Projection Spillovers and Exact Robust Variance Estimation

Yanping Chen*

November 14, 2026

Abstract

Conventional heteroskedasticity- and cluster-robust standard errors plug OLS residuals in for the unobserved errors and correct for its own leverage. Each residual, however, also carries the errors of other observations, and this projection spillover biases the classical corrections. For inference on $a'\beta$, we propose a level factor, computed from the design alone, that measures the size of this bias, which need not vanish as the sample grows. We estimate the variance of $a'\hat{\beta}$ directly rather than correct the residuals: since both the variance and the expected residual products are linear in the error covariance, estimation is a linear inverse problem, solved by a matrix-free algorithm. The resulting Riesz variance estimator is exactly unbiased whenever the design identifies the variance, and is consistent and yields an asymptotically normal t-statistic with many controls and growing clusters. It exists where block-inversion corrections (CR2, CR3, and leave-cluster-out) do not. A general theorem extends it to any estimator linear in the errors and any dependency graph, including multi-way clustering and network, spatial, and banded serial dependence. A degrees-of-freedom correction and restricted covariance classes handle designs with few clusters. For weighted and generalized least squares, the within-cluster products of the weighted residuals and the full set of products no longer give the same estimator, and a change of coordinates restores the OLS benchmark. In simulations the Riesz estimator attains nominal coverage where classical estimators do not. Two applications, one with many fixed effects and one with household weights, show it can change published conclusions.

Keywords: robust variance estimation; projection spillovers; cluster-robust inference; many controls; Riesz representation; linear inverse problem; generalized least squares.
JEL codes: C10, C12, C13, C23.

1 Introduction

Inference in linear regression requires the sampling variance of the ordinary least squares (OLS) estimator, which depends on the covariance of the unobserved errors. The heteroskedasticity-consistent (HC) and cluster-robust (CR) variance estimators (Eicker, 1963; White, 1980; Liang and Zeger, 1986; Arellano, 1987) rest on a plug-in principle: the observed OLS residuals stand in for the unobserved errors. Residuals, however, are not true errors. When OLS fits the regression, it also fits part of the errors. The fitted values are $\hat{y} = Hy$, where the hat matrix $H = (h_{ij})$ has the leverages h_{ii} on its

*Department of Economics, Indiana University Bloomington. Email: chenyanp@iu.edu. The identification results used here, including the Gram equivalence of Proposition 7.1, are developed in the companion paper Chen (2026); this paper develops estimation and inference.

diagonal, and the higher an observation's leverage, the more of its error the fit absorbs. Under homoskedasticity, $\mathbb{E}[\hat{u}_i^2] = (1 - h_{ii})\sigma^2$. The finite-sample refinements thus scale the residuals back up. HC1 and CR1 inflate all of them by a degrees-of-freedom factor. HC2 (MacKinnon and White, 1985) divides $\hat{u}_i^2$ by $1 - h_{ii}$, which undoes this shrinkage exactly under homoskedasticity, and CR2 (Bell and McCaffrey, 2002) does the same cluster by cluster. HC3 and CR3 divide $\hat{u}_i$ itself by $1 - h_{ii}$, which turns it into the prediction error it would have if the observation, or its cluster, had been left out of the fit — an approximation to the jackknife, conservative by design. These refinements underpin current recommendations (Imbens and Kolesar, 2016; Pustejovsky and Tipton, 2018; Colin Cameron and Miller, 2015; MacKinnon et al., 2023a).

But the OLS fit absorbs more than the observation's own error. Each residual is $\hat{u}_i = (1 - h_{ii})u_i - \sum_{j \neq i} h_{ij} u_j$, a mix of its own error and those of every other observation; we call the second term the *projection spillover*. The own-leverage corrections account only for the first term, and the spillover biases them. With independent errors, the bias of HC2 for observation i is exactly $h_{ii}(\tilde{\sigma}_i^2 - \sigma_i^2)$ (Section 2), where $\tilde{\sigma}_i^2 = \sum_{j \neq i} h_{ij}^2 \sigma_j^2 / \sum_{j \neq i} h_{ij}^2$ is the average variance of the errors that arrive from other observations, each weighted by how much it contributes to the residual. The leverage h_{ii} is the share of the squared residual that arrives from other observations, and the *contrast* $\tilde{\sigma}_i^2 - \sigma_i^2$ compares the arriving variance with the observation's own. Relative to the error variance, the contrast is scale free: it reflects how unequal the variances are rather than how large, and is bounded for a given degree of heteroskedasticity. The size of each observation's bias is therefore controlled by its leverage. Under homoskedasticity the contrast vanishes and HC2 is exact; otherwise it does not, because a weight built from h_{ii} fixes only how much error arrives but not where it comes from. With clusters, both factors become matrices, and in some designs the cluster corrections CR2 and CR3 cannot even be computed, because the fit absorbs some direction of a cluster's errors entirely.

This paper studies the sampling variance of a linear combination $a'\hat{\beta}$, such as a single coefficient. For such a target the observation-level biases add up, and we summarize their size by a *level factor* κ_n (8): the leverage of each observation or cluster, measured along the regressor of interest and averaged with weights given by its share of that regressor's variation. For a given degree of heteroskedasticity, the level factor sets the size of the relative bias of the classical estimators. It is close to p/n, the number of parameters relative to the sample size, when the regressors vary freely across observations, but otherwise depends on how the regressors are structured (figure 6), and it can stay large as the sample grows. Leading examples are a non-negligible p/n (Anatolyev and Yaskov, 2017; Anatolyev and Smirnov, 2024), regressors that vary mainly across clusters, and a regressor of interest carried by a few observations or clusters; in such designs the spillover bias does not vanish. Many current designs are of this kind: wage regressions with fixed effects for every worker and every firm, identified from the few workers who move between firms (Abowd et al., 1999; Card et al., 2013); selection on observables with long lists of controls (Belloni et al., 2014); and policy evaluations that assign treatment at the state level, often to a handful of states, and cluster by state (Bertrand et al., 2004; Conley and Taber, 2011; Abadie et al., 2023).

