Cross-Validation for Clustered Regression

Bruce E. Hansen*
University of Wisconsin†

February, 2025

Abstract

This paper explores cross-validation regression model selection under one-way clustered dependence. Clustered dependence is ubiquitous in current econometric applications, as evidence by the widespread use of cluster-robust variance estimation and standard errors. Cross-validation methods are also routinely used to compare and select between estimated models. There is no current theory, however, investigating the interaction of clustered dependence and model selection. We show that conventional cross-validation methods are inappropriate for model selection, as they do not account for within-cluster correlation. In contrast, we show that delete-cluster cross-validation is asymptotically optimal for model selection when evaluated by a regression mean-squared error criterion. This is because the delete-cluster criterion mimics the dependence structure of the observations. In contrast, we show that conventional cross-validation effectively impose a "too small" parameterization penalty similar to the "too small" standard errors obtained when clustered dependence is ignored in variance estimation. In a simulation experiment we investigate the performance of conventional and delete-cluster cross-valiation methods, and find that delete-cluster cross-validation is much preferred when the within-cluster dependence is high. We illustrate the method in a simple empirical application. The delete-cluster cross-validation criterion is simple to calculate, and is displayed by default by the author's `jregress` clustered regression R and Stata packages.

*Research support from the Phipps Chair is gratefully acknowledged.
†Department of Economics, 1180 Observatory Drive, University of Wisconsin, Madison WI 53706.

1 Introduction

Clustered samples are widespread in current applied econometric practice. Despite this dominance, there is no existing model-selection literature allowing for clustered dependence. This paper provides a formal asymptotic theory for model selection by cross-validation.

We evaluate model-selection performance by the fitted regression mean-squared error. This is the conventional criterion for the evaluation of regression model selection in the theoretical literature, including Li (1987) and Andrews (1991). Our assumptions are similar to those this literature, except that rather than assuming the observations are mutually independent, we assume that the observations are grouped in known clusters, that the observations are mutually independent across clusters, but dependence within clusters is unrestricted.

We show that the presence of within-cluster dependence can dramatically impact optimal model selection. The mechanism is the same as for standard error construction: clustered dependence affects estimation variance, and thereby affects the optimal fitted model. This arises even under a small degree of within-cluster correlation when the number of observations per cluster is large. This is similar to Moulton's (1986) result on the impact of clustered dependence on the variance of regression estimates.

Cluster-robust variance estimation was introduced to econometrics by Liang and Zeger (1986) and Arellano (1987), and became the dominant variance estimation method in empirical practice following the influential work of Bertrand, Duflo, and Mullainathan (2004). The distributional theory justifying cluster-robust inference has been been developed by many authors, including Moulton (1986, 1990), Wooldridge (2003), C. Hansen (2007), Cameron, Gelbach, and Miller (2008), Bester, Conley, and C. Hansen (2011), Conley and Taber (2011), Ibragimov and Müller (2016), Imbens and Kolesár (2016), Djogbenou, MacKinnon, and Nielsen (2019), Ferman and Pinto (2019), Hagemann (2019), Hansen and Lee (2019), MacKinnon and Webb (2020), Canay, Santos, and Shaikh (2021), and Hansen (2024).

Delete-one cross-validation for selection of regression models was introduced by Allen (1974), Stone (1974), Geisser (1974), and Wahba and Wold (1975). Its optimality for homoskedastic regression was demonstrated by Li (1987) and for heteroskedastic regression by Andrews (1991). A review of theory is presented in Shao (1997). A related theory for model averaging was developed by Hansen (2007) and Hansen and Racine (2012). For further theoretical results, see Hansen (2014, 2015).

Cross-validation is closely related to the jackknife. The jackknife estimator of variance was introduced by Tukey (1958) and was developed in the monographs of Efron (1982) and Shao and Tu (1995). Extensions to clustered samples were made by Cochran (1977), Rust and Rao (1996), Bell and McCaffrey (2002), MacKinnon, Nielsen, and Webb (2023), and Hansen (2024).

Our proposed delete-cluster cross-valiation criterion (CCV) is simple to calculate, especially when the number of clusters is small to moderate, which is typical in economic applications. Its calculation is also a simple by-product of delete-cluster jackknife variance estimation. The CCV criterion is displayed by default by the author's `jregress` Stata and R packages for regression with delete-cluster variance estimation.

This paper is organized as follows. Section 2 presents the clustered regression model. Section 3 presents cross-validation model selection methods. Section 4 presents a simple empirical illustration.

Section 5 presents a definition of overall model fit based on mean-squared error (MSE). Section 6 presents the main theory, demonstrating that model selection by delete-cluster cross-validation is asymptotically equivalent to the infeasible optimal regression model with respect to the MSE measure. Section 7 examines conventional delete-one cross-validation (CV), and shows that model selection by conventional CV is asymptotically not optimal. Section 8 extends the results to handle the presence of cluster fixed effects. Section 9 presents a simulation experiment. Section 10 presents two probability inequalities, generalizations of the those of Whittle (1960) to the cluster-dependent setting. The proofs of the main theorems follow.

2 Clustered Regression

The setting is a clustered sample, with observations grouped into G unbalanced independent clusters. We index the observations by ig, denoting the ith observation in the gth cluster, with $g = 1, ..., G$ and $i = 1, ..., n_g$. The total number of observations is $n = \sum_{g=1}^{G} n_g$.

A researcher is interested in estimating the conditional mean of a scalar dependent variable Y_{ig} given a large set of potential regressors X_{ig}. This structure can be written as

$$\begin{aligned} Y_{ig} &= \mu_{ig} + e_{ig} \\ \mu_{ig} &= \mathbb{E}\left[Y_{ig} \mid X_{ig}\right] \\ \mathbb{E}\left[e_{ig} \mid X_{ig}\right] &= 0. \end{aligned} \tag{1}$$

By construction, e_{ig} is the true regression error.

We can also write the regression structure using cluster-level notation. This is

$$\boldsymbol{Y}_g = \boldsymbol{\mu}_g + \boldsymbol{e}_g \tag{2}$$

$$\boldsymbol{\mu}_g = \mathbb{E}\left[\boldsymbol{Y}_g \mid \boldsymbol{X}_g\right] \tag{3}$$

$$\mathbb{E}\left[\boldsymbol{e}_g \mid \boldsymbol{X}_g\right] = 0. \tag{4}$$

where $\boldsymbol{Y}_g$, $\boldsymbol{\mu}_g$, and $\boldsymbol{e}_g$ are $n_g \times 1$. The error $\boldsymbol{e}_g$ has the $n_g \times n_g$ cluster-level covariance matrix

$$\mathbb{E}\left[\boldsymbol{e}_g \boldsymbol{e}_g' \mid \boldsymbol{X}_g\right] = \boldsymbol{\Sigma}_g. \tag{5}$$

We treat $\boldsymbol{\Sigma}_g$ as unknown and unstructured. This allows for unconditional and conditional heteroskedasticity, as well as arbitrary within-cluster correlation. The maintained assumption is that the cluster-level observations are mutually independent across clusters.

The researcher considers a set of M regression models, which we index as $m = 1, ..., M$. For the mth model, the researcher constructs[1] a $k(m) \times 1$ vector $X_{ig}(m)$ from the set X_{ig} and postulates the linear

[1] This can include individual variables, transformations, interactions, and series transformations.

regression

$$Y_{ig} = X_{ig}(m)'\beta(m) + u_{ig}(m) \tag{6}$$

where $\beta(m)$ is a $k(m) \times 1$ coefficient vector and $u_{ig}(m)$ is a projection error. We index the regressors, coefficients, and projection errors by the model m as they are model-dependent. In general, the model (6) is not equal to the true regression (1) but rather is a projection approximation.

The model (6) written at the level of the cluster is

$$\boldsymbol{Y}_g = \boldsymbol{X}_g(m)\beta(m) + \boldsymbol{u}_g(m) \tag{7}$$

where $\boldsymbol{X}_g(m)$ is $n_g \times k(m)$.

The coefficient $\beta(m)$ is estimated by least squares[2]

$$\begin{aligned}\widehat{\beta}(m) &= \left(\sum_{g=1}^{G}\sum_{i=1}^{n_g} X_{ig}(m)X_{ig}(m)'\right)^{-1}\left(\sum_{g=1}^{G}\sum_{i=1}^{n_g} X_{ig}(m)Y_{ig}\right) \\ &= \left(\sum_{g=1}^{G} \boldsymbol{X}_g(m)'\boldsymbol{X}_g(m)\right)^{-1}\left(\sum_{g=1}^{G} \boldsymbol{X}_g(m)'\boldsymbol{Y}_g\right).\end{aligned}$$

The least-squares fitted conditional mean written at the level of the cluster is

$$\widehat{\boldsymbol{\mu}}_g(m) = \boldsymbol{X}_g(m)\widehat{\beta}(m).$$

3 Cross-Validation

A conventional criterion for model selection is delete-one cross-validation (CV). For each individual observation ig, the delete-one least squares estimator is

$$\bar{\beta}_{-ig}(m) = \left(\sum_{g=1}^{G}\sum_{j\neq i} X_{jg}(m)X_{jg}(m)'\right)^{-1}\left(\sum_{g=1}^{G}\sum_{j\neq i} X_{jg}(m)Y_{jg}\right).$$

The delete-one prediction errors are

$$\bar{u}_{ig}(m) = Y_{ig} - X_{ig}(m)'\bar{\beta}_{-ig}(m).$$

The delete-one cross-validation criterion is

$$\mathrm{CV}(m) = \sum_{g=1}^{G}\sum_{i=1}^{n_g} \bar{u}_{ig}(m)^2. \tag{8}$$

An alternative criterion is delete-cluster cross-validation (CCV). For each cluster g, the delete-cluster

[2]The ideas apply to other estimators, but our theoretical treatment is confined to least squares estimation.

least squares estimator is

$$\widetilde{\beta}_{-g}(m) = \left(\sum_{j\neq g} \boldsymbol{X}_j(m)'\boldsymbol{X}_j(m)\right)^{-1}\left(\sum_{j\neq g} \boldsymbol{X}_j(m)'\boldsymbol{Y}_j\right). \tag{9}$$

The delete-cluster prediction errors are

$$\widetilde{u}_{ig}(m) = Y_{ig} - X_{ig}(m)'\bar{\beta}_{-g}(m).$$

These equal

$$\widetilde{\boldsymbol{u}}_g(m) = \boldsymbol{Y}_g - \boldsymbol{X}_g(m)\widetilde{\beta}_{-g}(m) \tag{10}$$

when grouped by cluster. The delete-cluster cross-validation (CCV) criterion is

$$\mathrm{CCV}(m) = \sum_{g=1}^{G}\sum_{i=1}^{n_g} \widetilde{u}_{ig}(m)^2. \tag{11}$$

The cross-validation criteria (8) and (11) may be used to compare and select models. The CV-selected model is the one with the smallest value of $\mathrm{CV}(m)$. This is

$$\widehat{m}_{\mathrm{CV}} = \operatorname*{argmin}_{1\leq m\leq M} \mathrm{CV}(m).$$

Similarly, the CCV-selected model is the one with the smallest value of $\mathrm{CCV}(m)$, which is

$$\widehat{m}_{\mathrm{CCV}} = \operatorname*{argmin}_{1\leq m\leq M} \mathrm{CCV}(m).$$

Calculation of the CCV criterion (11) requires computation of G least-squares regressions. This is roughly a G-fold computational increase over a single regression estimation, though some computational reductions can be achieved by storage of the within-cluster cross-product matrices. This is distinct from the delete-one criterion (8), which does not require explicit re-estimations. In most economic applications the number of clusters G is small to moderate, so this computational cost is mild.

Furthermore, when covariance matrix estimation is performed by the cluster jackknife (as recommended by MacKinnon, Nielsen, and Webb (2023abc) and Hansen (2024)) the CCV criterion is a simple by-product of jackknife calculation, so CCV calculation does not require any additional computational burden. The criterion (11) is displayed (by default) by the Stata and R `jregress` jackknife regression command written by the author.

4 Empirical Illustration

We illustrate the use of the CCV criterion (8) by an application to one of the regression models from Duflo, Dupas, and Kramer (2011). Our sample, taken from the authors' replication file, is 5269 Kenyan

first grade students[3] at 111 schools.

These authors investigate the impact of *tracking* (splitting students based on beginning-of-year achievement) on end-of-year *testscores*, conditional on a set of controls. The observations are clustered by school.

An ideal regression model includes a sufficient set of controls to minimize omitted variables bias, but not more controls than necessary so that estimation error is minimized. We consider four control variables: *gender*, *age*, contract *teacher*, and *percentile* (initial student achievement reported as a percentile). Two of these variables (*gender* and *teacher*) are binary, the third (*age*) is discrete with a large number of distinct values, and the fourth (*percentile*) is continuous. Regression models which control for these variables can take many forms, and it is not a priori obvious which is correct. Plausible models include linear specifications, nonlinear, nonparametric, and with and without interactions.

Table 1 reports five nested regression models which vary the specifications. All models include *tracking*. Regression 1 includes no other variables. Regression 2 includes the four control variables linearly. Regression 3 includes a fourth-order polynomial in *percentile*. Regression 4 replaces the linear control *age* with dummy indicators for each year of *age*. Regression 5 includes six interactions between the linear variables.

To compare the regressions, we report delete-one CV and the delete-cluster cross-validation CCV for each estimated model. Smaller values indicate better fit. Focusing on CCV, we can see that Regression 4 (the model with flexible specifications for *percentile* and *age*, but no interactions) has the best fit among the five regressions. If the CCV criterion is used as a guide for model selection among the reported models, then this regression is selected, resulting in an estimated coefficient for tracking of 0.171 with a standard error of 0.078.

In contrast, if conventional CV is used for model selection, then the largest estimated model (Regression 5) is selected.

5 MSE

A theoretical measure of overall model fit is the weighted mean (expected) squared error. Let $\boldsymbol{X}$ denote the full set of regressors. We define

$$\mathrm{MSE}(m) = \sum_{g=1}^{G} \mathbb{E}\left[\left(\widehat{\boldsymbol{\mu}}_g(m) - \boldsymbol{\mu}_g\right)' \left(\widehat{\boldsymbol{\mu}}_g(m) - \boldsymbol{\mu}_g\right) \mid \boldsymbol{X}\right]. \tag{12}$$

This is the expected sum of squared deviations between the least-squares fitted values and the true conditional mean.

Define the covariance matrix $\boldsymbol{\Sigma} = \mathrm{diag}\{\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_G\}$ where $\boldsymbol{\Sigma}_g$ is defined in (5), the $n \times 1$ conditional mean $\boldsymbol{\mu}$ by stacking the cluster-level means $\boldsymbol{\mu}_g$, the full-sample regressor matrices $\boldsymbol{X}(m)$ by stacking the

[3]The original sample has 7022 students at 121 schools. We reduced the sample to 5269 students to eliminate observations with missing relevant variables.

