcal{T}_{g,n}\times\mathcal{T}_{g,n} : \mathbb{E}[U_{i,n}U_{j,n}|\mathcal{X}_n,\mathcal{W}_n]\neq 0\}, \\
\mathcal{R}_{g,i,n} &= \{j\in\mathcal{T}_{g,n} : \mathbb{E}[U_{i,n}U_{j,n}|\mathcal{X}_n,\mathcal{W}_n]\neq 0\},
\end{aligned}$$

and let L_n be the number of non-zero elements contained in $\mathbf{\Omega}_{U,n} = \mathbb{E}[\mathbf{U}_n\mathbf{U}_n'|\mathcal{X}_n,\mathcal{W}_n]$. The generalized version of our proposed class of cluster-robust variance estimators reads:

$$\hat{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n) = \frac{1}{n}\sum_{g_1=1}^{G_n}\sum_{g_2=1}^{G_n}\sum_{(i_1,j_1)\in\mathcal{V}_{g_1,n}}\sum_{(i_2,j_2)\in\mathcal{V}_{g_2,n}} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{\mathbf{v}}_{i_1,n}\hat{\mathbf{v}}_{j_1,n}'\hat{u}_{i_2,n}\hat{u}_{j_2,n},$$

where $\kappa_{g_1,g_2,i_1,j_1,i_2.j_2,n}$ corresponds to the $(h(g_1,i_1,j_1),h(g_2,i_2,j_2))$ entry of the $L_n\times L_n$ symmetric matrix $\boldsymbol{\kappa}_n$, where $h(g,i,j) = [\sum_{k=0}^{(g-1)}(\#\mathcal{V}_{k,n}) + \sum_{k=0}^{i-1}(\#\mathcal{R}_{g,k,n}) + \overline{j(i)_{g,n}}]$ with $\overline{j(i)_{g,n}} = \#\{k\in\mathcal{T}_{g,n} : \mathbb{E}[U_{i,n}U_{j,n}|\mathcal{X}_n,\mathcal{W}_n]\neq 0 \quad \text{and} \quad k\leq j\}$ and we adopt the convention that $\#\mathcal{V}_{0,n}=0$ and $\#\mathcal{R}_{g,0,n}=0 \quad \forall g$.

A consistent estimator under Assumptions 1*-3* is then defined as $\hat{\mathbf{\Sigma}}(\boldsymbol{\kappa}_n^{\mathtt{CR}})$, where $\boldsymbol{\kappa}_n^{\mathtt{CR}} = (\mathbf{S}_n'(\mathbf{M}_n\otimes\mathbf{M}_n)\mathbf{S}_n)^{-1}$.