The paper makes five contributions. *First, an exact variance estimator* (Sections 4 to 6). Rather than correcting the residuals one by one, we estimate the variance $V = \mathrm{Var}(a'\hat{\beta})$ directly. The key observation is that both sides of the problem are linear in the error covariance: the variance is a linear functional of the covariance, and the expected products of the residuals, within and across clusters,

are a known linear transformation of it. Estimating V from the residuals is therefore a linear inverse problem, and we take its minimum-norm solution. We compute it with LSQR (Paige and Saunders, 1982), a matrix-free algorithm that needs only the action of the linear map and its adjoint, and so scales to large designs. The resulting *Riesz variance estimator* is exactly unbiased, conditional on the design, for every covariance the maintained model admits, whenever the design identifies the variance, that is, whenever every feature of the covariance on which V depends leaves a trace in the expected residual products (Theorem 4.6). This is a condition on the design, which the paper characterizes exactly, and the stopping rule of LSQR verifies it up to its tolerance.

With one-way clustering, the benchmark case, the OLS residuals come from an orthogonal projection, so the within-cluster products carry all the information. The estimator therefore keeps the familiar cluster-robust form, a weighted sum of within-cluster residual products, but chooses the weights for all clusters jointly. It does not rely on the cluster leverage corrections, so it exists in designs where CR2, CR3 and the leave-cluster-out estimator of Anatolyev and Ng (2026) do not. It is consistent under heteroskedasticity, arbitrary within-cluster dependence, growing clusters and many controls, and the t-statistic it produces is asymptotically normal, and so is the one computed from the LSQR iterate when the solver's tolerance shrinks fast enough. In simulations built on the designs of Cattaneo et al. (2018) and Djogbenou et al. (2019) with errors clustered within clusters, the Riesz estimator covers 0.946 with normal critical values and 0.950 with t_ν critical values from its own degrees of freedom, while CR1, CR2 with the Bell–McCaffrey adjustment, CR3 and the leave-cluster-out estimator cover 0.939, 0.942, 0.967 and 0.944 (figure 2 and table 3). With rare-indicator controls, CR2, CR3 and the leave-cluster-out estimator are undefined in 60% of the designs, while the Riesz estimator exists in all of them and covers 0.945 (table 5). In the colonial-governor regressions of Xu (2018) where CR2 and CR3 are undefined, the Riesz estimator changes the conclusions for three of five connection measures (table 10) and for three of the fiscal-performance results (table 11).

Second, a diagnostic of when and by how much the classical corrections fail (Sections 2 and 7). We show that the classical estimators are restricted, approximate inverses of the same system as the Riesz estimator, so their bias comes entirely from inverting it cluster by cluster. The level of that bias is measured by κ_n, which is computed from the design alone: under homoskedasticity CR0 is biased by exactly $-\kappa_n$. The degrees-of-freedom factor of CR1 corrects for p/n instead, which is right only when the regressors vary freely across observations. The simulations confirm that κ_n, not p/n, governs the bias. Across $4,940$ designs, the bias of CR0 has correlation -0.94 with κ_n, against -0.82 with p/n. CR2 removes the leverage but not the heteroskedasticity contrast, so its bias is unrelated to κ_n under homoskedasticity (correlation -0.18) and grows with it under heteroskedasticity (-0.68). CR3 over-corrects in proportion to κ_n (correlation 0.88; figures 7 and 8). In the application to Xu (2018), the change from the CR1 to the Riesz standard error has a correlation of 0.98 with κ_n across the five connection measures (table 10).

Third, a general theory (Section 8). Because the construction uses only linearity, the framework extends well beyond the benchmark. A general theorem covers any estimator linear in the errors, any residual map that removes the regressors, and any dependence described by a dependency graph, including multi-way clustering, network and spatial dependence, and serially correlated errors, and it covers every linear class of maintained covariance structures. Under the same identification condition the estimator is exactly unbiased, and it is consistent as long as the residuals carry enough

information relative to the covariance features to be estimated, that is, as long as the dependence neighbourhoods are small relative to the sample and the inverse problem is well conditioned enough that its weights do not amplify the sampling noise in the residual products. Whether the t-statistic is approximately normal is a separate question: it is in the leading cases, but not in multi-way designs whose components are degenerate (Menzel, 2021), where the variance estimator remains exact but normal critical values do not apply.

Fourth, inference when the variance rests on little information (Sections 5.2 and 8.3). With few clusters, concentrated leverage or two-way dependence, the Riesz estimator remains unbiased but is imprecise, so the t-statistic inherits its noise and is heavier-tailed than normal. A Satterthwaite degrees-of-freedom correction computed from the estimator's own representer, as in the Bell–McCaffrey adjustment of CR2, accounts for this noise; estimators exact on declared covariance classes reduce the noise, trading robustness for precision, and keep close to nominal coverage.

Fifth, weighted and generalized least squares (Section 9). The weighted fit is not an orthogonal projection, so the within-cluster products of the weighted residuals need not carry all the information: the restricted system (Riesz-P(T)) can differ from the full system (Riesz-F(T)). With special weights that couple clusters, the restriction can lose information, and the full system can be unstable, so neither dominates. The remedy is a change of coordinates. The weighted and OLS residuals determine each other, so nothing is lost by moving to coordinates in which the fit is again an orthogonal projection, where the benchmark applies. We propose two such variants. As OLS and GLS are both unbiased for β but differ in precision, both are exactly unbiased whatever the covariance, and the coordinates affect only their precision. Riesz-M works in the OLS residuals and is the most precise when the errors are close to homoskedastic and independent; Riesz-Υ, for weights that respect the cluster partition, works in the regression transformed by the square root of the weights and is the most precise when the weights model the covariance. These properties are proved; how the variants compare outside these two cases is simulation evidence. With t critical values from their own degrees of freedom, Riesz-M and Riesz-Υ cover between 0.931 and 0.969, against 0.724 to 0.897 for the cluster-robust sandwich computed from the weighted residuals (table 8). In the county regressions of Mian and Sufi (2014), household weights raise κ_n to about 0.5 although $p/n < 0.03$, and Riesz-M roughly doubles the published standard errors in six of the seven regressions (table 12).

Related literature. The paper connects six literatures. The first estimates variances from residuals by unbiased quadratic forms. Rao (1970, 1971) and Horn et al. (1975) construct estimators unbiased for heteroskedastic variances and variance components, and Hinkley (1977) and Chesher and Jewitt (1987) study the bias of the robust estimator. The Riesz estimator is a minimum-norm unbiased quadratic estimator in this tradition. It targets the variance of a contrast rather than the covariance itself, and its identification is a condition on the design.