Table 1: Regression Models

	(1)	(2)	(3)	(4)	(5)
Tracking	0.149	0.174	0.174	0.171	0.173
Jackknife s.e.	(0.078)	(0.078)	(0.078)	(0.078)	(0.078)
Linear controls	No	Yes	Yes	Yes	Yes
Quartic in percentile	No	No	Yes	Yes	Yes
Age (discrete)	No	No	No	Yes	Yes
Interactions	No	No	No	No	Yes
CV	5254	3969	3936	3916	3914*
CCV	5282	4005	3973	3953*	3961

Notes: The sample includes 60 tracking and 61 nontracking schools. The dependent variables are normalized test scores, with mean 0 and standard deviation 1 in the nontracking schools. Jackknife standard errors clustered at the school level are presented in parentheses. Basic controls are: age (linear), gender, being assigned to the contract teacher, and initial attainment percentile. Age (discrete) is a set of 13 dummies indicating student age from 6 through 19. Interactions are the products between the basic controls.

cluster-level regressor matrices $\boldsymbol{X}_g(m)$, and the projection matrix

$$\boldsymbol{P}(m) = \boldsymbol{X}(m)\left(\boldsymbol{X}(m)'\boldsymbol{X}(m)\right)^{-1}\boldsymbol{X}(m)'. \tag{13}$$

The MSE has the following characterization.

Theorem 1 *Under Assumption 1*

$$\mathrm{MSE}(m) = \boldsymbol{\mu}'\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\boldsymbol{\mu} + \mathrm{tr}\left(\boldsymbol{P}(m)\boldsymbol{\Sigma}\right).$$

This decomposes (12) into two components. The first component $\boldsymbol{\mu}'\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\boldsymbol{\mu}$ is the bias due to model misspecification, and equals the sum of squared residuals from the regression of the true conditional mean on the model regressors. If a regression is correctly specified in the sense that there are no omitted regressors, this component will equal zero. On the other hand, this component will be strictly positive if there are relevant omitted variables. In general, as the number of regressors are increased, the bias component decreases. The second component $\mathrm{tr}\left(\boldsymbol{P}(m)\boldsymbol{\Sigma}\right)$ is due to estimation variance. In general, as the number of regressors are increased, the variance component increases.

To gain some insight into this expression we consider a stylized example.

Example 1 *The cluster sizes are homogeneous with $n_g = N$. The regressors are constant within clusters and orthonormal across clusters. The true coefficients equal $\beta_j = b\sqrt{a}\,j^{-(1+a)/2}$ for some $b \neq 0$ and $a > 0$. The errors satisfy the random effects structure $e_{ig} = u_g + \varepsilon_{ig}$ with u_g and ε_{ig} i.i.d., $\mathbb{E}\left[u_g\right] = 0$, $\mathbb{E}\left[e_{ig}\right] = 0$, $\mathrm{var}\left[u_g\right] = \sigma_u^2$, and $\mathrm{var}\left[\varepsilon_{ig}\right] = \sigma_\varepsilon^2$. The models are nested with $k(m) = m$. Then*

$$\mathrm{MSE}(m) \simeq \frac{nb^2}{m^a} + \left(\sigma_\varepsilon^2 + N\sigma_u^2\right)m. \tag{14}$$

In (14) we have a precise characterization of how the model dimension m impacts MSE through bias and variance. Most notably, we can see that the cluster-level error u_g impacts the variance component of MSE through the multiplicative factor $N\sigma_u^2$. This shows that cluster size magnifies the impact of within-cluster dependence. Consequently, cluster dependence increases the variance cost of estimating larger models, while having no impact on bias.

The MSE-optimal model in the set $m = 1, ..., M$ is the one with the smallest value of $\text{MSE}(m)$. This is

$$m_{\text{opt}} = \underset{1 \le m \le M}{\text{argmin}} \, \text{MSE}(m). \tag{15}$$

The optimal value of MSE is

$$\min_{1 \le m \le M} \text{MSE}(m) = \text{MSE}(m_{\text{opt}}). \tag{16}$$

As $\text{MSE}(m)$ is unknown, m_{opt} is unknown, and the optimal MSE is not achievable.

We can calculate the optimal model in our stylized example.

Example 2 *In the context of Example 1*

$$m_{\text{opt}} \simeq \left(\frac{anb^2}{\sigma_\varepsilon^2 + N\sigma_u^2} \right)^{1/(1+a)}. \tag{17}$$

Equation (17) shows that the optimal model m_{opt} increases with sample size n, but is decreasing with cluster size N. This means that cluster dependence leads to more parsimonious optimal model choices.

Furthermore, from (17) we can see that the optimal model order m_{opt} rate (as a function of sample size) can be significantly affected by cluster dependence. If $N \sim n^\eta$ is increasing with n for some $0 \le \eta < 1$, then $m_{\text{opt}} \sim n^{(1-\eta)/(1+a)}$. This means that the optimal model rate slows as cluster sizes increase.

6 Asymptotic Optimality of CCV

Our main theoretical result is to provide conditions under which the CCV-selected model $\widehat{m}_{\text{CCV}}$ is asymptotically equivalent to the MSE-optimal model m_{opt}. To describe our assumptions, it will be useful to define the matrices

$$\boldsymbol{P}_g(m) = \boldsymbol{X}_g(m) \left(\boldsymbol{X}(m)' \boldsymbol{X}(m) \right)^{-1} \boldsymbol{X}_g(m)' \tag{18}$$

$$\boldsymbol{M}_g(m) = \boldsymbol{I}_{n_g} - \boldsymbol{P}_g(m) \tag{19}$$

which are the diagonal blocks of $\boldsymbol{P}(m)$ and $\boldsymbol{I}_n - \boldsymbol{P}(m)$. We also define

$$Q_n = \text{MSE}(m_{\text{opt}})$$

$$N_n = \max_{1 \le g \le G} n_g.$$

Q_n is the infeasible optimal MSE. If one of the models contains the true regression, so that the bias term equals zero, then Q_n will equal the variance of this model and will be bounded as n diverges. On the

other hand, if none of the models contain the true regression (which occurs in a nonparametric context such as Example 1) then Q_n will diverge to positive infinity as n diverges. N_n is the largest cluster size. We allow $N_n \to \infty$ to incorporate applications with large cluster sizes.

For any matrix $\boldsymbol{A}$, let $\|\boldsymbol{A}\| = \lambda_{\max}\left(\boldsymbol{A}'\boldsymbol{A}\right)^{1/2}$ denote the spectral norm.

Assumption 1

1. *The observations* $\left(\boldsymbol{Y}_g, \boldsymbol{X}_g\right)$ *are mutually independent across clusters.*
2. $\mathbb{E}\left[\left|e_{ig}\right|^{4r} \mid \boldsymbol{X}\right] \leq D < \infty$ *for some* $r > 1$.
3. $\lambda_{\min}\left(\boldsymbol{\Sigma}_g\right) \geq \underline{\lambda} > 0$, *almost surely.*
4. $\lambda_{\max}\left(\boldsymbol{\Sigma}_g\right) \leq \bar{\lambda} < \infty$, *almost surely.*
5. $\boldsymbol{M}_g(m)$ *is invertible for all* $1 \leq g \leq G$ *and* $1 \leq m \leq M$, *almost surely.*
6. $\max_{1\leq m\leq M} \max_{1\leq g\leq G} \left\|\boldsymbol{P}_g(m)\right\| = o_p(1)$.
7. $k(m) \geq Am^{\phi}$ *for some* $A > 0$ *and* $1/r < \phi \leq 1$.
8. $\dfrac{N_n^{2r}}{Q_n^{r-1/\phi}} = o_p(1)$.

Assumption 1.1 states that the clusters are mutually independent, which means that our analysis is confined to oneway clustering. Assumption 1.2 states that the regression errors have a uniformly bounded conditional moment above four. Assumption 1.3 and 1.4 specify that the eigenvalues of the cluster-level covariance matrices $\boldsymbol{\Sigma}_g$ are uniformly bounded away from zero and infinity. Assumption 1.5 is essentially equivalent to the uniqueness of the delete-one-cluster estimators (9). In Hansen (2024) this condition is called "clusterwise invertibility". This assumption rules out dummy variables which indicate individual clusters, and thus excludes cluster-level treatment with one treated cluster and cluster-level fixed effects from the model regressors $X_{ig}(m)$. Assumption 1.6 states that uniformly across models, cluster-level regressors are asymptotically negligible relative to the full sample. Assumption 1.7 states that the number of regressors is increasing across models at some rate. If the models are conventionally nested then $\phi = 1$. Assumption 1.7 allows for non-nested models, but restricts the number of considered subset models. The constraint on ϕ (and thus the growth rate of subset models) is less restrictive as the number of bounded moments r increases. Assumption 1.8 is a technical condition which relates the rate of growth of the maximal cluster size N_n to the rate of growth of the infeasible optimal MSE Q_n. A necessary condition for Assumption 1.8 is $Q_n \to \infty$. As discussed above, this holds when the true regression function is nonparametric in the sense that all linear regression models are misspecified. If cluster sizes N_n are bounded, then $Q_n \to \infty$ is sufficient for Assumption 1.8. When cluster sizes increase with n, then Assumption 1.8 restricts the rate of growth of the cluster sizes. This rate is less restrictive when Q_n diverges at a fast rate and when ϕr is large. This implies a trade-off between the number of bounded moments r, the growth rate ϕ of the number of subset models, the growth rate of the maximal cluster size N_n, and the growth rate of Q_n.

Theorem 2 *Under Assumption 1, as $n \to \infty$,*

$$\max_{1 \le m \le M} \left| \frac{\mathrm{CCV}(m) - \mathrm{MSE}(m) - \boldsymbol{e}'\boldsymbol{e}}{\mathrm{MSE}(m)} \right| = o_p(1). \tag{20}$$

Theorem 2 shows that the CCV criterion is asymptotically equivalent to the mean square error (12). Thus CCV can be used as a guide to compare and select regression models under clustered dependence.

If CCV is used for model selection, the achieved MSE will be close to the theoretical optimal, as we now show. This is the main result of the paper.

Theorem 3 *Under Assumption 1, as $n \to \infty$,*

$$\frac{\mathrm{MSE}(\widehat{m}_{\mathrm{CCV}})}{\mathrm{MSE}(m_{\mathrm{opt}})} \xrightarrow[p]{} 1.$$

Theorem 3 demonstrates that the mean squared error of the regression model selected by minimizing the CCV criterion is asymptotically equivalent to the infeasible optimal mean squared error. Thus CCV model selection is asymptotically optimal.

7 Conventional CV

We have shown that the delete-cluster cross-validation criterion is asymptotically equivalent to the model MSE. We now show that the conventional delete-one cross-validation criterion is asymptotically equivalent to something different.

Theorem 4 *Under Assumption 1, as $n \to \infty$,*

$$\max_{1 \le m \le M} \left| \frac{\mathrm{CV}(m) - \mathrm{C}(m) - \boldsymbol{e}'\boldsymbol{e}}{\mathrm{MSE}(m)} \right| = o_p(1), \tag{21}$$

where

$$\mathrm{C}(m) = \boldsymbol{\mu}' \left(\boldsymbol{I}_n - \boldsymbol{P}(m) \right) \boldsymbol{\mu} - \mathrm{tr}\left(\boldsymbol{P}(m) \boldsymbol{\Sigma} \right) + 2\,\mathrm{tr}\left(\boldsymbol{P}(m) \boldsymbol{\Sigma}_0 \right) \tag{22}$$

and $\boldsymbol{\Sigma}_0$ is the $n \times n$ diagonal matrix consisting of the diagonal elements of $\boldsymbol{\Sigma}$.

Theorem 4 shows that the conventional cross-validation criterion $\mathrm{CV}(m)$ is asymptotically equivalent to the criterion $\mathrm{C}(m)$. Therefore, the CV-selected model $\widehat{m}_{\mathrm{CV}}$ is asymptotically similar to the minimizer of $\mathrm{C}(m)$, which we define as

$$m_{\mathrm{cv}} = \operatorname*{argmin}_{1 \le m \le M} \mathrm{C}(m). \tag{23}$$

In the absence of clustering, $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}$, so $\mathrm{C}(m) = \mathrm{MSE}(m) + \mathrm{tr}\left(\boldsymbol{\Sigma}\right)$ and $m_{\mathrm{cv}} = m_{\mathrm{opt}}$. Otherwise, it can be quite different. This is easiest to see if we reexamine the example from the previous section .

Example 3 *In the context of Example 1*

$$\mathrm{C}(m) \simeq n\frac{b^2}{m^a} + \left(\sigma_\varepsilon^2 - (N-2)\sigma_u^2\right) m.$$

If $\sigma_\varepsilon^2 - (N-2)\sigma_u^2 > 0$*, the minimizer (23) of (22) equals*

$$m_{\mathrm{cv}} = \min\left[\left(\frac{anb^2}{\sigma_\varepsilon^2 - (N-2)\sigma_u^2}\right)^{1/(1+a)}, M\right].$$

If $\sigma_\varepsilon^2 - (N-2)\sigma_u^2 \leq 0$ *then*

$$m_{\mathrm{cv}} = M.$$

We can see that the criterion $\mathrm{C}(m)$ differs substantially from $\mathrm{MSE}(m)$. The bias terms are the same, but the variances terms are different, with that for $C(m)$ smaller. Thus, $C(m)$ places a reduced cost on model order m. For small levels of clustered correlation the impact is to "select" a larger model than the optimal model m_{opt}. For larger levels of clustered correlation, however, the variance penalty can be decreasing in model dimension m, so the criterion $\mathrm{C}(m)$ is strictly decreasing in m and "selects" the largest model M. The practical implication is that model-selection by delete-one CV leads to over-selection.

8 Fixed Effects

Many estimated clustered regressions include cluster-level fixed effects. Instead of (2), this regression framework can be written as

$$\boldsymbol{Y}_g = \boldsymbol{i}_g \alpha_g + \boldsymbol{\mu}_g + \boldsymbol{e}_g \tag{24}$$

where $\boldsymbol{i}_g$ is an $n_g \times 1$ vector of ones. Similarly, instead of (7), the mth regression model takes the form

$$\boldsymbol{Y}_g = \boldsymbol{i}_g \alpha_g + \boldsymbol{X}_g(m)\beta(m) + \boldsymbol{u}_g(m). \tag{25}$$

This model falls outside of the framework covered by Theorems 2-3 for two interrelated reasons. First, if we interpret (25) as expanding the list of regressors from $\boldsymbol{X}_g(m)$ to $(\boldsymbol{i}_g, \boldsymbol{X}_g(m))$, then Assumption 1.5 fails. Second, if we expand the list of coefficients of the model from β to $(\alpha_1, \ldots \alpha_G, \beta)$, then the delete-one-cluster estimator (9) is not well-defined. Essentially, when the observations in cluster g are omitted, the fixed effect α_g cannot be estimated. These two problems are interrelated, as Assumption 1.5 is precisely the condition needed for existence of the delete-one-cluster estimators.

This problem can be sidestepped if we focus on the model after applying the within transformation. Define the $n_g \times n_g$ within transformation matrix

$$\boldsymbol{W}_g = \boldsymbol{I}_{n_g} - \boldsymbol{i}_g \left(\boldsymbol{i}_g' \boldsymbol{i}_g\right)^{-1} \boldsymbol{i}_g'$$

which subtracts cluster-specific means from cluster-level vectors. For example, the within-transformed

dependent variable equals

$$\dot{\boldsymbol{Y}}_g = \boldsymbol{Y}_g - \frac{1}{n_g}\sum_{i=1}^{n_g} Y_{ig} = \boldsymbol{W}_g \boldsymbol{Y}_g.$$

The within-transformation applied to the regression structure (24) equals

$$\dot{\boldsymbol{Y}}_g = \dot{\boldsymbol{\mu}}_g + \dot{\boldsymbol{e}}_g$$

where $\dot{\boldsymbol{\mu}}_g = \boldsymbol{W}_g \boldsymbol{\mu}_g$ and $\dot{\boldsymbol{e}}_g = \boldsymbol{W}_g \boldsymbol{e}_g$. The transformed regression error satisfies

$$\begin{aligned}\mathbb{E}\left[\dot{\boldsymbol{e}}_g \mid \boldsymbol{X}_g\right] &= 0 \\ \mathbb{E}\left[\dot{\boldsymbol{e}}_g \dot{\boldsymbol{e}}_g' \mid \boldsymbol{X}_g\right] &= \dot{\boldsymbol{\Sigma}}_g = \boldsymbol{W}_g \boldsymbol{\Sigma}_g \boldsymbol{W}_g.\end{aligned}$$

The within-transformation applied to the mth regression model (25) equals

$$\dot{\boldsymbol{Y}}_g = \dot{\boldsymbol{X}}_g(m)\beta(m) + \dot{\boldsymbol{u}}_g(m) \tag{26}$$

where $\dot{\boldsymbol{X}}_g(m) = \boldsymbol{W}_g \boldsymbol{X}_g(m)$ and $\dot{\boldsymbol{u}}_g(m) = \boldsymbol{W}_g \boldsymbol{u}_g(m)$.

The regression model is estimated by least squares. As is well known, this can be achieved either by estimation of the full regression (25) or the transformed equation (26). The least squares esitmate of $\beta(m)$ equals

$$\widehat{\beta}(m) = \left(\sum_{g=1}^{G} \dot{\boldsymbol{X}}_g(m)' \dot{\boldsymbol{X}}_g(m)\right)^{-1} \left(\sum_{g=1}^{G} \dot{\boldsymbol{X}}_g(m)' \dot{\boldsymbol{Y}}_g\right).$$

Write the fitted values as $\widehat{\dot{\boldsymbol{\mu}}}_g(m) = \dot{\boldsymbol{X}}_g(m)\widehat{\beta}(m)$.

The delete-one-cluster coefficient estimates are

$$\widetilde{\beta}_{-g}(m) = \left(\sum_{j\neq g} \dot{\boldsymbol{X}}_j(m)' \dot{\boldsymbol{X}}_j(m)\right)^{-1} \left(\sum_{j\neq g} \dot{\boldsymbol{X}}_j(m)' \dot{\boldsymbol{Y}}_j\right)$$

with prediction errors

$$\widetilde{\dot{\boldsymbol{u}}}_g(m) = \dot{\boldsymbol{Y}}_g - \dot{\boldsymbol{X}}_g(m)\widetilde{\beta}_{-g}(m).$$

The delete-one-cluster within cross-validation (WCCV) criterion is

$$\mathrm{WCCV}(m) = \sum_{g=1}^{G} \widetilde{\dot{\boldsymbol{u}}}_g(m)' \widetilde{\dot{\boldsymbol{u}}}_g(m).$$

Let $\widehat{m}$ be the model which minimizes $\mathrm{WCCV}(m)$.

The MSE of the transformed regression is

$$\mathrm{WMSE}(m) = \sum_{g=1}^{G} \mathbb{E}\left[\left(\widehat{\dot{\boldsymbol{\mu}}}_g(m) - \dot{\boldsymbol{\mu}}_g\right)' \left(\widehat{\dot{\boldsymbol{\mu}}}_g(m) - \dot{\boldsymbol{\mu}}_g\right) \mid \boldsymbol{X}\right]. \tag{27}$$

Let $\dot{m}_{\text{opt}}$ be the model which minimizes $\text{WMSE}(m)$. We also define

$$\begin{aligned}\dot{Q}_n &= \text{WMSE}(\dot{m}_{\text{opt}}) \\ \dot{\boldsymbol{P}}_g(m) &= \dot{\boldsymbol{X}}_g(m)\left(\sum_{g=1}^{G}\dot{\boldsymbol{X}}_g(m)'\dot{\boldsymbol{X}}_g(m)\right)^{-1}\dot{\boldsymbol{X}}_g(m)' \\ \dot{\boldsymbol{M}}_g(m) &= \boldsymbol{I}_{n_g} - \dot{\boldsymbol{P}}_g(m).\end{aligned}$$

Assumption 2 *Assumption 1 holds, with parts 1.5-1.7 replaced by*

5' $\dot{\boldsymbol{M}}_g(m)$ *is invertible for all* $1 \le g \le G$ *and* $1 \le m \le M$.

6' $\max_{1\le m\le M}\max_{1\le g\le G}\left\|\dot{\boldsymbol{P}}_g(m)\right\| = o_p(1)$.

7' $\dfrac{N_n^{2r}}{\dot{Q}_n^{r-1/\phi}} = o_p(1)$.

Theorem 5 *Under Assumption 2, as* $n \to \infty$,

$$\frac{\text{WMSE}(\widehat{m})}{\text{WMSE}(\dot{m}_{\text{opt}})} \xrightarrow[p]{} 1.$$

Theorem 5 shows that model selection by minimization of the within cross-validation criterion is asymptotically optimal, when optimality is assessed by the MSE of the within-transformed regression. Therefore, WCCV can be used to compare and select among regressions estimated with fitted fixed effects, but cannot be used to compare models with fixed effects with models without fixed effects.

9 Simulation

We present a simulation experiment to investigate the performance of model selection methods under clustered dependence. The model is the linear regression

$$Y_{ig} = \alpha + \sum_{j=1}^{J}\beta_j X_{jig} + e_{ig}$$

with the observations grouped into G clusters of size N, so the full sample has $n = NG$ observations. We vary $G \in \{20, 100\}$ and $N \in \{20, 100\}$. The number of regressors is set at $J = \sqrt{n}$. The errors e_{ig} are independent of the regressors and are distributed $N(0,1)$ with within-cluster correlation ρ. We vary $\rho \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. In the baseline model the regressors X_{jig} are distributed $N(0,1)$ with within-cluster correlation ρ. The coefficients are set as $\alpha = 0$ and $\beta_j = b\sqrt{a}\,j^{-(1+a)/2}$, with $b \in \{0.3, 1\}$ and $a \in \{1, 2, 3\}$. For each parameter configuration we generate 20,000 samples.

For each simulation sample we estimate $M = 20$ nested models. The mth includes the first $k_m = mJ/M$ regressors. The Mth includes all J regressors. We select among the M models by CCV and CV,

Table 2: Normal Regressors, $a = 1$

			$b = 0.3$					$b = 1.0$				
G	N	ρ	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
20	20	m_{opt}	6	6	5	3	2	20	19	16	10	6
		$\mathbb{E}[\widehat{m}_{\text{CCV}}]$	7	7	6	6	5	16	16	14	11	6
		$\mathbb{E}[\widehat{m}_{\text{CV}}]$	7	8	12	18	19	16	17	17	19	19
		RMSE(CCV)	1.30	1.29	1.35	1.57	2.29	1.12	1.12	1.13	1.16	1.39
		RMSE(CV)	1.30	1.33	1.74	2.65	5.24	1.12	1.09	1.05	1.16	1.78
20	100	m_{opt}	14	10	6	4	2	40	38	24	12	6
		$\mathbb{E}[\widehat{m}_{\text{CCV}}]$	15	13	10	10	12	36	32	25	20	18
		$\mathbb{E}[\widehat{m}_{\text{CV}}]$	15	23	38	39	40	36	37	39	40	40
		RMSE(CCV)	1.18	1.19	1.34	1.63	2.55	1.08	1.08	1.09	1.12	1.35
		RMSE(CV)	1.18	1.43	2.19	2.82	4.32	1.07	1.02	1.06	1.15	1.40
100	20	m_{opt}	14	14	10	6	4	40	40	32	20	12
		$\mathbb{E}[\widehat{m}_{\text{CCV}}]$	15	14	11	8	6	36	35	30	22	14
		$\mathbb{E}[\widehat{m}_{\text{CV}}]$	15	17	29	38	40	36	36	37	39	40
		RMSE(CCV)	1.19	1.18	1.23	1.33	1.52	1.07	1.07	1.08	1.14	1.24
		RMSE(CV)	1.19	1.23	1.87	3.12	5.75	1.07	1.05	1.04	1.25	1.98
100	100	m_{opt}	30	25	10	5	5	100	80	40	20	15
		$\mathbb{E}[\widehat{m}_{\text{CCV}}]$	32	25	15	9	7	89	77	48	29	17
		$\mathbb{E}[\widehat{m}_{\text{CV}}]$	32	60	99	100	100	89	95	99	100	100
		RMSE(CCV)	1.10	1.11	1.17	1.26	1.44	1.05	1.06	1.10	1.18	1.26
		RMSE(CV)	1.11	1.64	3.29	5.23	11.69	1.04	1.04	1.32	1.91	3.74

denoting the selected models by $\widehat{m}_{\text{CCV}}$ and $\widehat{m}_{\text{CV}}$. Given the selected models we calculate the model-selected coefficient vectors $\widehat{\beta}_{\text{ccv}} = \widehat{\beta}(\widehat{m}_{\text{CCV}})$ and $\widehat{\beta}_{\text{cv}} = \widehat{\beta}(\widehat{m}_{\text{CV}})$, where the non-estimated coefficients are set to zero.

We assess estimation accuracy by coefficient mean-squared error. For the individual model estimators $\widehat{\beta}(m)$ we define

$$\text{MSE}(m) = \mathbb{E}\left[\left(\widehat{\beta}(m) - \beta\right)'\left(\widehat{\beta}(m) - \beta\right)\right]$$

which is similar (but not identical) to the expected squared error (12). Among the estimated models the MSE-optimal model is

$$m_{\text{opt}} = \underset{1 \le m \le M}{\text{argmin}}\, \text{MSE}(m).$$

We compare the model selection methods by contrasting m_{opt} with the average selected models $\mathbb{E}[\widehat{m}_{\text{CCV}}]$ and $\mathbb{E}[\widehat{m}_{\text{CV}}]$.

For the selection estimators $\widehat{\beta}_{\text{ccv}}$ and $\widehat{\beta}_{\text{cv}}$ we define the coefficient mean-squared error

$$\text{MSE(CCV)} = \mathbb{E}\left[\left(\widehat{\beta}_{\text{ccv}} - \beta\right)'\left(\widehat{\beta}_{\text{ccv}} - \beta\right)\right]$$
$$\text{MSE(CV)} = \mathbb{E}\left[\left(\widehat{\beta}_{\text{cv}} - \beta\right)'\left(\widehat{\beta}_{\text{cv}} - \beta\right)\right]$$

and the relative mean-squared error

$$\mathrm{RMSE(CCV)} = \frac{\mathrm{MSE(CCV)}}{\mathrm{MSE}(m_{\mathrm{opt}})}$$
$$\mathrm{RMSE(CV)} = \frac{\mathrm{MSE(CV)}}{\mathrm{MSE}(m_{\mathrm{opt}})}.$$

By construction the RMSE measures are greater than 1. Values close to 1 imply that the selection estimator has near-optimal accuracy. Large values of RMSE indicate that the selection estimator has poor accuracy.

Table 3: Normal Regressors, $a = 3$

			$b = 0.3$					$b = 1.0$				
G	N	ρ	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
20	20	m_{opt}	4	4	3	3	2	6	6	5	4	4
		$\mathbb{E}[\widehat{m}_{\mathrm{CCV}}]$	4	4	5	5	5	7	7	7	7	7
		$\mathbb{E}[\widehat{m}_{\mathrm{CV}}]$	4	5	10	18	19	7	7	11	18	19
		RMSE(CCV)	1.55	1.47	1.57	1.79	2.79	1.34	1.32	1.37	1.53	1.95
		RMSE(CV)	1.55	1.58	2.55	3.83	6.97	1.34	1.37	1.82	2.51	3.89
20	100	m_{opt}	6	4	4	2	2	10	8	6	4	4
		$\mathbb{E}[\widehat{m}_{\mathrm{CCV}}]$	6	6	7	8	11	10	10	10	11	13
		$\mathbb{E}[\widehat{m}_{\mathrm{CV}}]$	6	14	37	39	40	10	17	37	39	40
		RMSE(CCV)	1.32	1.26	1.61	1.84	3.12	1.23	1.23	1.42	1.64	2.27
		RMSE(CV)	1.34	2.09	3.84	3.99	5.88	1.23	1.71	2.72	2.85	3.34
100	20	m_{opt}	6	6	4	4	4	10	10	8	6	6
		$\mathbb{E}[\widehat{m}_{\mathrm{CCV}}]$	6	6	6	5	4	10	10	9	8	8
		$\mathbb{E}[\widehat{m}_{\mathrm{CV}}]$	6	7	23	38	40	10	11	25	38	40
		RMSE(CCV)	1.38	1.31	1.36	1.63	1.69	1.25	1.23	1.28	1.40	1.44
		RMSE(CV)	1.38	1.46	3.85	6.71	9.82	1.25	1.32	2.59	4.17	6.18
100	100	m_{opt}	10	5	5	5	5	15	10	10	5	5
		$\mathbb{E}[\widehat{m}_{\mathrm{CCV}}]$	9	8	6	6	6	15	13	11	8	7
		$\mathbb{E}[\widehat{m}_{\mathrm{CV}}]$	9	31	99	100	100	15	36	99	100	100
		RMSE(CCV)	1.26	1.14	1.27	1.39	1.51	1.18	1.11	1.26	1.28	1.47
		RMSE(CV)	1.26	2.86	9.05	11.33	17.72	1.18	2.28	5.74	6.95	13.87

We report the simulation results in Tables 2-3, with Table 2 reporting results for $a = 1$ and Table 3 for $a = 3$. In each table the first five columns are for $b = 0.3$ and the second set of five columns for $b = 1.0$. The columns vary by the within-cluster correlation ρ. The results are displayed in four groups, with the first group for the setting $G = N = 20$ and the final group for the setting $G = N = 100$. Within each group we report the MSE-optimal model m_{opt}, the average selected models $\mathbb{E}[\widehat{m}_{\mathrm{CCV}}]$ and $\mathbb{E}[\widehat{m}_{\mathrm{CV}}]$, and the relative mean-squared errors of the two model selection methods.