The second corrects heteroskedasticity-robust inference for many regressors. Cattaneo et al. (2018) solve exactly for each observation's error variance from the expected squared residuals, and Kline et al. (2020), Jochmans (2022) and Anatolyev and Ng (2026) use leave-out constructions, the last for clustered designs; related work studies fixed-effects panels (Stock and Watson, 2008; Verdier, 2020) and the entries of projection matrices (Anatolyev and Yaskov, 2017; Anatolyev and Smirnov, 2024). The Riesz estimator is the exact inverse in the scalar case and extends it to clusters without

requiring the cluster leverage corrections to exist.

The third measures leverage and influence (Hoaglin and Welsch, 1978; Cook and Weisberg, 1982; Chatterjee and Hadi, 1988), which make robust standard errors understate uncertainty in practice (Young, 2019). With clusters, Carter et al. (2017) and MacKinnon et al. (2023a) propose the effective number of clusters and partial leverages as warnings. The level factor κ_n is a block leverage in the direction of the contrast, and it sets the level of the classical bias exactly rather than warning of it.

The fourth treats dependence beyond one-way clustering: multiway clustering (Cameron et al., 2011; Thompson, 2011; Davezies et al., 2021; MacKinnon et al., 2021; Menzel, 2021; Chiang et al., 2024) and spatial and network dependence (Conley, 1999; Kojevnikov et al., 2021). The general theorem covers these through a dependency graph.

The fifth concerns small-sample and few-cluster inference with clustered data: the Satterthwaite and Bell–McCaffrey adjustments (Satterthwaite, 1946; Bell and McCaffrey, 2002; Imbens and Kolesar, 2016; Pustejovsky and Tipton, 2018), the wild cluster bootstrap, jackknife and randomization inference (Cameron et al., 2008; MacKinnon and Webb, 2017, 2020; MacKinnon et al., 2023b), and asymptotics with few or heterogeneous clusters (Hansen, 2007a; Bester et al., 2011; Carter et al., 2017; Djogbenou et al., 2019). Our degrees-of-freedom correction and restricted classes address the same problem for an estimator that is exactly unbiased.

The sixth concerns weights. Weighted and generalized least squares are used to correct for sampling designs, to gain efficiency, or under a working covariance model (DuMouchel and Duncan, 1983; Liang and Zeger, 1986; Solon et al., 2015; Romano and Wolf, 2017), with cluster-robust standard errors computed from the weighted residuals. We show that the within-cluster products of these residuals can lose information that the OLS residuals retain, though only with special weights that couple clusters.

The rest of the paper is organized as follows. Section 2 explains the geometry of projection spillovers informally. Section 3 introduces the model and variance functional, and Section 4 the operator framework. Section 5 presents the solver and the degrees-of-freedom correction, and Section 6 the asymptotic results. Section 7 prices the classical restrictions, makes the diagnosis of Section 2 exact, and gives practical guidance. Section 8 states the general theorem and its extensions to dependence that does not partition the sample, including two-way clustering, and to restricted covariance classes. Section 9 treats weighted and generalized least squares. Section 10 reports simulations, Section 11 two empirical applications, and Section 12 concludes.

2 Motivation

This section develops the geometry behind the failure of classical variance estimators and the worth of the cross-residual moments which are typically discarded. The failure is rooted in bias: residualization contaminates each observation's residual with other observations' error terms, and correcting for it requires knowing *whose* errors arrive in what amounts, not merely the total. The precise statements are in Section 7. Under ordinary least squares (OLS), the cross-residual moments provide no extra information beyond the diagonal moments; but under weighted least squares (WLS), they can matter, for precision and, with special weights, for identification, which motivates the estimators of Section 9.

OLS splits $\mathbb{R}^n$ into two orthogonal subspaces: the column space $\mathrm{col}(X)$ of dimension p, onto which $H = X(X'X)^{-1}X'$ projects, and its orthogonal complement $\mathrm{col}(X)^\perp$ of dimension $n-p$, onto which $M = I - H$ projects. Fitted values $\hat{y} = Hy$ live entirely in $\mathrm{col}(X)$; residuals $\widehat{U} = Mu$ live entirely in $\mathrm{col}(X)^\perp$. The true errors u are confined to neither: the component Hu lies in $\mathrm{col}(X)$, drives $\hat{\beta} - \beta = (X'X)^{-1}X'u$, and is annihilated by M, leaving no trace in $\widehat{U}$. The sampling uncertainty in $\hat{\beta}$ therefore originates in $\mathrm{col}(X)$ while the only observable evidence lives in $\mathrm{col}(X)^\perp$, and inference requires using the observed component $\widehat{U}$ to learn about the unobserved one $\mathrm{Var}(\hat{\beta})$.

With $h_{ij} = [X(X'X)^{-1}X']_{ij}$ and $M_{ij} = \mathbf{1}_{i=j} - h_{ij}$, the residual satisfies

$$\hat{u}_i \;=\; M_{ii}\,u_i \;+\; \sum_{j\neq i} M_{ij}\,u_j, \tag{1}$$

so $\hat{u}_i$ both shrinks u_i — the more observation i pulls the fit toward itself, the less of its own error survives — and imports the errors of the observations the regression cannot cleanly separate from it. Under heteroskedasticity with $\mathrm{Var}(u_k) = \sigma_k^2$, the exact finite-sample moments are

$$\mathbb{E}[\hat{u}_i^2] = M_{ii}^2\,\sigma_i^2 \;+\; \underbrace{\sum_{j\neq i} M_{ij}^2\,\sigma_j^2}_{\text{projection spillover}}, \tag{2}$$

$$\mathbb{E}[\hat{u}_i\hat{u}_j] = \sum_{k=1}^{n} M_{ik}M_{jk}\,\sigma_k^2, \qquad i \neq j. \tag{3}$$

The spillover term weights the other variances $\{\sigma_j^2\}_{j\neq i}$, not σ_i^2; this is what makes the problem one of allocation rather than of scale.