First examine Table 2 and the displayed values of m_{opt}. What you can see is that the optimal model order is strongly affected by the parameters. As should be generally expected, the optimal model order is increasing in regression signal (b), number of clusters (G), and cluster size (N). More importantly,

the optimal model order is strongly decreasing in the within-cluster correlation ρ. For example, when $G = N = 100$ and $b = 1$, the optimal model order is $m_{\text{opt}} = 100$ for $\rho = 0.1$ but is $m_{\text{opt}} = 15$ for $\rho = .9$. This illustrates how important it is to incorporate within-cluster correlation when considering model selection.

Next examine the displayed values of $\mathbb{E}[\widehat{m}_{\text{CCV}}]$, the average value of the CCV-selected model. What we can see is that it tracks the optimal model order m_{opt} well in large samples, and moderately well in smaller samples.

Next examine the displayed values of $\mathbb{E}[\widehat{m}_{\text{CV}}]$, the average value of the conventional CV-selected model. We can see that it is similar to $\mathbb{E}[\widehat{m}_{\text{CCV}}]$ for small values of ρ, but they are very different for large values of ρ. While the optimal model order m_{opt} is decreasing with ρ, $\mathbb{E}[\widehat{m}_{\text{CV}}]$ is increasing in ρ. This discrepancy is particular striking in large samples. For example, when $G = N = 100$ and $\rho \geq 0.5$, CV-selection essentially picks the largest model ($m = 100$) in all simulation replications, while the optimal model order is small.

Next example the relative mean-squared error of the model selection methods. In most cases the RMSE of CCV selection is slightly elevated over the optimal value of 1.0, and the values are descreasing with sample size. In contrast, the RMSE of CV selection has high variation. It is low for small values of ρ, but is high for large values of ρ. The latter values are increasing with sample size. For example, when $G = N = 100$ and $\rho = 0.9$, the RMSE of CV exceeds 11, meaning that the MSE of CV-selection is 11 times the optimal value.

Comparing CCV and CV, we can see that CCV has equal or lower MSE than CV in every single case examined. The two methods are near equivalents when $\rho = 0.1$, but differ meaningfully for larger values of ρ.

Now example Table 3 which displays the results for $a = 3$. In this setting the regression coefficients decay more quickly, reducing the bias of low-dimension models, and thus altering the model-selection trade-offs. We can see that the optimal model orders m_{opt} are smaller than in Table 2, but display the same patterns. The other results are similar to Table 2. CCV-selection performs similarly to the optimal model order in most cases, and achieves low RMSE in most cases, especially in large samples. In contrast, conventional CV leads to sever over-selection when ρ is large, leading to large values of RMSE. These latter values are considerably larger than in Table 2, reaching as high as 17.7. In all cases, CCV-selection has equal or lower MSE than CV-selection.

Many papers, including MacKinnon, Nielsen, and Webb (2023c) and Hansen (2024) have argued that the context where conventional robust covariance matrix estimation fails to perform well, and jackknife covariance matrix estimation performs better, is in the presence of regressor leverage, which arises when regressors have heavy-tailed distributions. Given the parallel between cross-validation and the jackknife, we now explore the impact of heavy-tailed regressors. We augment the model by taking the regressors X_{jig} as originally generated (as $N(0,1)$ with within-cluster correlation ρ) and replacing them with the transformed regressors $Z_{jig} = (\exp(X_{jig}) - \exp(1/2))/\sqrt{\exp(2) - \exp(1)}$. These regressors are log-normally distributed, but scaled to have zero mean and unit variance, and preserve the within-cluster dependence of the original regressors. Otherwise, the experience is unaltered. The results (not reported)

are remarkably similar to Tables 2-3. Essentially, the presence of leveraged (heavy-tailed) regressors has no meaningful impact on the performance of CCV and CV model selection.

The simulation evidence supports the use of CCV model selection for clustered regression. Model selection methods which neglect dependence should be avoided.

10 Conclusion

Model selection methods are routinely used for model evaluation, especially with the rise of high-dimensional and machine-learning estimation methods. At the same time, economic data sets typically display clustered sampling dependence. This paper investigates cross-valiation model-selection methods under clustered dependence, demonstrates that a computationally inexpensive cluster cross-validation (CCV) method is asymptotically optimal, while conventional CV methods can lead to severe over-selection.

The proposed CCV method, similarly to cluster-robust covariance matrix estimation, requires the knowledge of the appropriate clustering structure. If the clustering structure is mis-specified, then the CCV method will be mis-specified and can lead to over-selection similarly to conventional CV.

The results of this paper are confined to one-way clustering. Extensions to two-way and multi-way clustering would be desirable.

11 Appendix

11.1 Whittle Inequalities

Whittle (1960) provided two powerful inequalities which bound real-valued sums and quadratic forms of independent random variables. We generalize these inequalities to the clustered dependence setting.

The following is a generalization of Whittle's first inequality.

Lemma 1 *For any $p \geq 2$, any $n \times d$ matrix $\boldsymbol{A}$, and any $n \times 1$ random vector $\boldsymbol{e} = (\boldsymbol{e}_1', ..., \boldsymbol{e}_G')'$ with $n_g \times 1$ $\boldsymbol{e}_g$ mutually independent across g and $\mathbb{E}\left[\boldsymbol{e}_g\right] = 0$, then*

$$\mathbb{E}\left\|\boldsymbol{A}'\boldsymbol{e}\right\|^p \leq 2M_p\left(\operatorname{tr}\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{p/2} \max_{1\leq g\leq G} \mathbb{E}\left\|\boldsymbol{e}_g\right\|^p \tag{28}$$

where $M_p < \infty$ is the matrix Rosenthal constant from B. E. Hansen (2015) which only depends on p.

The following is a generalization of Whittle's (1960) second inequality.

Lemma 2 *For any $q \geq 2$, any $n \times n$ matrix $\boldsymbol{P}$, and any $n \times 1$ random vector $\boldsymbol{e} = (\boldsymbol{e}_1', ..., \boldsymbol{e}_G')'$ with $n_g \times 1$ $\boldsymbol{e}_g$ mutually independent across g and $\mathbb{E}\left[\boldsymbol{e}_g\right] = 0$, then*

$$\mathbb{E}\left|\boldsymbol{e}'\boldsymbol{P}\boldsymbol{e} - \mathbb{E}\left[\boldsymbol{e}'\boldsymbol{P}\boldsymbol{e}\right]\right|^q \leq C_q\left(\operatorname{tr}\left(\boldsymbol{P}'\boldsymbol{P}\right)\right)^{q/2} \max_{1\leq g\leq G} \mathbb{E}\left\|\boldsymbol{e}_g\right\|^{2q}. \tag{29}$$

where

$$C_q = \sqrt{\frac{2^{1+5q}}{\pi}}\Gamma((q+1)/2)M_{2q}^{1/2} \tag{30}$$

and $M_{2q} < \infty$ is the matrix Rosenthal constant from B. E. Hansen (2015).

The bounds (28) and (29) are written in terms of the moments of the cluster-level errors $\boldsymbol{e}_g$. Typically we will want to use bounds written in terms of the moments of the individual errors. For this, the following result is useful.

Lemma 3 *Write $\boldsymbol{e}_g = (e_{1g}, ..., e_{n_g g})'$. If $n_g \le N_n$ and for some $p \ge 2$, $\mathbb{E}\left|e_{ig}\right|^p \le D < \infty$, then*

$$\max_{1\le g\le G} \mathbb{E}\left\|\boldsymbol{e}_g\right\|^p \le N_n^{p/2} D.$$

Proof of Lemma 1. Make the partition $\boldsymbol{A}' = \left(\boldsymbol{A}_1', ..., \boldsymbol{A}_G'\right)$ conformably with $\boldsymbol{e} = (\boldsymbol{e}_1', ..., \boldsymbol{e}_G')'$. The vectors $\boldsymbol{A}_g'\boldsymbol{e}_g$ are mean zero and independent across g. Let $\|\boldsymbol{A}\|_F = \left(\mathrm{tr}\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2}$ denote the Frobenius norm of a matrix $\boldsymbol{A}$. By the matrix Rosenthal inequality of B. E. Hansen (2015), there is a constant $M_p < \infty$ depending only on p such that

$$\begin{aligned}
\mathbb{E}\left\|\boldsymbol{A}'\boldsymbol{e}\right\|^p &= \mathbb{E}\left\|\sum_{g=1}^G \boldsymbol{A}_g'\boldsymbol{e}_g\right\|^p \\
&\le M_p\left(\left(\sum_{g=1}^G \mathbb{E}\left\|\boldsymbol{A}_g'\boldsymbol{e}_g\right\|^2\right)^{p/2} + \sum_{g=1}^G \mathbb{E}\left\|\boldsymbol{A}_g'\boldsymbol{e}_g\right\|^p\right) \qquad (31)\\
&\le 2M_p\left(\sum_{g=1}^G \left(\mathbb{E}\left\|\boldsymbol{A}_g'\boldsymbol{e}_g\right\|^p\right)^{2/p}\right)^{p/2} \\
&\le 2M_p\left(\sum_{g=1}^G \left\|\boldsymbol{A}_g\right\|_F^2 \left(\mathbb{E}\left\|\boldsymbol{e}_g\right\|^p\right)^{2/p}\right)^{p/2} \\
&\le 2M_p\left(\mathrm{tr}\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{p/2} \max_{1\le g\le G} \mathbb{E}\left\|\boldsymbol{e}_g\right\|^p. \qquad (32)
\end{aligned}$$

This is (28). The second inequality holds by applying Lyapunov's inequality to the first term in (31) and applying the c_r inequality to the second term. The third inequality is the Schwarz matrix inequality. The fourth uses the relationship $\sum_{g=1}^G \left\|\boldsymbol{A}_g\right\|_F^2 = \mathrm{tr}\left(\boldsymbol{A}'\boldsymbol{A}\right)$. ■

Proof of Lemma 2. Let $\boldsymbol{u} = (\boldsymbol{u}_1', ..., \boldsymbol{u}_G')'$ be an independent copy of $\boldsymbol{e}$ and let $\mathbb{E}_u$ denote expectations over

$\boldsymbol{u}$. By the equality $\mathbb{E}\left[\boldsymbol{e}'\boldsymbol{P}\boldsymbol{e}\right] = \mathbb{E}_u\left[\boldsymbol{u}'\boldsymbol{P}\boldsymbol{u}\right]$, Jensen's inequality, and iterated expectations,

$$\begin{aligned}
\mathbb{E}\left|\boldsymbol{e}'\boldsymbol{P}\boldsymbol{e} - \mathbb{E}\left[\boldsymbol{e}'\boldsymbol{P}\boldsymbol{e}\right]\right|^q &= \mathbb{E}\left|\boldsymbol{e}'\boldsymbol{P}\boldsymbol{e} - \mathbb{E}_u\left[\boldsymbol{u}'\boldsymbol{P}\boldsymbol{u}\right]\right|^q \\
&= \mathbb{E}\left|\mathbb{E}_u\left[\boldsymbol{e}'\boldsymbol{P}\boldsymbol{e} - \boldsymbol{u}'\boldsymbol{P}\boldsymbol{u}\right]\right|^q \\
&\leq \mathbb{E}\left[\mathbb{E}_u\left|\boldsymbol{e}'\boldsymbol{P}\boldsymbol{e} - \boldsymbol{u}'\boldsymbol{P}\boldsymbol{u}\right|^q\right] \\
&= \mathbb{E}\left|\boldsymbol{e}'\boldsymbol{P}\boldsymbol{e} - \boldsymbol{u}'\boldsymbol{P}\boldsymbol{u}\right|^q \\
&= \mathbb{E}\left|(\boldsymbol{e}+\boldsymbol{u})'\boldsymbol{P}(\boldsymbol{e}-\boldsymbol{u})\right|^q \\
&= \mathbb{E}\left|\sum_{g=1}^{G}(\boldsymbol{e}+\boldsymbol{u})'\boldsymbol{P}_g\left(\boldsymbol{e}_g - \boldsymbol{u}_g\right)\right|^q \\
&= \mathbb{E}\left|\sum_{g=1}^{G} a_g\right|^q, \qquad (33)
\end{aligned}$$

where the second-to-last equality makes the partition $\boldsymbol{P} = (\boldsymbol{P}_1, ..., \boldsymbol{P}_G)$ conformably with $\boldsymbol{e} = (\boldsymbol{e}_1', ..., \boldsymbol{e}_G')'$ and the final equality sets $a_g = (\boldsymbol{e}+\boldsymbol{u})'\boldsymbol{P}_g\left(\boldsymbol{e}_g - \boldsymbol{u}_g\right)$.

Because $\boldsymbol{e}_g$ and $\boldsymbol{u}_g$ have the same distribution, $\boldsymbol{e}_g - \boldsymbol{u}_g$ has the same distribution as $\boldsymbol{u}_g - \boldsymbol{e}_g$ and a_g has the same distribution as $-a_g$. Thus a_g has the same distribution as $a_g\varepsilon_g$, where ε_g is an independent Rademacher random variable. Since the a_g are independent across g, it follows that $(a_1, ..., a_G)$ has the same distribution as $(a_1\varepsilon_1, ..., a_G\varepsilon_G)$. Let $\mathbb{E}_\varepsilon$ denote expectations over ε. By iterated expectations, (33) equals

$$\mathbb{E}\left[\mathbb{E}_\varepsilon\left|\sum_{g=1}^{G} a_g\varepsilon_g\right|^q\right] \leq \mathbb{E}\left[K_q\left(\sum_{g=1}^{G} a_g^2\right)^{q/2}\right] \leq K_q\left(\sum_{g=1}^{G}\left(\mathbb{E}\left|a_g\right|^q\right)^{2/q}\right)^{q/2}. \qquad (34)$$

The first inequality is Khintchine's inequality (with $K_q = 2^{q/2}\Gamma((q+1)/2)/\pi^{1/2}$, see equation (B.45) in B. E. Hansen (2022)) and the second is Minkowski's inequality.

By multiple application of Minkowski's inequality, identical distributions, the Schwarz inequality, the Cauchy-Schwarz inequality, and finally (28) with $p = 2q$ and $\boldsymbol{A} = \boldsymbol{P}_g$,

$$\begin{aligned}
\left(\mathbb{E}\left|a_g\right|^q\right)^{1/q} &= \left(\mathbb{E}\left|(\boldsymbol{e}+\boldsymbol{u})'\boldsymbol{P}_g\left(\boldsymbol{e}_g - \boldsymbol{u}_g\right)\right|^q\right)^{1/q} \\
&\leq 2\left(\mathbb{E}\left|\boldsymbol{e}'\boldsymbol{P}_g\boldsymbol{e}_g\right|^q\right)^{1/q} + 2\left(\mathbb{E}\left|\boldsymbol{e}'\boldsymbol{P}_g\boldsymbol{u}_g\right|^q\right)^{1/q} \\
&\leq 2\left(\mathbb{E}\left[\left\|\boldsymbol{P}_g'\boldsymbol{e}\right\|^q\left\|\boldsymbol{e}_g\right\|^q\right]\right)^{1/q} + 2\left(\mathbb{E}\left[\left\|\boldsymbol{P}_g'\boldsymbol{e}\right\|^q\left\|\boldsymbol{u}_g\right\|^q\right]\right)^{1/q} \\
&\leq 4\left(\mathbb{E}\left\|\boldsymbol{P}_g'\boldsymbol{e}\right\|^{2q}\right)^{1/2q}\left(\mathbb{E}\left\|\boldsymbol{e}_g\right\|^{2q}\right)^{1/2q} \\
&\leq 4\left(2M_{2q}\left(\operatorname{tr}(\boldsymbol{P}_g'\boldsymbol{P}_g)\right)^q \max_{1\leq g\leq G}\mathbb{E}\left\|\boldsymbol{e}_g\right\|^{2q}\right)^{1/2q}\left(\mathbb{E}\left\|\boldsymbol{e}_g\right\|^{2q}\right)^{1/2q} \\
&\leq 2^{2+1/2q}M_{2q}^{1/2q}\left(\operatorname{tr}(\boldsymbol{P}_g'\boldsymbol{P}_g)\right)^{1/2}\left(\max_{1\leq g\leq G}\mathbb{E}\left\|\boldsymbol{e}_g\right\|^{2q}\right)^{1/q}. \qquad (35)
\end{aligned}$$

Together, (33), (34), (35), and $\sum_{g=1}^{G} \text{tr}(\boldsymbol{P}_g'\boldsymbol{P}_g) = \text{tr}\left(\boldsymbol{P}'\boldsymbol{P}\right)$ establish that

$$\begin{aligned}\mathbb{E}\left|\boldsymbol{e}'\boldsymbol{P}\boldsymbol{e} - \mathbb{E}\left[\boldsymbol{e}'\boldsymbol{P}\boldsymbol{e}\right]\right|^q &\leq K_q \left(\sum_{g=1}^{G}\left(2^{2+1/2q} M_{2q}^{1/2q}\left(\text{tr}(\boldsymbol{P}_g'\boldsymbol{P}_g)\right)^{1/2}\right)^2\right)^{q/2} \\ &= 2^{2q+1/2} K_q M_{2q}^{1/2}\left(\text{tr}\left(\boldsymbol{P}'\boldsymbol{P}\right)\right)^{q/2} \max_{1\leq g\leq G} \mathbb{E}\left\|\boldsymbol{e}_g\right\|^{2q},\end{aligned}$$

which is (29). ■

Proof of Lemma 3. By the definitition of the Euclidean norm and the c_r inequality,

$$\mathbb{E}\left\|\boldsymbol{e}_g\right\|^p = \mathbb{E}\left[\left(\sum_{i=1}^{n_g} e_{ig}^2\right)^{p/2}\right] \leq n_g^{p/2-1}\sum_{i=1}^{n_g}\mathbb{E}\left|e_{ig}\right|^p \leq n_g^{p/2} D \leq N_n^{p/2} D. \tag{36}$$

The final two inequalities are the assumptions $\mathbb{E}\left|e_{ig}\right|^p \leq D$ and $n_g \leq N_n$. ■

11.2 Intermediate Results

In this section we list some intermediate results which will be used for the main proofs. Define the delete-cluster fitted values $\widetilde{\boldsymbol{\mu}}_g(m) = \boldsymbol{X}_g(m)\widetilde{\beta}_{-g}(m)$ and stack into the full-sample vectors $\widetilde{\boldsymbol{\mu}}(m)$. Similarly, define the delete-one fitted values $\bar{\mu}_i(m) = X_i'(m)\bar{\beta}_{-i}(m)$ and stack into the full-sample vectors $\bar{\boldsymbol{\mu}}(m)$.

Lemma 4 *Under Assumption 1.5, the vector of delete-cluster and delete-one fitted values can be written as*

$$\widetilde{\boldsymbol{\mu}}(m) = \widetilde{\boldsymbol{P}}(m)\boldsymbol{Y} \tag{37}$$

and

$$\bar{\boldsymbol{\mu}}(m) = \bar{\boldsymbol{P}}(m)\boldsymbol{Y} \tag{38}$$

where

$$\widetilde{\boldsymbol{P}}(m) = \boldsymbol{P}(m) - \widetilde{\boldsymbol{D}}(m)\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right) \tag{39}$$

$$\widetilde{\boldsymbol{D}}(m) = \text{diag}\left\{\boldsymbol{M}_g(m)^{-1} - \boldsymbol{I}_{n_g}\right\} \tag{40}$$

$$\bar{\boldsymbol{P}}(m) = \boldsymbol{P}(m) - \bar{\boldsymbol{D}}(m)\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right) \tag{41}$$

$$\bar{\boldsymbol{D}}(m) = \text{diag}\left\{\frac{P_{ig}(m)}{1 - P_{ig}(m)}\right\}, \tag{42}$$

$\boldsymbol{P}(m)$ and $\boldsymbol{M}_g(m)$ are defined in (13) and (19), and $P_{ig}(m) = X_{ig}(m)'\left(\boldsymbol{X}(m)'\boldsymbol{X}(m)\right)^{-1}X_{ig}(m)$. The $n\times n$ matrix $\widetilde{\boldsymbol{P}}(m)$ has the property that its diagonal $n_g\times n_g$ blocks (corresponding to each cluster) contain only 0's. The $n\times n$ matrix $\bar{\boldsymbol{P}}(m)$ has the property that its diagonal elements equal 0. The matrices $\widetilde{\boldsymbol{D}}(m)$ and $\bar{\boldsymbol{D}}(m)$ are symmetric.

Proof. We demonstrate (37), as (38) is the special case where each cluster has one observation. To simplify the notation for ease of reading, we omit notational dependence on the model m. Using the Woodbury matrix identity, the delete-cluster fitted values equal

$$\begin{aligned}
\widetilde{\boldsymbol{\mu}}_g &= \boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{X}_g'\boldsymbol{Y}_g\right) \\
&= \boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\boldsymbol{Y}_g \\
&+ \boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\left(\boldsymbol{I}_{n_g} - \boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\right)^{-1}\boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y} \\
&- \boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\left(\boldsymbol{I}_{n_g} - \boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\right)^{-1}\boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\boldsymbol{Y}_g \\
&= \boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y} - \left(\boldsymbol{I}_{n_g} - \boldsymbol{M}_g\right)\boldsymbol{Y}_g \\
&+ \left(\boldsymbol{I}_{n_g} - \boldsymbol{M}_g\right)\boldsymbol{M}_g^{-1}\boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y} - \left(\boldsymbol{I}_{n_g} - \boldsymbol{M}_g\right)\boldsymbol{M}_g^{-1}\left(\boldsymbol{I}_{n_g} - \boldsymbol{M}_g\right)\boldsymbol{Y}_g \\
&= \boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y} - \left(\boldsymbol{M}_g^{-1} - \boldsymbol{I}_{n_g}\right)\left(\boldsymbol{Y}_g - \boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}\right).
\end{aligned}$$

The third equality makes multiple use of (19). The fourth equality simplifies terms.

Stacking over $g = 1, ..., G$ and using (39)-(40) we obtain

$$\begin{aligned}
\widetilde{\boldsymbol{\mu}} &= \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y} - \widetilde{\boldsymbol{D}}\left(\boldsymbol{Y} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}\right) \\
&= \left(\boldsymbol{P} - \widetilde{\boldsymbol{D}}\left(\boldsymbol{I}_n - \boldsymbol{P}\right)\right)\boldsymbol{Y} \\
&= \widetilde{\boldsymbol{P}}\boldsymbol{Y}
\end{aligned}$$

which is (37).

To show that the diagonal blocks of $\widetilde{\boldsymbol{P}}$ contain only 0's, observe that the diagonal blocks of $\boldsymbol{I}_n - \boldsymbol{P}$ are $\boldsymbol{M}_g$, so those of $\widetilde{\boldsymbol{D}}\left(\boldsymbol{I}_n - \boldsymbol{P}\right)$ are

$$\left(\boldsymbol{M}_g^{-1} - \boldsymbol{I}_{n_g}\right)\boldsymbol{M}_g = \boldsymbol{I}_{n_g} - \boldsymbol{M}_g$$

which are the diagonal blocks of $\boldsymbol{P}$. We deduce that the diagonal blocks of $\boldsymbol{P}$ and $\widetilde{\boldsymbol{D}}\left(\boldsymbol{I}_n - \boldsymbol{P}\right)$ are identical, so those of $\widetilde{\boldsymbol{P}}$ are 0's, as claimed.

The matrices $\boldsymbol{M}_g(m)$ are symmetric which implies that the diagonal blocks in (40) are symmetric and thus $\widetilde{\boldsymbol{D}}(m)$ is symmetric as claimed.

To show that the diagonal elements of $\bar{\boldsymbol{P}}(m)$ equal 0, observe that since $\bar{\boldsymbol{D}}(m)$ is diagonal and the diagonal elements of $\boldsymbol{P}(m)$ are $P_{ig}(m)$, the diagonal elements of $\bar{\boldsymbol{D}}(m)\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)$ equal

$$\frac{P_{ig}(m)}{\left(1 - P_{ig}(m)\right)}\left(1 - P_{ig}(m)\right) = P_{ig}(m),$$

which equal the diagonal elements of $\boldsymbol{P}(m)$. Hence, the diagonal elements of $\bar{\boldsymbol{P}}(m) = \boldsymbol{P}(m) - \bar{\boldsymbol{D}}(m)\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)$ equal zero, as claimed. ■

Lemma 5 *Under Assumption 1.5-1.6, as $n \to \infty$,*

$$\max_{1\le m\le M} \left\| \widetilde{\boldsymbol{D}}(m) \right\| = o_p(1) \tag{43}$$

$$\max_{1\le m\le M} \left\| \widetilde{\boldsymbol{P}}(m) \right\| = 1 + o_p(1) \tag{44}$$

$$\max_{1\le m\le M} \left\| \boldsymbol{I}_n - \widetilde{\boldsymbol{P}}(m) \right\| = 1 + o_p(1) \tag{45}$$

$$\max_{1\le m\le M} \left\| \tilde{\boldsymbol{D}}(m) \right\| = o_p(1) \tag{46}$$

$$\max_{1\le m\le M} \left\| \tilde{\boldsymbol{P}}(m) \right\| = 1 + o_p(1) \tag{47}$$

$$\max_{1\le m\le M} \left\| \boldsymbol{I}_n - \tilde{\boldsymbol{P}}(m) \right\| = 1 + o_p(1) \tag{48}$$

Proof. Assumption 1.6 states that

$$\max_{1\le m\le M} \max_{1\le g\le G} \left\| \boldsymbol{I}_{n_g} - \boldsymbol{M}_g(m) \right\| = o_p(1). \tag{49}$$

This implies

$$\max_{1\le m\le M} \left\| \widetilde{\boldsymbol{D}}(m) \right\| = \max_{1\le m\le M} \max_{1\le g\le G} \left\| \boldsymbol{M}_g(m)^{-1} - \boldsymbol{I}_{n_g} \right\| = o_p(1)$$

which is (43). Hence

$$\begin{aligned}
\max_{1\le m\le M} \left\| \widetilde{\boldsymbol{P}}(m) \right\| &= \max_{1\le m\le M} \left\| \boldsymbol{P}(m) - \widetilde{\boldsymbol{D}}(m) \left(\boldsymbol{I}_n - \boldsymbol{P}(m) \right) \right\| \\
&\le \max_{1\le m\le M} \| \boldsymbol{P}(m) \| + \max_{1\le m\le M} \left\| \widetilde{\boldsymbol{D}}(m) \right\| \| \boldsymbol{I}_n - \boldsymbol{P}(m) \| \\
&= 1 + \max_{1\le m\le M} \left\| \widetilde{\boldsymbol{D}}(m) \right\| \\
&= 1 + o_p(1)
\end{aligned}$$

where the second equality uses $\| \boldsymbol{P}(m) \| = 1$ and $\| \boldsymbol{I}_n - \boldsymbol{P}(m) \| = 1$. This is (44). Similarly,

$$\begin{aligned}
\max_{1\le m\le M} \left\| \boldsymbol{I}_n - \widetilde{\boldsymbol{P}}(m) \right\| &= \max_{1\le m\le M} \left\| \boldsymbol{I}_n - \boldsymbol{P}(m) + \widetilde{\boldsymbol{D}}(m) \left(\boldsymbol{I}_n - \boldsymbol{P}(m) \right) \right\| \\
&\le \max_{1\le m\le M} \| \boldsymbol{I}_n - \boldsymbol{P}(m) \| + \max_{1\le m\le M} \left\| \widetilde{\boldsymbol{D}}(m) \right\| \| \boldsymbol{I}_n - \boldsymbol{P}(m) \| \\
&= 1 + o_p(1),
\end{aligned}$$

which is (45).