2.1 The projection spillover

Because M is symmetric and idempotent, $M_{ii} = \sum_j M_{ij}^2$, so the total off-diagonal mass arriving at observation i is a function of its leverage alone:

$$\sum_{j\neq i} M_{ij}^2 \;=\; M_{ii} - M_{ii}^2 \;=\; h_{ii}(1-h_{ii}). \tag{4}$$

Normalizing, $w_{ij} := M_{ij}^2 / \sum_{k\neq i} M_{ik}^2$ is a weighting of the other observations, and $\tilde{\sigma}_i^2 := \sum_{j\neq i} w_{ij}\,\sigma_j^2$ is the average variance of the errors that arrive at i. Substituting into (2) and dividing by M_{ii} gives the bias of HC2 in closed form:

$$\mathbb{E}\left[\frac{\hat{u}_i^2}{1-h_{ii}}\right] - \sigma_i^2 \;=\; h_{ii}\left(\tilde{\sigma}_i^2 - \sigma_i^2\right). \tag{5}$$

The bias factors into a *leverage* term and a *contrast* term.[1] The leverage h_{ii} is the share of $\hat{u}_i^2$ that arrives from other observations; and the contrast $\tilde{\sigma}_i^2 - \sigma_i^2$ measures how much those errors differ in variance from i's own. Both factors must be non-negligible for the bias to matter: under homoskedas-

[1] The scaling by M_{ii} is canonical rather than a choice of estimator. For $\hat{u}_i^2/(1-h_{ii})^\delta$ the coefficients on $(\sigma_i^2, \tilde{\sigma}_i^2)$ sum to $(1-h_{ii})^{1-\delta}$, so $\delta = 1$ is the unique case in which they sum to one and hence the unique case in which the bias vanishes under homoskedasticity. At $\delta = 2$ (HC3) it is $h_{ii}\tilde{\sigma}_i^2/(1-h_{ii})$, positive by construction.

ticity the contrast is zero and HC2 is exactly unbiased, which is the classical result.

Equation (5) also isolates what a leverage correction can and cannot do. The scalar h_{ii} determines how much error arrives but not from whom; and the contrast relies on both the allocation $\{w_{ij}\}$ and the variances $\{\sigma_j^2\}$. No weight computed from i's own leverage can remove the bias term.

For inference on a linear combination $a'\beta$, such as a single coefficient, what matters is the bias relative to the variance. With $\Gamma = n^{-1}\sum_i x_i x_i'$ and $V = n^{-1}\sum_i x_i x_i' \sigma_i^2$, the variance of $a'\hat{\beta}$ is $n^{-1}c'Vc$, where $c := \Gamma^{-1}a$ is the direction that a picks out in the space of regressors. The relative bias of HC2 for it satisfies

$$\frac{|c'Bc|}{c'Vc} \;\le\; \frac{\max_i \left|\tilde{\sigma}_i^2 - \sigma_i^2\right|}{\min_i \sigma_i^2} \sum_{i=1}^{n} s_i h_{ii}, \qquad s_i = \frac{(c'x_i)^2}{\sum_j (c'x_j)^2}, \tag{6}$$

where $B = n^{-1}\sum_i x_i x_i' h_{ii}(\tilde{\sigma}_i^2 - \sigma_i^2)$. The first factor is scale free: it measures how unequal the variances are, not how large. The second is an average of leverage weighted by how much each observation contributes to V. The bias is therefore small only if the observations that drive V have low leverage, and this fails in two broad regimes.

The first is when there are many parameters relative to the sample, so that $p/n \to \alpha \in (0,1)$. Average leverage then stays at α rather than vanishing, and leverage is non-negligible throughout the sample, so $\sum_i s_i h_{ii}$ does not go to zero however the weights are spread. Three settings fall in this regime. With many controls or saturated interactions, Cattaneo et al. (2018) show that the usual robust estimator is inconsistent and propose a corrected one. With many instruments, the problem is studied by Verdier (2020); empirically, this setting includes judge or examiner designs that use many assignment indicators as instruments (Kling, 2006; Dobbie et al., 2018). With fixed effects estimated from few observations each, Stock and Watson (2008) show that the robust estimator is inconsistent in panels with few periods, and Kline et al. (2020) develop leave-out estimators for models with many regressors, with two-way fixed effects as the leading case; empirical examples include worker and firm fixed-effects models of wages (Abowd et al., 1999; Card et al., 2013) and teacher value-added models (Chetty et al., 2014).

The second is when the direction c is informed by only a few observations, even if p/n is small. Those observations dominate V, and since $h_{ii} \ge s_i$ they must also have high leverage: the weighted leverage is at least $\sum_i s_i^2$, the inverse of the effective number of observations behind c. A treatment given to T units gives each treated unit $s_i = 1/T$, so with five or ten treated units the bias is first order even in a sample of thousands. Imbens and Kolesar (2016) illustrate the problem with unbalanced binary regressors, and Young (2019) finds that published randomized experiments often have a few highly influential observations, which makes robust standard errors understate uncertainty.

Clustered dependence. With clusters, the same decomposition holds block by block:

$$\mathbb{E}[\widehat{U}_g \widehat{U}_g' \mid X] \;=\; M_{gg}\Sigma_g M_{gg} \;+\; \underbrace{\sum_{k \neq g} M_{gk}\Sigma_k M_{kg}}_{\text{cross-cluster spillover}}, \tag{7}$$

with the same two parts: what survives of cluster g's own errors and what arrives from other clusters. The block counterparts of $1 - h_{ii}$ and h_{ii} are M_{gg} and $H_{gg} = I - M_{gg}$, and the off-block mass reaching cluster g is $\sum_{k \neq g} M_{gk}M_{kg} = M_{gg}H_{gg}$. CR2 and CR3 normalize $\widehat{U}_g\widehat{U}_g'$ by $M_{gg}^{-1/2}$ and M_{gg}^{-1}, treating

it as informative about Σ_g alone. Normalized as in (5), CR2 is left with $H_{gg}^{1/2}(\tilde{\Sigma}_g - \Sigma_g)H_{gg}^{1/2} + R_g$ (Section 7.3), where $\tilde{\Sigma}_g$ is an average of the neighbouring $\{\Sigma_k\}$ and R_g a remainder: again a mass times a contrast, but now both are matrices, which changes the problem in three ways.