As $P_{ig}(m)$ is a diagonal element of $\boldsymbol{P}_g(m)$,

$$\max_{1\le m\le M} \max_{i,g} \left| P_{ig}(m) \right| \le \max_{1\le m\le M} \max_{1\le g\le G} \left\| \boldsymbol{P}_g(m) \right\| = o_p(1) \tag{50}$$

where the convergence is (49). This is the delete-one analog of (49). Just as the three results (43)-(45) follow from (49), (46)-(48) follow from (50). ■

11.3 Proof of Theorem 1

Stack $\boldsymbol{Y}_g$, $\boldsymbol{e}_g$, and $\widehat{\boldsymbol{\mu}}_g(m)$ into the full-sample vectors $\boldsymbol{Y}$, $\boldsymbol{e}$, and $\widehat{\boldsymbol{\mu}}_g(m)$. We have

$$\begin{aligned}\widehat{\boldsymbol{\mu}}(m) &= \boldsymbol{P}(m)\boldsymbol{Y} = \boldsymbol{P}(m)\boldsymbol{\mu} + \boldsymbol{P}(m)\boldsymbol{e} \\ \widehat{\boldsymbol{\mu}}(m) - \boldsymbol{\mu} &= -\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\boldsymbol{\mu} + \boldsymbol{P}(m)\boldsymbol{e}.\end{aligned}$$

Using the properties

$$\boldsymbol{P}(m)\boldsymbol{P}(m) = \boldsymbol{P}(m) \tag{51}$$

$$\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\boldsymbol{P}(m) = 0 \tag{52}$$

$$\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right) = \boldsymbol{I}_n - \boldsymbol{P}(m) \tag{53}$$

and $\mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right] = \boldsymbol{\Sigma}$ we find

$$\begin{aligned}\mathrm{MSE}(m) &= \mathbb{E}\left[\left(\widehat{\boldsymbol{\mu}}(m) - \boldsymbol{\mu}\right)'\left(\widehat{\boldsymbol{\mu}}(m) - \boldsymbol{\mu}\right) \mid \boldsymbol{X}\right] \\ &= \boldsymbol{\mu}'\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\boldsymbol{\mu} + \mathbb{E}\left[\boldsymbol{e}'\boldsymbol{P}(m)\boldsymbol{e} \mid \boldsymbol{X}\right] \\ &= \boldsymbol{\mu}'\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\boldsymbol{\mu} + \mathrm{tr}\left(\boldsymbol{P}(m)\boldsymbol{\Sigma}\right)\end{aligned}$$

as claimed. ■

11.4 Proof of Theorem 2

The $n \times 1$ vector of delete-cluster errors can be written as

$$\widetilde{\boldsymbol{u}}(m) = \left(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}(m)\right)\boldsymbol{Y} = \left(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}(m)\right)\left(\boldsymbol{\mu} + \boldsymbol{e}\right).$$

Expanding the quadratic,

$$\begin{aligned}\mathrm{CCV}(m) &= \widetilde{\boldsymbol{u}}(m)'\widetilde{\boldsymbol{u}}(m) \\ &= \left(\boldsymbol{\mu} + \boldsymbol{e}\right)'\left(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}(m)\right)'\left(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}(m)\right)\left(\boldsymbol{\mu} + \boldsymbol{e}\right) \\ &= \boldsymbol{\mu}'\left(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}(m)\right)'\left(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}(m)\right)\boldsymbol{\mu} + \boldsymbol{e}'\boldsymbol{e} - 2\boldsymbol{e}'\widetilde{\boldsymbol{P}}(m)\boldsymbol{e} + \boldsymbol{e}'\widetilde{\boldsymbol{P}}(m)'\widetilde{\boldsymbol{P}}(m)\boldsymbol{e} + 2\boldsymbol{\mu}'\left(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}(m)\right)'\left(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}(m)\right)\boldsymbol{e}.\end{aligned}$$

Using (39) and (51),

$$\mathrm{tr}\left(\widetilde{\boldsymbol{P}}(m)'\widetilde{\boldsymbol{P}}(m)\boldsymbol{\Sigma}\right) = \mathrm{tr}\left(\boldsymbol{P}(m)\boldsymbol{\Sigma}\right) - 2\,\mathrm{tr}\left(\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\widetilde{\boldsymbol{D}}(m)\boldsymbol{P}(m)\boldsymbol{\Sigma}\right) + \mathrm{tr}\left(\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\widetilde{\boldsymbol{D}}(m)\widetilde{\boldsymbol{D}}(m)\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\boldsymbol{\Sigma}\right). \tag{54}$$

Using Theorem 1, (54), and some rearrangment we find

$$\frac{\mathrm{CCV}(m) - \mathrm{MSE}(m) - \boldsymbol{e}'\boldsymbol{e}}{\mathrm{MSE}(m)} = T_1(m) - 2T_2(m) + T_3(m) - 2T_4(m) + T_5(m) + 2T_6(m) \tag{55}$$

where

$$T_1(m) = \frac{\boldsymbol{\mu}'\left(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}(m)\right)'\left(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}(m)\right)\boldsymbol{\mu} - \boldsymbol{\mu}'\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\boldsymbol{\mu}}{\mathrm{MSE}(m)} \tag{56}$$

$$T_2(m) = \frac{\mathrm{tr}\left(\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\widetilde{\boldsymbol{D}}(m)\boldsymbol{P}(m)\boldsymbol{\Sigma}\right)}{\mathrm{MSE}(m)} \tag{57}$$

$$T_3(m) = \frac{\mathrm{tr}\left(\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\widetilde{\boldsymbol{D}}(m)\widetilde{\boldsymbol{D}}(m)\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\boldsymbol{\Sigma}\right)}{\mathrm{MSE}(m)} \tag{58}$$

$$T_4(m) = \frac{\boldsymbol{e}'\widetilde{\boldsymbol{P}}(m)\boldsymbol{e}}{\mathrm{MSE}(m)} \tag{59}$$

$$T_5(m) = \frac{\boldsymbol{e}'\widetilde{\boldsymbol{P}}(m)'\widetilde{\boldsymbol{P}}(m)\boldsymbol{e} - \mathrm{tr}\left(\widetilde{\boldsymbol{P}}(m)'\widetilde{\boldsymbol{P}}(m)\boldsymbol{\Sigma}\right)}{\mathrm{MSE}(m)} \tag{60}$$

$$T_6(m) = \frac{\boldsymbol{\mu}'\left(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}(m)\right)'\left(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}(m)\right)\boldsymbol{e}}{\mathrm{MSE}(m)}. \tag{61}$$

We will show that the components (56)-(61) are each $o_p(1)$ uniformly in $1 \leq m \leq M$, which will establish (20).

It will be useful to observe that Theorem 1, $\mathrm{tr}\left(\boldsymbol{P}(m)\boldsymbol{\Sigma}\right) \geq 0$, and $\boldsymbol{\mu}'\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\boldsymbol{\mu} \geq 0$ imply the two inequalities

$$\boldsymbol{\mu}'\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\boldsymbol{\mu} \leq \mathrm{MSE}(m) \tag{62}$$

and

$$\mathrm{tr}\left(\boldsymbol{P}(m)\boldsymbol{\Sigma}\right) \leq \mathrm{MSE}(m). \tag{63}$$

We will make multiple use of the inequalities

$$\mathrm{tr}\left(\boldsymbol{AB}\right) \leq \|\boldsymbol{A}\|\,\mathrm{tr}\left(\boldsymbol{B}\right) \tag{64}$$

$$\boldsymbol{b}'\boldsymbol{Ab} \leq \boldsymbol{b}'\boldsymbol{b}\,\|\boldsymbol{A}\| \tag{65}$$

for any square matrices $\boldsymbol{A}$, positive semi-definite matrices $\boldsymbol{B}$, and vectors $\boldsymbol{b}$. In addition, (63) and an analog of (64) using Assumption 1.3 imply

$$\mathrm{MSE}(m) \geq \mathrm{tr}\left(\boldsymbol{P}(m)\boldsymbol{\Sigma}\right) \geq \mathrm{tr}\left(\boldsymbol{P}(m)\right)\underline{\lambda} = k(m)\underline{\lambda}. \tag{66}$$

Take (56). Using (39), some algebra, (65), (53), (62), the triangle inequality, and finally (43) we find

$$\begin{aligned}
|T_1(m)| &= \frac{\boldsymbol{\mu}'\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\left(2\widetilde{\boldsymbol{D}}(m) + \widetilde{\boldsymbol{D}}(m)\widetilde{\boldsymbol{D}}(m)\right)\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\boldsymbol{\mu}}{\mathrm{MSE}(m)} \\
&\leq \frac{\left\|2\widetilde{\boldsymbol{D}}(m) + \widetilde{\boldsymbol{D}}(m)\widetilde{\boldsymbol{D}}(m)\right\|\boldsymbol{\mu}'\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\boldsymbol{\mu}}{\mathrm{MSE}(m)} \\
&\leq 2\left\|\widetilde{\boldsymbol{D}}(m)\right\| + \left\|\widetilde{\boldsymbol{D}}(m)\right\|^2 \\
&\leq o_p(1)
\end{aligned} \tag{67}$$

uniformly in m.

Take (57). Rearranging, applying (64), the Schwarz matrix inequality, (63), and finally (43) and (45),

$$\begin{aligned} T_2(m) &= \frac{\operatorname{tr}\left(\boldsymbol{\Sigma}^{1/2}\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\widetilde{\boldsymbol{D}}(m)\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{P}(m)\boldsymbol{\Sigma}^{1/2}\right)}{\mathrm{MSE}(m)} \\ &\leq \frac{\left\|\boldsymbol{\Sigma}^{1/2}\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\widetilde{\boldsymbol{D}}(m)\boldsymbol{\Sigma}^{-1/2}\right\|\operatorname{tr}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{P}(m)\boldsymbol{\Sigma}^{1/2}\right)}{\mathrm{MSE}(m)} \\ &\leq \frac{\left\|\boldsymbol{I}_n - \boldsymbol{P}(m)\right\|\left\|\widetilde{\boldsymbol{D}}(m)\right\|\operatorname{tr}\left(\boldsymbol{P}(m)\boldsymbol{\Sigma}\right)}{\mathrm{MSE}(m)} \\ &\leq \left\|\boldsymbol{I}_n - \boldsymbol{P}(m)\right\|\left\|\widetilde{\boldsymbol{D}}(m)\right\| \\ &\leq o_p(1) \end{aligned} \tag{68}$$

uniformly in m.

Take (58). We first observe that by applying (53), (64), and the fact from Theorem 4 that the diagonal blocks of $\widetilde{\boldsymbol{D}}(m)\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)$ equal those of $\boldsymbol{P}(m)$

$$\begin{aligned} \operatorname{tr}\left(\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\widetilde{\boldsymbol{D}}(m)\widetilde{\boldsymbol{D}}(m)\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\right) &= \operatorname{tr}\left(\widetilde{\boldsymbol{D}}(m)^{1/2}\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\widetilde{\boldsymbol{D}}(m)^{1/2}\widetilde{\boldsymbol{D}}(m)\right) \\ &\leq \operatorname{tr}\left(\widetilde{\boldsymbol{D}}(m)^{1/2}\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\widetilde{\boldsymbol{D}}(m)^{1/2}\right)\left\|\widetilde{\boldsymbol{D}}(m)\right\| \\ &= \operatorname{tr}\left(\widetilde{\boldsymbol{D}}(m)\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\right)\left\|\widetilde{\boldsymbol{D}}(m)\right\| \\ &= \operatorname{tr}\left(\boldsymbol{P}(m)\right)\left\|\widetilde{\boldsymbol{D}}(m)\right\| \\ &\leq \frac{\mathrm{MSE}(m)}{\underline{\lambda}}o_p(1) \end{aligned} \tag{69}$$

uniformly in m, with the final inequality by (66) and (43).

Applying (64) and (69), we find

$$T_3(m) \leq \frac{\operatorname{tr}\left(\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\widetilde{\boldsymbol{D}}(m)\widetilde{\boldsymbol{D}}(m)\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\right)}{\mathrm{MSE}(m)}\bar{\lambda} \leq o_p(1) \tag{70}$$

uniformly in m.

Using (39), (51), (52), (66), and (69), we calculate that

$$\begin{aligned} \frac{\operatorname{tr}\left(\widetilde{\boldsymbol{P}}(m)'\widetilde{\boldsymbol{P}}(m)\right)}{\mathrm{MSE}(m)} &= \frac{\operatorname{tr}\left(\boldsymbol{P}(m)\right) + \operatorname{tr}\left(\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\widetilde{\boldsymbol{D}}(m)\widetilde{\boldsymbol{D}}(m)\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\right)}{\mathrm{MSE}(m)} \\ &\leq \frac{1}{\underline{\lambda}}\left(1 + o_p(1)\right) \end{aligned} \tag{71}$$

uniformly in m.

Take (59). As established by Lemma 4, the diagonal blocks of $\widetilde{\boldsymbol{P}}(m)$ are 0's. Since $\boldsymbol{\Sigma}$ is block diagonal, this implies that the diagonal blocks of $\widetilde{\boldsymbol{P}}(m)\boldsymbol{\Sigma}$ are 0's as well. Hence

$$\mathbb{E}\left[\boldsymbol{e}'\widetilde{\boldsymbol{P}}(m)\boldsymbol{e} \mid \boldsymbol{X}\right] = \operatorname{tr}\left(\widetilde{\boldsymbol{P}}(m)\boldsymbol{\Sigma}\right) = 0. \tag{72}$$

Using (72), applying Lemma 2 with $q = 2r$, Lemma 3 under Assumption 1.2, and (71)

$$\begin{aligned}
\mathbb{E}\left|T_4(m) \mid \boldsymbol{X}\right|^{2r} &= \mathbb{E}\left|\frac{\boldsymbol{e}'\widetilde{\boldsymbol{P}}(m)\boldsymbol{e} - \mathbb{E}\left[\boldsymbol{e}'\widetilde{\boldsymbol{P}}(m)\boldsymbol{e} \mid \boldsymbol{X}\right]}{\mathrm{MSE}(m)} \mid \boldsymbol{X}\right|^{2r} \\
&\leq \frac{C_{2r}\left(\dfrac{\mathrm{tr}\left(\widetilde{\boldsymbol{P}}(m)'\widetilde{\boldsymbol{P}}(m)\right)}{\mathrm{MSE}(m)}\right)^{r} \max\limits_{1\leq g\leq G}\mathbb{E}\left\|\boldsymbol{e}_g\right\|^{4r}}{\mathrm{MSE}(m)^r} \\
&\leq \frac{C_{2r}DN_n^{2r}\left(1+o_p(1)\right)}{\underline{\lambda}^r\mathrm{MSE}(m)^r}
\end{aligned} \tag{73}$$

uniformly in m.