The mass is a matrix, and its size is κ_n. H_{gg} weights the directions within a cluster unequally. When the regressors vary mostly across clusters, it is concentrated on the cluster mean, which the regression fits closely, and nearly zero on the deviations from it, so no single number per cluster describes the mass. For the direction c of (6), its size in the directions the regressor takes is

$$\kappa_n := \frac{\sum_{g=1}^{G}(X_g c)' H_{gg} (X_g c)}{\sum_{g=1}^{G} \|X_g c\|^2}. \tag{8}$$

Writing $S_g = \|X_g c\|^2 / \sum_k \|X_k c\|^2$ for cluster g's share of the variation in direction c and $\rho_g = (X_g c)' H_{gg} (X_g c) / \|X_g c\|^2$ for its leverage in that direction, $\kappa_n = \sum_g S_g \rho_g$ is the counterpart of $\sum_i s_i h_{ii}$, to which it reduces when $N_g = 1$. Under homoskedasticity the relative bias of CR0 is exactly $-\kappa_n$ (Proposition 7.3). Figure 1 shows the bias of CR0 across the simulated designs scattered against p/n and collapsing onto κ_n.

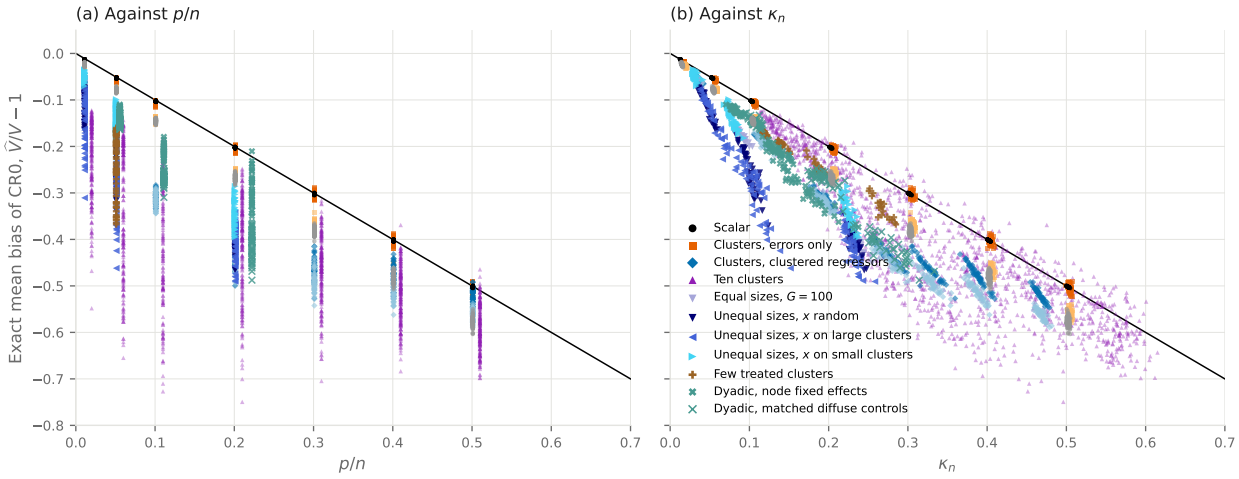


Figure 1: The level of the classical bias. Exact mean bias of CR0 (HC0 when $N_g = 1$), against p/n (left) and against the block leverage κ_n of (8) (right), in simulated designs with scalar, clustered, few-cluster, few-treated-cluster and dyadic structure (Section 10, Section 10.2). The line is the bias under $\Sigma = \sigma^2 I_n$: exact on the right for every design (Proposition 7.3), on the left only in the scalar case. Departures from the line are the contrast term: heteroskedasticity aligned with the regressor ($\vartheta = 1$) and the within-cluster random effect ($\rho_u = 0.3$) push the bias mostly below $-\kappa_n$. For the scalar and Exhibit 1 cluster designs, dark markers are homoskedastic ($\vartheta = 0$) and light markers heteroskedastic ($\vartheta = 1$).

As in the scalar case, $\rho_g \geq S_g$,[2] the block analogue of $h_{ii} \geq s_i$, so $\kappa_n \geq \sum_g S_g^2 \geq 1/G$. What differs is the size of this floor. The scalar floor $1/n$ is negligible, but $1/G$ is not: the variance estimator is built from G cluster scores rather than n residuals, and the fit forces the scores in direction c to sum to zero, absorbing about one cluster's worth of variation whatever the cluster sizes. The two regimes of the scalar case carry over, with clusters in place of observations.

The first is when many parameters compete for the same variation. Regressors that vary only

[2]Let $v := Xc$, and let $\tilde{v}_g$ agree with v on cluster g and be zero elsewhere. Since $v \in \mathrm{col}(X)$, $Hv = v$, so by Cauchy–Schwarz $\|X_g c\|^2 = \langle \tilde{v}_g, Hv \rangle = \langle H\tilde{v}_g, v \rangle \leq \|H\tilde{v}_g\| \, \|v\|$. As H is a symmetric projection, $\|H\tilde{v}_g\|^2 = (X_g c)' H_{gg} (X_g c) = \rho_g \|X_g c\|^2$, and therefore $\rho_g \geq \|X_g c\|^2 / \|v\|^2 = S_g$.

within clusters add about p/n, so κ_n stays bounded away from zero when $p/n \to \alpha \in (0,1)$. Parameters identified from variation across clusters cost more, about $1/G$ each, because they compete for the G cluster scores rather than the n residuals. When $X_g c$ is constant within clusters, ρ_g is the leverage of cluster g's mean. These leverages sum to at most p, with equality when all regressors are constant within clusters, so κ_n is then about p/G (Proposition 7.3(iii)) and is large when $p/G \to \alpha \in (0,1)$. Fixed effects shared by several clusters fall in this regime. In gravity models of trade clustered by country pair, exporter and importer effects are constant within each pair but common to all pairs involving the same country (Anderson and Van Wincoop, 2003). Each is estimated from the roughly C pairs involving that country, so κ_n is of order $2/C$ rather than $1/G \approx 1/C^2$, and the bias is non-negligible when few countries are observed. The same holds for node effects in other dyadic and network data.