Fix $\varepsilon > 0$. By Boole's inequality, Markov's inequality, and (73),

$$\begin{aligned}
\mathbb{P}\left[\sup_{1\leq m\leq M}|T_4(m)| > \varepsilon \mid \boldsymbol{X}\right] &\leq \sum_{m=1}^{M}\mathbb{P}\left[|T_4(m)| > \varepsilon \mid \boldsymbol{X}\right] \\
&\leq \frac{1}{\varepsilon^{2r}}\sum_{m=1}^{M}\mathbb{E}\left|T_4(m) \mid \boldsymbol{X}\right|^{2r} \\
&\leq \frac{C_{2r}D}{\varepsilon^{2r}\underline{\lambda}^r}\sum_{m=1}^{M}\frac{1}{\mathrm{MSE}(m)^r}N_n^{2r}\left(1+o_p(1)\right) \\
&\leq \frac{C_{2r}D}{\varepsilon^{2r}\underline{\lambda}^r}\left(\sum_{m=1}^{Q_n^{1/\phi}}\frac{1}{\mathrm{MSE}(m)^r} + \frac{1}{\underline{\lambda}^r A^r}\sum_{m=Q_n^{1/\phi}+1}^{M}\frac{1}{m^{\phi r}}\right)N_n^{2r}\left(1+o_p(1)\right) \\
&\leq \frac{C_{2r}D}{\varepsilon^{2r}\underline{\lambda}^r}\left(1+\frac{1}{\underline{\lambda}^r A^r\left(\phi r-1\right)}\right)\frac{N_n^{2r}}{Q_n^{r-1/\phi}}\left(1+o_p(1)\right) \\
&\leq o_p(1).
\end{aligned} \tag{74}$$

For the fourth inequality we apply (66) and Assumption 1.7 to obtain

$$\mathrm{MSE}(m) \geq \underline{\lambda}k(m) \geq \underline{\lambda}Am^{\phi}.$$

For the fifth inequality in (74) we use $\mathrm{MSE}(m) \geq Q_n$ for the first sum, and in the second sum we use

$$\sum_{m=t+1}^{M}\frac{1}{m^{\alpha}} \leq \sum_{m=t+1}^{\infty}\frac{1}{m^{\alpha}} \leq \frac{1}{(\alpha-1)\,t^{\alpha-1}}$$

which holds for $\alpha > 1$ and integer $t \geq 0$. The final inequality in (74) is Assumption 1.8. As ε is arbitrary, (74) establishes

$$T_4(m) = o_p(1) \tag{75}$$

by Markov's inequality.

Take (60). Recognizing that $\mathrm{tr}\left(\widetilde{\boldsymbol{P}}(m)'\widetilde{\boldsymbol{P}}(m)\boldsymbol{\Sigma}\right) = \mathbb{E}\left[\boldsymbol{e}'\widetilde{\boldsymbol{P}}(m)'\widetilde{\boldsymbol{P}}(m)\boldsymbol{e}\right]$, applying Lemma 2 with $q = 2r$,

Lemma 3 under Assumption 1.2, (64), (44), and (71)

$$\begin{aligned}
\mathbb{E}\left|T_5(m)\mid \boldsymbol{X}\right|^{2r} &= \mathbb{E}\left|\frac{\boldsymbol{e}'\widetilde{\boldsymbol{P}}(m)'\widetilde{\boldsymbol{P}}(m)\boldsymbol{e}-\mathbb{E}\left[\boldsymbol{e}'\widetilde{\boldsymbol{P}}(m)'\widetilde{\boldsymbol{P}}(m)\boldsymbol{e}\mid \boldsymbol{X}\right]}{\mathrm{MSE}(m)}\mid \boldsymbol{X}\right|^{2r} \\
&\leq \frac{C_{2r}\left(\frac{\mathrm{tr}\left(\widetilde{\boldsymbol{P}}(m)'\widetilde{\boldsymbol{P}}(m)\widetilde{\boldsymbol{P}}(m)'\widetilde{\boldsymbol{P}}(m)\right)}{\mathrm{MSE}(m)}\right)^r \max_{1\leq g\leq G}\mathbb{E}\left\|\boldsymbol{e}_g\right\|^{4r}}{\mathrm{MSE}(m)^r} \\
&\leq \frac{C_{2r}\left(\frac{\mathrm{tr}\left(\widetilde{\boldsymbol{P}}(m)'\widetilde{\boldsymbol{P}}(m)\right)}{\mathrm{MSE}(m)}\right)^r \left\|\widetilde{\boldsymbol{P}}(m)\right\|^{2r} DN_n^{2r}}{\mathrm{MSE}(m)^r} \\
&\leq \frac{C_{2r}DN_n^{2r}\left(1+o_p(1)\right)}{\underline{\lambda}^r\mathrm{MSE}(m)^r}
\end{aligned} \tag{76}$$

uniformly in m. This is identical to (73). By the same steps as in (74)-(75), we conclude that $T_5(m)=o_p(1)$ uniformly in m.

Take (61). It is useful to observe that by norm monotonicity and Assumption 1.2,

$$\mathbb{E}\left[\left|e_{ig}\right|^{2r}\mid \boldsymbol{X}\right] \leq \left(\mathbb{E}\left[\left|e_{ig}\right|^{4r}\mid \boldsymbol{X}\right]\right)^{1/2} \leq D^{1/2}. \tag{77}$$

By Lemma 1 with $p=2r$, Lemma 3 under (77), and (65),

$$\begin{aligned}
\mathbb{E}\left|T_6(m)\mid \boldsymbol{X}\right|^{2r} &\leq \frac{B_{2r}\left(\frac{\boldsymbol{\mu}'\left(\boldsymbol{I}_n-\widetilde{\boldsymbol{P}}(m)\right)'\left(\boldsymbol{I}_n-\widetilde{\boldsymbol{P}}(m)\right)\left(\boldsymbol{I}_n-\widetilde{\boldsymbol{P}}(m)\right)'\left(\boldsymbol{I}_n-\widetilde{\boldsymbol{P}}(m)\right)\boldsymbol{\mu}}{\mathrm{MSE}(m)}\right)^r \max_{1\leq g\leq G}\mathbb{E}\left\|\boldsymbol{e}_g\right\|^{2r}}{\mathrm{MSE}(m)^r} \\
&\leq \frac{B_{2r}\left(\frac{\boldsymbol{\mu}'\left(\boldsymbol{I}_n-\widetilde{\boldsymbol{P}}(m)\right)'\left(\boldsymbol{I}_n-\widetilde{\boldsymbol{P}}(m)\right)\boldsymbol{\mu}}{\mathrm{MSE}(m)}\right)^r \left\|\boldsymbol{I}_n-\widetilde{\boldsymbol{P}}(m)\right\|^{2r} D^{1/2}N_n^r}{\mathrm{MSE}(m)^r} \\
&\leq \frac{B_{2r}D^{1/2}N_n^r\left(\frac{\boldsymbol{\mu}'\left(\boldsymbol{I}_n-\boldsymbol{P}(m)\right)\boldsymbol{\mu}}{\mathrm{MSE}(m)}+o_p(1)\right)^r \left\|\boldsymbol{I}_n-\widetilde{\boldsymbol{P}}(m)\right\|^{2r}}{\mathrm{MSE}(m)^r} \\
&\leq \frac{B_{2r}D^{1/2}N_n^{2r}\left(1+o_p(1)\right)}{\mathrm{MSE}(m)^r}
\end{aligned} \tag{78}$$

uniformly in m. The third inequality follows from $T_1(m)=o_p(1)$. The final inequality uses (62), (45), and $N_n^r\leq N_n^{2r}$.

The bound (78) is of the same form as (73). By the same steps as in (74)-(75), we conclude that

$$T_6(m)=o_p(1)$$

uniformly in m.

This completes the proof. ■

11.5 Proof of Theorem 3

Define the normalized delete-one-cluster cross-validation criterion

$$\mathrm{CCV}^*(m) = \mathrm{CCV}(m) - \boldsymbol{e}'\boldsymbol{e}. \tag{79}$$

Since the regression error $\boldsymbol{e}$ is independent of the model m, the minimizers of $\mathrm{CCV}^*(m)$ and $\mathrm{CCV}(m)$ are the same, thus

$$\widehat{m}_{\mathrm{CCV}} = \operatorname*{argmin}_{1\le m\le M} \mathrm{CCV}^*(m). \tag{80}$$

Definition (15) and (80) imply that

$$\mathrm{MSE}(\widehat{m}_{\mathrm{CCV}}) \ge \mathrm{MSE}(m_{\mathrm{opt}}) \tag{81}$$

$$\mathrm{CCV}^*(\widehat{m}_{\mathrm{CCV}}) \le \mathrm{CCV}^*(m_{\mathrm{opt}}). \tag{82}$$

By construction, $\mathrm{MSE}(m) > 0$. Hence

$$\begin{aligned}
0 &\le \frac{\mathrm{MSE}(\widehat{m}_{\mathrm{CCV}}) - \mathrm{MSE}(m_{\mathrm{opt}})}{\mathrm{MSE}(\widehat{m}_{\mathrm{CCV}})} \\
&\le \left(\frac{\mathrm{CCV}^*(m_{\mathrm{opt}}) - \mathrm{MSE}(m_{\mathrm{opt}})}{\mathrm{MSE}(\widehat{m}_{\mathrm{CCV}})}\right) - \left(\frac{\mathrm{CCV}^*(\widehat{m}_{\mathrm{CCV}}) - \mathrm{MSE}(\widehat{m}_{\mathrm{CCV}})}{\mathrm{MSE}(\widehat{m}_{\mathrm{CCV}})}\right) \\
&\le \left|\frac{\mathrm{CCV}^*(m_{\mathrm{opt}}) - \mathrm{MSE}(m_{\mathrm{opt}})}{\mathrm{MSE}(\widehat{m}_{\mathrm{CCV}})}\right| + \left|\frac{\mathrm{CCV}^*(\widehat{m}_{\mathrm{CCV}}) - \mathrm{MSE}(\widehat{m}_{\mathrm{CCV}})}{\mathrm{MSE}(\widehat{m}_{\mathrm{CCV}})}\right| \\
&\le \left|\frac{\mathrm{CCV}^*(m_{\mathrm{opt}}) - \mathrm{MSE}(m_{\mathrm{opt}})}{\mathrm{MSE}(m_{\mathrm{opt}})}\right| + \left|\frac{\mathrm{CCV}^*(\widehat{m}_{\mathrm{CCV}}) - \mathrm{MSE}(\widehat{m}_{\mathrm{CCV}})}{\mathrm{MSE}(\widehat{m}_{\mathrm{CCV}})}\right| \\
&\le 2 \max_{1\le m\le M} \left|\frac{\mathrm{CCV}^*(m) - \mathrm{MSE}(m)}{\mathrm{MSE}(m)}\right| \\
&\xrightarrow[p]{} 0
\end{aligned} \tag{83}$$

as $G \to \infty$. The first and fourth inequalities use (81). The second inequality uses (82). The final convergence is (20). (83) shows that

$$\frac{\mathrm{MSE}(m_{\mathrm{opt}})}{\mathrm{MSE}(\widehat{m}_{\mathrm{CCV}})} \xrightarrow[p]{} 1.$$

Theorem 3 follows by the continuous mapping theorem. ■

11.6 Proof of Theorem 4

Similarly to (55),

$$\frac{\mathrm{CV}(m) - \mathrm{C}(m) - \boldsymbol{e}'\boldsymbol{e}}{\mathrm{MSE}(m)} = S_1(m) - 2S_2(m) + S_3(m) - 2S_4(m) + S_5(m) + 2S_6(m) + 2S_7(m) \tag{84}$$

where

$$S_1(m) = \frac{\boldsymbol{\mu}'\left(\boldsymbol{I}_n - \bar{\boldsymbol{P}}(m)\right)'\left(\boldsymbol{I}_n - \bar{\boldsymbol{P}}(m)\right)\boldsymbol{\mu} - \boldsymbol{\mu}'\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\boldsymbol{\mu}}{\mathrm{MSE}(m)} \tag{85}$$

$$S_2(m) = \frac{\mathrm{tr}\left(\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\bar{\boldsymbol{D}}(m)\boldsymbol{P}(m)\boldsymbol{\Sigma}\right)}{\mathrm{MSE}(m)} \tag{86}$$

$$S_3(m) = \frac{\mathrm{tr}\left(\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\bar{\boldsymbol{D}}(m)\bar{\boldsymbol{D}}(m)\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\boldsymbol{\Sigma}\right)}{\mathrm{MSE}(m)} \tag{87}$$

$$S_4(m) = \frac{\boldsymbol{e}'\bar{\boldsymbol{P}}(m)\boldsymbol{e} - \mathrm{tr}\left(\bar{\boldsymbol{P}}(m)\boldsymbol{\Sigma}\right)}{\mathrm{MSE}(m)} \tag{88}$$

$$S_5(m) = \frac{\boldsymbol{e}'\bar{\boldsymbol{P}}(m)'\bar{\boldsymbol{P}}(m)\boldsymbol{e} - \mathrm{tr}\left(\bar{\boldsymbol{P}}(m)'\bar{\boldsymbol{P}}(m)\boldsymbol{\Sigma}\right)}{\mathrm{MSE}(m)} \tag{89}$$

$$S_6(m) = \frac{\boldsymbol{\mu}'\left(\boldsymbol{I}_n - \bar{\boldsymbol{P}}(m)\right)'\left(\boldsymbol{I}_n - \bar{\boldsymbol{P}}(m)\right)\boldsymbol{e}}{\mathrm{MSE}(m)} \tag{90}$$

$$S_7(m) = \frac{\mathrm{tr}\left(\boldsymbol{P}(m)\boldsymbol{\Sigma}\right) - \mathrm{tr}\left(\bar{\boldsymbol{P}}(m)\boldsymbol{\Sigma}\right) - \mathrm{tr}\left(\boldsymbol{P}(m)\boldsymbol{\Sigma}_0\right)}{\mathrm{MSE}(m)}. \tag{91}$$

We now show that (85)-(91) are each $o_p(1)$ uniformly in m. The arguments for (85)-(90) are similar to those for (56)-(61)

The proof that (85) is $o_p(1)$ is identical to (67), with $\bar{\boldsymbol{D}}(m)$ replacing $\widetilde{\boldsymbol{D}}(m)$ and using (46).

The proof that (86) is $o_p(1)$ is identical to (68), with $\bar{\boldsymbol{D}}(m)$ replacing $\widetilde{\boldsymbol{D}}(m)$, using (46) and (48).

Analogously to (69), using the fact (Theorem 4) that the diagonal elements of $\bar{\boldsymbol{D}}(m)\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)$ equal those of $\boldsymbol{P}(m)$, and (46),

$$\mathrm{tr}\left(\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\bar{\boldsymbol{D}}(m)\bar{\boldsymbol{D}}(m)\left(\boldsymbol{I}_n - \boldsymbol{P}(m)\right)\right) \leq \frac{\mathrm{MSE}(m)}{\underline{\lambda}}o_p(1). \tag{92}$$

This implies that (87) is $o_p(1)$.

Similarly to (71), using (41) and (92),

$$\frac{\mathrm{tr}\left(\bar{\boldsymbol{P}}(m)'\bar{\boldsymbol{P}}(m)\right)}{\mathrm{MSE}(m)} \leq \frac{1}{\underline{\lambda}}\left(1 + o_p(1)\right). \tag{93}$$

Analogously to (73) using (93),

$$\mathbb{E}\left|S_4(m) \mid \boldsymbol{X}\right|^{2r} \leq \frac{C_{2r}DN_n^{2r}\left(1 + o_p(1)\right)}{\underline{\lambda}^r \mathrm{MSE}(m)^r}. \tag{94}$$

By the same calculation as (74) we deduce that (88) is $o_p(1)$. Analogously to (76) using (47) and (93), $S_5(m)$ satisfies the same bound as in (94), and by the same steps as for $T_5(m)$, we find that (89) is $o_p(1)$.