The second is when few effective clusters inform the direction c. The effective number $G^{\text{eff}} = 1/\sum_g S_g^2$ equals G only when every cluster carries the same share of the variation, and falls below it whenever the shares are uneven. Since $\|X_g c\|^2$ is the size of cluster g times the average of $(c'x_i)^2$ within it, the shares depend both on how large the clusters are and on how much the regressor moves in each, so G^{eff} can be small for three reasons. The first is that G itself is small, whether or not the regressor varies within clusters: Bell and McCaffrey (2002) introduce CR2 to correct the resulting bias, and Cameron et al. (2008), Imbens and Kolesar (2016) and Pustejovsky and Tipton (2018) show that inference based on CR0 or CR1 over-rejects with few clusters. The second is that the regressor varies in only a few clusters. When it is carried by T clusters, $\kappa_n \geq 1/T$ however large G is, as in difference-in-differences designs with one or a handful of policy changes (Conley and Taber, 2011; MacKinnon and Webb, 2018; Ferman and Pinto, 2019), such as studies of a single state's minimum wage increase or a single city's immigration shock (Card and Krueger, 2000; Card, 1990). The third is that clusters differ greatly in size or in how much of the regressor they carry, so that a few large clusters, or clusters with extreme values of the regressor, dominate $\sum_g \|X_g c\|^2$ (MacKinnon and Webb, 2017; Carter et al., 2017).

The contrast reweights and rotates. $\tilde{\Sigma}_g$ averages the neighbouring $\{\Sigma_k\}$, so the contrast is nonzero when covariances differ across clusters in a way aligned with the design. But it averages matrices, and the averaging can mix their directions: equal Σ_k no longer give $\tilde{\Sigma}_g = \Sigma_g$. The remainder R_g has no scalar counterpart. It is nonzero when Σ_g does not share the eigenvectors of H_{gg}, as with serial correlation within a cluster while the regression absorbs the cluster mean, and so it biases CR2 even when every cluster has the same covariance matrix. With $\Sigma_k = \sigma^2 I$ both vanish, and CR2 is exact however much mass crosses, as HC2 is. Section 7.3 gives the exact decomposition.

The correction may not exist. CR2 and CR3 require M_{gg} to be nonsingular, which fails if and only if some $b \in \mathbb{R}^p$ has $X_g b \neq 0$ and $X_k b = 0$ for all $k \neq g$, that is, some vector supported on cluster g lies in $\mathrm{col}(X)$. Cluster fixed effects do this for every cluster, since they put ι_g in $\mathrm{col}(X)$. In Section 11, we revisit Xu (2018), where CR2 and CR3 do not exist.

2.2 The cross-residual moments

Stacking the squared-residual moments (2) over i gives a linear system in $\mathbf{s} = (\sigma_1^2, \ldots, \sigma_n^2)'$,

$$A\mathbf{s} = \mathbf{q}, \qquad A_{ik} = M_{ik}^2, \quad q_i = \mathbb{E}[\hat{u}_i^2], \tag{9}$$

with $A = M \odot M$ the Hadamard square. Solving it undoes the spillover of (5) for all observations at once: instead of correcting each $\hat{u}_i^2$ by its own leverage, it accounts for every error that reaches every residual. Whenever A is nonsingular the system recovers $\mathbf{s}$ exactly, spillover included; this is essentially the many-covariate correction of Cattaneo et al. (2018). The cross-products (3) are discarded, and it is natural to suspect that they carry identifying information: there are about $n^2/2$ of them against n squares, and their coefficients $M_{ik}M_{jk}$ are different from M_{ik}^2.

Under OLS they do not. The cross-products depend on the same variances as the squares, and because M is symmetric and idempotent the squared-residual system is exactly the normal equations of the full system that includes them (Proposition 7.1). Keeping the cross-products therefore buys no identification: solved exactly, the two give the same estimate in every sample (Corollary 7.2). What they change is conditioning, which the restricted system squares (Proposition E.1), just as forming the normal equations of a least-squares problem squares its condition number; that costs iterations, not accuracy.

Under weighted least squares they can. The residual maker of a weighted fit is idempotent but not symmetric — it projects obliquely — so the collapse above fails, and the two systems written in the weighted residuals are no longer one estimator. Both are unbiased when the squares identify the variance, which fails only with special weights that couple observations (Proposition 9.1); which is more precise depends on what the weights are for: the restricted system when the weights model the error covariance, the full system when they do not (Proposition 9.6). The estimators proposed for weighted fits work in coordinates in which the fit is again an orthogonal projection: Riesz-M organizes the cross-products through the OLS projection (Proposition 9.2), and Riesz-Υ uses the regression transformed by the square root of the weights (Proposition 9.4), so the geometry of the OLS case carries over (Section 9).

3 Econometric Model

We consider a clustered linear regression model as benchmark. Let clusters be indexed by $g = 1, \ldots, G$, each containing N_g observations, with total sample size $n = \sum_{g=1}^{G} N_g$. For a generic variable Z, let z_{gi} denote the ith observation in cluster g, Z_g the stacked observations in cluster g, and $Z = (Z_1', \ldots, Z_G')'$ the full sample stacked across clusters. Following Cattaneo et al. (2018), the data satisfy

$$Y_g = X_g\beta + W_g\gamma + U_g, \tag{10}$$

where $Y_g \in \mathbb{R}^{N_g}$ is the outcome vector, $X_g \in \mathbb{R}^{N_g \times k}$ collects the regressors of primary interest, $W_g \in \mathbb{R}^{N_g \times \ell}$ denotes a (possibly high-dimensional) set of control variables, and $U_g \in \mathbb{R}^{N_g}$ represents unobserved disturbances that may exhibit arbitrary dependence within clusters. The parameter of interest is $\beta \in \mathbb{R}^k$, while $\gamma \in \mathbb{R}^\ell$ is treated as a nuisance parameter. The dimension ℓ may increase with the sample size n, accommodating high-dimensional control designs. We assume clusters are independent conditional on (X, W).

This specification encompasses several well-known cases. When $\ell = 0$, it reduces to the classical low-dimensional regression model; when ℓ/n does not vanish, it corresponds to high-dimensional regression settings. The framework includes heteroskedasticity-robust inference as the case $G = n$ and $N_g = 1$, and standard cluster-robust inference when $G \to \infty$ with bounded cluster sizes N_g.

We handle the high-dimensional controls by partialling out W. For a matrix Z with n rows, let Π_Z denote the orthogonal projector onto $\mathrm{col}(Z)$.[3] Let $M_W := I_n - \Pi_W$ denote the residual-maker projecting onto the orthogonal complement of $\mathrm{col}(W)$; W need not have full column rank. Everything below depends on W only through $\mathrm{span}(W)$, and is therefore invariant to rescaling or re-parametrizing the controls. Define $\widetilde{X} = M_W X$ and $\widetilde{Y} = M_W Y$. Applying OLS to the residualized model yields

$$\widehat{\beta} = (X' M_W X)^{-1}(X' M_W Y) = \beta + \big(\widetilde{X}'\widetilde{X}\big)^{-1}\big(\widetilde{X}' U\big), \tag{11}$$

and the residuals

$$\widehat{U} = M_W(Y - X\widehat{\beta}) = MU, \tag{12}$$

where $M := M_{\widetilde{X}} M_W$ with $M_{\widetilde{X}} = I_n - \widetilde{X}(\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'$ is the effective OLS residual-maker. By the Frisch–Waugh–Lovell theorem $M = I_n - \Pi_{[X\ W]}$, so M is the residual-maker of the joint regression and is in particular symmetric and idempotent. All coupling across observations or clusters arises through the projection geometry of M.

To conduct inference, an appropriate normalization for $(\widehat{\beta} - \beta)$ is required, which generally depends on cluster sizes, within-cluster dependence, and regressor variation (Carter et al., 2017; Hansen and Lee, 2019; Djogbenou et al., 2019). Let μ_n denote the effective scaling of the estimator, with $\mu_n = n$ under heteroskedastic sampling and μ_n equal to G, n, or an intermediate rate under clustered dependence. To avoid estimating μ_n directly, we focus on the scalar t-statistic

$$t = \frac{\sqrt{\mu_n}\, a'(\widehat{\beta} - \beta)}{\sqrt{\mu_n \widehat{V}}} = \widehat{V}^{-1/2} a'(\widehat{\beta} - \beta), \tag{13}$$

where $a \in \mathbb{R}^k$ is a fixed unit vector and $\widehat{V}$ is a variance estimator satisfying

$$\widehat{V} = V + o_p(\mu_n^{-1}), \tag{14}$$

with

$$V = a'\Gamma^{-1}\left(\frac{1}{n^2}\sum_{g=1}^{G} \widetilde{X}_g' \Sigma_g \widetilde{X}_g\right)\Gamma^{-1} a, \quad \Gamma = \frac{1}{n}\sum_{g=1}^{G} \widetilde{X}_g'\widetilde{X}_g, \quad \Sigma_g = \mathbb{E}[U_g U_g' \mid X, W]. \tag{15}$$

Under suitable regularity conditions, $t \xrightarrow{d} \mathcal{N}(0,1)$.

In practice, variance estimation is challenging because the covariance blocks $\{\Sigma_g\}$ are unknown and the errors $\{U_g\}$ are unobserved, so V in (15) cannot be computed directly. By (12),

$$\widehat{U}_g = [MU]_g = \sum_{h=1}^{G} M_{gh} U_h = M_{gg} U_g + \sum_{h \neq g} M_{gh} U_h,$$

where M_{gh} denotes the (g,h)-th block of M conformable with the cluster structure. Classical HC and CR variance estimators correct only through the diagonal blocks $\{M_{gg}\}$ and so treat $\widehat{U}_g \widehat{U}_g'$ as informative about Σ_g alone. As Section 2 shows, this fails in three ways: the spillover has a level, κ_n of (8), which biases CR0 even under homoskedasticity (Proposition 7.3); it has a contrast between Σ_g

[3] Explicitly, $\Pi_Z = Z(Z'Z)^{+}Z'$, where $(\cdot)^{+}$ is the Moore–Penrose inverse; any generalized inverse of $Z'Z$ gives the same matrix, so Z need not have full column rank.

and the covariance blocks arriving from other clusters, which biases CR2 and which equal blocks no longer remove; and when M_{gg} is singular, CR2 and CR3 do not exist (Proposition 7.4). The level κ_n is close to p/n when the regressors vary within clusters, to p/G when they vary only across clusters, and to $1/T$ when the contrast is carried by T clusters, so first-order spillover is not exotic.

Rather than attempting to estimate the latent covariance blocks $\{\Sigma_g\}$ or to reconstruct the unobserved errors $\{U_g\}$, the approach developed in this paper targets the variance functional V directly, exploiting the structure of the full residual moment system implied by the projection geometry. For a scalar contrast, the variance admits the linear representation

$$V = \frac{1}{n^2}\sum_{g=1}^{G} \operatorname{tr}\big(\Sigma_g \widetilde{X}_g Q \widetilde{X}_g'\big), \qquad Q := \Gamma^{-1} a a' \Gamma^{-1}. \tag{16}$$

This representation rewrites V as a quantity that is *linear* in the unknown covariance array $\{\Sigma_g\}$, rather than as the quadratic form in (15). Once $\{\Sigma_g\}$ is equipped with a suitable inner product, linearity is precisely what permits V to admit a Riesz representer, which we construct in Section 4.

Although $\{\Sigma_g\}$ is unobserved, the feasible residual moments

$$\widehat{C}_{hh'} := \widehat{U}_h \widehat{U}_{h'}'$$

have population counterparts. For cluster indices $h, h' \in \{1, \dots, G\}$,

$$C_{hh'} = \mathbb{E}[\widehat{C}_{hh'} \mid X, W] = \sum_{g=1}^{G} M_{hg} \Sigma_g M_{gh'},$$

using cluster independence and symmetry of M. This linear system links the full residual moment array $\{C_{hh'}\}$ — including both within- and cross-cluster moments — to latent covariance blocks $\{\Sigma_g\}$ through the projection geometry of M, motivating the Riesz representation framework developed in the next section. In summary, this section has introduced three objects on which everything that follows depends: the projection geometry of M, governing how errors mix across observations and clusters; the unknown target $\{\Sigma_g\}$, which we never attempt to estimate directly; and the observable link $\{C_{hh'}\}$, the full residual moment array through which $\{\Sigma_g\}$ is related to V.

The development below uses OLS and the block-diagonal covariance space for concreteness, and they remain the leading case throughout. The construction depends on them only through three facts — the estimator is linear in the errors, the residuals are a known linear map of the errors, and the maintained covariance model is a linear space — and Section 8 states and proves it at that level of generality, which covers weighted and generalized least squares and maintained classes smaller than the block-diagonal one.