By a similar calculation to (78) we can show that

$$\mathbb{E}\left|S_6(m) \mid \boldsymbol{X}\right|^{2r} \leq \frac{B_{2r}D^{1/2}N_n^{2r}\left(1 + o_p(1)\right)}{\mathrm{MSE}(m)^r}$$

and by the same steps as for $T_6(m)$ we deduce that (90) is $o_p(1)$.

We now take (91). Using (41)

$$S_7(m) = \frac{\operatorname{tr}\left(\bar{\boldsymbol{D}}(m)\boldsymbol{\Sigma}\right) - \operatorname{tr}\left(\boldsymbol{P}(m)\boldsymbol{\Sigma}_0\right)}{\mathrm{MSE}(m)} - \frac{\operatorname{tr}\left(\bar{\boldsymbol{D}}(m)\boldsymbol{P}(m)\boldsymbol{\Sigma}\right)}{\mathrm{MSE}(m)}. \tag{95}$$

Let σ_i^2 be the diagonal elements of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_0$. Observe that $\sigma_i^2 \leq \bar{\lambda}$ by Assumption 1.4. Using (66) and (50)

$$\begin{aligned}
\frac{\operatorname{tr}\left(\bar{\boldsymbol{D}}(m)\boldsymbol{\Sigma}\right) - \operatorname{tr}\left(\boldsymbol{P}(m)\boldsymbol{\Sigma}_0\right)}{\mathrm{MSE}(m)} &= \frac{1}{\mathrm{MSE}(m)}\left(\sum_{i=1}^{n}\frac{P_{ii}(m)}{1-P_{ii}(m)}\sigma_i^2 - \sum_{i=1}^{n}P_{ii}(m)\sigma_i^2\right) \\
&= \frac{1}{\mathrm{MSE}(m)}\sum_{i=1}^{n}\frac{P_{ii}^2(m)}{1-P_{ii}(m)}\sigma_i^2 \\
&\leq \frac{\operatorname{tr}\left(\boldsymbol{P}(m)\right)}{\mathrm{MSE}(m)}\max_{1\leq i\leq n}\frac{P_{ii}(m)}{1-P_{ii}(m)}\bar{\lambda} \\
&\leq o_p(1).
\end{aligned}$$

Using (64), (66), (46), and Assumption 1.4,

$$\frac{\operatorname{tr}\left(\bar{\boldsymbol{D}}(m)\boldsymbol{P}(m)\boldsymbol{\Sigma}\right)}{\mathrm{MSE}(m)} \leq \frac{\operatorname{tr}\left(\boldsymbol{P}(m)\right)\left\|\bar{\boldsymbol{D}}(m)\boldsymbol{\Sigma}\right\|}{\mathrm{MSE}(m)} \leq o_p(1).$$

Thus both components on the right side of (95) are $o_p(1)$ and we conclude that (91) is $o_p(1)$.

We have shown that (85)-(91), and hence (84), are $o_p(1)$ uniformly in m. This completes the proof.

■

11.7 Proof of Theorem 5

The proof is an analog of that of Theorem 3, which is based on Theorems 1 and 2.

The same arguments leading to Theorem 1 can establish that

$$\mathrm{WMSE}(m) = \dot{\boldsymbol{\mu}}'\left(\boldsymbol{I}_n - \dot{\boldsymbol{P}}(m)\right)\dot{\boldsymbol{\mu}} + \operatorname{tr}\left(\dot{\boldsymbol{P}}(m)\dot{\boldsymbol{\Sigma}}\right)$$

where

$$\dot{\boldsymbol{P}}(m) = \dot{\boldsymbol{X}}(m)\left(\dot{\boldsymbol{X}}(m)'\dot{\boldsymbol{X}}(m)\right)^{-1}\dot{\boldsymbol{X}}(m)'$$

and $\dot{\boldsymbol{X}}(m)$ consists of the stacked $\dot{\boldsymbol{X}}_g(m)$ and $\dot{\boldsymbol{\Sigma}} = \operatorname{diag}\left(\dot{\boldsymbol{\Sigma}}_g\right)$ Define $\boldsymbol{W} = \operatorname{diag}\{\boldsymbol{W}_g\}$. Then $\dot{\boldsymbol{X}}(m) = \boldsymbol{W}\boldsymbol{X}(m)$. The matrices $\boldsymbol{W}_g$ and $\boldsymbol{W}$ are idempotent. This implies $\boldsymbol{W}\dot{\boldsymbol{X}}(m) = \boldsymbol{W}\boldsymbol{W}\boldsymbol{X}(m) = \dot{\boldsymbol{X}}(m)$ and

$$\begin{aligned}
\boldsymbol{W}\dot{\boldsymbol{P}}(m)\boldsymbol{W} &= \boldsymbol{W}\dot{\boldsymbol{X}}(m)\left(\dot{\boldsymbol{X}}(m)'\dot{\boldsymbol{X}}(m)\right)^{-1}\dot{\boldsymbol{X}}(m)'\boldsymbol{W} \\
&= \dot{\boldsymbol{X}}(m)\left(\dot{\boldsymbol{X}}(m)'\dot{\boldsymbol{X}}(m)\right)^{-1}\dot{\boldsymbol{X}}(m)' \\
&= \dot{\boldsymbol{P}}(m).
\end{aligned}$$

Since $\dot{\boldsymbol{\Sigma}} = \boldsymbol{W}\dot{\boldsymbol{\Sigma}}\boldsymbol{W}$ we find

$$\operatorname{tr}\left(\dot{\boldsymbol{P}}(m)\dot{\boldsymbol{\Sigma}}\right) = \operatorname{tr}\left(\dot{\boldsymbol{P}}(m)\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}\right) = \operatorname{tr}\left(\boldsymbol{W}\dot{\boldsymbol{P}}(m)\boldsymbol{W}\boldsymbol{\Sigma}\right) = \operatorname{tr}\left(\dot{\boldsymbol{P}}(m)\boldsymbol{\Sigma}\right)$$

and

$$\mathrm{WMSE}(m) = \dot{\boldsymbol{\mu}}' \left(\boldsymbol{I}_n - \dot{\boldsymbol{P}}(m) \right) \dot{\boldsymbol{\mu}} + \mathrm{tr}\left(\dot{\boldsymbol{P}}(m) \boldsymbol{\Sigma} \right). \tag{96}$$

We therefore use (96) in place of Theorem 1.

We next prove an analog of Theorem 2, with WCCV(m) replacing CCV(m) and WMSE(m) replacing MSE(m). Theorem 2 holds unde Assumption 1, so we need to establish that these conditions hold for the within-transformed model under Assumption 2. All these conditions apply with the exception of Assumption 1.3. To see that Assumption 1.2 holds, by Minkowski's inequality and Assumption 2.2

$$\begin{aligned}
\mathbb{E}\left[\left| \dot{e}_{ig} \right|^{4r} \mid \boldsymbol{X} \right] &= \mathbb{E}\left[\left| e_{ig} - \frac{1}{n_g} \sum_{j=1}^{n_g} e_{jg} \right|^{4r} \;\middle|\; \boldsymbol{X} \right] \\
&\leq \left(\left(\mathbb{E}\left[\left| e_{ig} \right|^{4r} \mid \boldsymbol{X} \right] \right)^{1/4r} + \frac{1}{n_g} \sum_{j=1}^{n_g} \left(\mathbb{E}\left[\left| e_{jg} \right|^{4r} \mid \boldsymbol{X} \right] \right)^{1/4r} \right)^{4r} \\
&\leq 2^{4r} D.
\end{aligned}$$

so Assumption 1.2 holds with D replaced with $2^{4r} D$ when applied to (26).

Assumption 1.3, however, does not hold when applied to (26) because $\dot{\boldsymbol{\Sigma}}$ is singular. However, Assumption 1.3 is only used at one point in the proof of Theorem 2, and this is for the proof of equation (66). The needed analog for the within-transformed model is

$$\mathrm{WMSE}(m) \geq k(m) \underline{\lambda}.$$

This however, follows from (96) by the same argument as for (66).

Otherwise, the details of the proofs of Theorems 1-2 carry over to the within-transformed model, establishing the stated result. ■

References

[1] Allen, David M. (1974): "The relationship between variable selection and data augmentation and a method for prediction," *Technometrics*, 16, 125-127.

[2] Andrews, Donald W. K. (1991): "Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors, *Journal of Econometrics*, 47, 359-377.

[3] Arellano, Manuel (1987): "Computing robust standard errors for within groups estimators," *Oxford Bulletin of Economics and Statistics* 49, 431-434.

[4] Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004): "How much should we trust difference-in-differences estimates?" *Quarterly Journal of Economics*, 119, 249-275.

[5] Bell, Robert M., and Daniel F. McCaffrey (2002): "Bias reduction in standard errors for linear regression with multi-stage samples," *Survey Methodology*, 28, 169-181.

[6] Bester, C. Alan, Timothy G. Conley, Christian B. Hansen (2011): "Inference with dependent data using cluster covariance estimators," *Journal of Econometrics*, 165, 137-151.

[7] Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller (2008): "Bootstrap-based improvements for inference with clustered errors," *Review of Economics and Statistics*, 90, 414-427.

[8] Canay, Ivan A., Andres Santos, and Azeem M. Shaikh (2021): "The wild bootstrap with a small number of large clusters," *Review of Economics and Statistics*, 103, 346-363.

[9] Conley, Timothy G. and Christopher R. Taber (2011): "Inference with 'difference in differences' with a small number of policy changes," *Review of Economics and Statistics*, 93, 113-125.

[10] Djogbenou, Antoine. A., James G. MacKinnon, and Morten Ørregaard Nielsen (2019): "Asymptotic theory and wild bootstrap inference with clustered errors," *Journal of Econometrics*, 212, 393-412.

[11] Duflo, Esther, Pascaline Dupas, and Michael Kremer (2011): "Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya," *American Economic Review*, 101, 1739-1774.

[12] Efron, Bradley (1982): *The Jackknife, the Bootstrap, and Other Resampling Plans*, Society for Industrial and Applied Mathematics.

[13] Ferman, Bruno and Cristine Pinto (2019): "Inference in differences-in-differences with few treated groups and heteroskedasticity," *Review of Economics and Statistics*, 101, 452-467.

[14] Geisser, Seymour (1974): "The predictive sample resue method with applications," *Journal of the American Statistical Association*, 70, 320-328.

[15] Hagemann, Andreas (2019): "Placebo inference on treatment effects when the number of clusters is small," *Journal of Econometrics*, 213, 190-209.

[16] Hansen, Bruce E. (2007): "Least squares model averaging," *Econometrica*, 75, 1175-1189.

[17] Hansen, Bruce E. (2014): "Nonparametric Sieve Regression: Least Squares, Averaging Least Squares, and Cross-Validation," *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, Jeffrey S. Racine, Liangjun Su, Aman Ullah, editors, 214-248.

[18] Hansen, Bruce E. (2015): "The integrated mean squared error of series regression and a Rosenthal Hilbert-space inequality," *Econometric Theory*, 31, 337-361.

[19] Hansen, Bruce E. (2022): *Econometrics*, Princeton University Press.

[20] Hansen, Bruce E. (2024): "Jackknife standard errors for clustered regression," working paper.

[21] Hansen, Bruce E. and Seojeong Lee (2019): "Asymptotic theory for clustered samples," *Journal of Econometrics*, 210, 268-290.

[22] Hansen, Bruce E. and Jeffrey S. Racine (2012): "Jackknife model averaging," *Journal of Econometrics*, 167, 38-46.

[23] Hansen, Christian B. (2007): "Asymptotic properties of a robust variance matrix estimator for panel data when T is large," *Journal of Econometrics*, 141, 597-620.

[24] Ibragimov, Rustam and Ulrich K. Müller (2016): "Inference with a few heterogeneous clusters," *Review of Economics and Statistics*, 98, 83-96.

[25] Imbens, Guido W. and Michal Kolesár (2016): "Robust standard errors in small samples: Some practical advice," *Review of Economics and Statistics*, 98, 701-712.

[26] Li, Ker-Chau (1987): "Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set," *The Annals of Statistics*, 15, 958-975.

[27] Liang, Kung-Yee, and Scott L. Zeger (1986): "Longitudinal data analysis using generalized linear models," *Biometrika*, 73, 13-22.

[28] MacKinnon, James G., Morten Ørregaard Nielsen, and Matthew D. Webb (2023a): "Cluster-robust inference: A guide to empirical practice," *Journal of Econometrics*, 232, 272-299.

[29] MacKinnon, James G., Morten Ørregaard Nielsen, and Matthew D. Webb (2023b): "Fast and reliable jackknife and bootstrap methods for cluster-robust inference," *Journal of Applied Econometrics*.

[30] MacKinnon, James G., Morten Ørregaard Nielsen, and Matthew D. Webb (2023c): "Leverage, influence, and the jackknife in clustered regression models: Reliable inference using summclust," *Stata Journal*, 4, 942-982.

[31] MacKinnon, James G., and Matthew D. Webb (2020): "Randomization inference for difference-in-differences with few treated clusters," *Journal of Econometrics*, 218, 435-450.

[32] Moulton, Brent R. (1986): "Random group effects and the precision of regression estimates," *Journal of Econometrics*, 32, 385-397.

[33] Moulton, Brent R. (1990): "An illustration of a pitfall in estimating the effects of aggregate variables on micro units," *Review of Economics and Statistics*, 72, 334-338.

[34] Rust, Keith F. and J. N. K. Rao (1996): "Variance estimation for complex surveys using replication techniques," *Statistical Methods in Medical Research*, 5, 283-310.

[35] Shao, Jun (1997): "An asymptotic theory for linear model selection," *Statistica Sinica*, 7, 221-264.

[36] Shao, Jun and Dongsheng Tu (1995): *The Jackknife and Bootstrap*, Springer.

[37] Stone, Mervyn (1974): "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society, Series B*, 36, 276-278.

[38] Tukey, John (1958): "Bias and confidence in not quite large samples," *Annals of Mathematical Statistics*, 29, 614.

[39] Wahba, Grace and Svante Wold (1975): " A completely automatic French curve: Fitting spline functions by cross-validation," *Communications in Statistics*, 4, 1-17.

[40] Whittle, P. (1960): "Bounds for the moments of linear and quadratic forms in independent variables," *Theory of Probability and Its Applications*, 5, 302-305.

[41] Wooldridge, Jeffrey M. (2003): "Cluster-sample methods in applied econometrics," *American Economic Review*, 93, 133-138.