4 The Riesz Representation of the Variance Functional

Our goal is to recover the scalar variance functional $V = \operatorname{Var}(a'\widehat{\beta})$ directly, without estimating the latent covariance blocks $\{\Sigma_g\}_{g=1}^{G}$. The key observation is that V is a linear functional of $\{\Sigma_g\}$ and therefore admits a Riesz representer under a suitable inner product. We then exploit the linear

operator linking $\{\Sigma_g\}$ to the residual moment array $\{C_{hh'}\}$ to obtain a feasible representation and develop a corresponding estimator via observable sample residual moments $\{\widehat{C}_{hh'}\}$.

Throughout this section, the sample size n and cluster structure $\{N_g\}_{g=1}^G$ are fixed, and we suppress the subscript n on the spaces $\mathcal{H}$ and $\mathcal{K}$, and the objects built from them, restoring it from Section 6 onward when asymptotic behavior as $n \to \infty$ is the object of interest. Proofs for this section are collected in Section B.

4.1 Latent covariance space and variance functional

For a given sample of size n with known cluster structure $\{N_g\}_{g=1}^G$, let $\mathbb{S}^m$ denote the space of real symmetric $m \times m$ matrices and define the finite-dimensional Hilbert space

$$\mathcal{H} := \bigoplus_{g=1}^{G} \mathbb{S}^{N_g}, \qquad \langle A, B\rangle_{\mathcal{H}} := \frac{1}{n^2}\sum_{g=1}^{G} \operatorname{tr}(A_g' B_g), \qquad \|A\|_{\mathcal{H}}^2 := \langle A, A\rangle_{\mathcal{H}}.$$

The covariance array $\Sigma := \{\Sigma_g\}_{g=1}^G$ is an element of $\mathcal{H}$. Restricting to symmetric blocks is not cosmetic: every admissible Σ_g is symmetric, and the restriction shrinks $\operatorname{Ker}(\mathcal{A})$ and hence weakens the range condition imposed in Assumption A5 below. We identify $\Sigma \in \mathcal{H}$ with the block-diagonal matrix $\operatorname{diag}(\Sigma_1, \ldots, \Sigma_G) \in \mathbb{S}^n$ and use the two readings interchangeably. The identification is isometric, $\|\Sigma\|_{\mathcal{H}} = n^{-1}\|\operatorname{diag}(\Sigma_1, \ldots, \Sigma_G)\|_F$, so that expressions such as $M\Sigma M$ are well defined.

Proposition 4.1 (Riesz applicability)**.** *If Γ is nonsingular, as holds with probability approaching one under Assumption A4, the variance functional $\Phi : \mathcal{H} \to \mathbb{R}$ defined by*

$$\Phi(\Sigma) := V = \frac{1}{n^2}\sum_{g=1}^{G} \operatorname{tr}\big(\Sigma_g \widetilde{X}_g Q \widetilde{X}_g'\big)$$

is linear and bounded on $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$. Hence, by the Riesz representation theorem, there exists a unique element $R^\star = \{R_g^\star\} \in \mathcal{H}$ such that

$$\Phi(\Sigma) = \langle \Sigma, R^\star\rangle_{\mathcal{H}} \quad \textit{for all } \Sigma \in \mathcal{H}.$$

Moreover, $R_g^\star = \widetilde{X}_g Q \widetilde{X}_g'$ for all g.

Proposition 4.1 expresses the target variance as an inner product between the latent covariance array and a known representer $R^\star$. To obtain a feasible representation, we next link Σ to observable residual moments via the projection geometry.

4.2 Residual moment space and the projection operator

Let M denote the effective OLS residual-maker from Section 3, so that $\widehat{U} = MU$. Define the space of the full residual moment array

$$\mathcal{K} := \mathbb{S}^n, \qquad \langle C, D\rangle_{\mathcal{K}} := \frac{1}{n^2}\operatorname{tr}(C'D) = \frac{1}{n^2}\sum_{h,h'=1}^{G} \operatorname{tr}(C_{hh'}' D_{hh'}),$$

whose elements we write blockwise under the cluster partition as $C = \{C_{hh'}\}_{h,h'=1}^{G}$ with $C_{hh'} \in \mathbb{R}^{N_h \times N_{h'}}$; symmetry of C means $C_{h'h} = C'_{hh'}$, so the diagonal blocks are symmetric and each off-diagonal pair carries one free block.[4] In particular $\|C\|_{\mathcal{K}} = n^{-1}\|C\|_F$. Under the identification of Section 4.1, $\mathcal{H}$ is the block-diagonal subspace of $\mathcal{K}$ and $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is the restriction of $\langle\cdot,\cdot\rangle_{\mathcal{K}}$; the two are kept distinct in the notation because $\mathcal{A}$ maps one to the other. The population residual moments $C_{hh'} := \mathbb{E}[\widehat{U}_h\widehat{U}'_{h'} \mid X, W]$ and their sample counterparts $\widehat{C}_{hh'} := \widehat{U}_h\widehat{U}'_{h'}$ both lie in $\mathcal{K}$. Under cluster independence the former satisfy the linear system

$$C_{hh'} = \sum_{g=1}^{G} M_{hg}\Sigma_g M_{gh'},$$

which motivates the linear operator $\mathcal{A} : \mathcal{H} \to \mathcal{K}$ defined by

$$[\mathcal{A}(\Sigma)]_{hh'} := \sum_{g=1}^{G} M_{hg}\Sigma_g M_{gh'}, \tag{17}$$

so that $C = \mathcal{A}(\Sigma)$: the operator carries the latent covariance array to the residual moments it induces, and every cross-cluster block of C is generated by the projection geometry. Its null space $\mathcal{N} := \mathrm{Ker}(\mathcal{A}) \subseteq \mathcal{H}$ collects the covariance directions that residualization annihilates, and $\Pi_{\mathcal{N}}$ denotes the orthogonal projector onto it in $\mathcal{H}$; with the subscript n restored, $\mathcal{N}_n := \mathrm{Ker}(\mathcal{A}_n)$.

Lemma 4.2 (Adjoint operator)**.** *The operator $\mathcal{A}$ admits a unique Hilbert-space adjoint $\mathcal{A}^* : \mathcal{K} \to \mathcal{H}$ satisfying*[5]

$$\langle \mathcal{A}(\Sigma), D\rangle_{\mathcal{K}} = \langle \Sigma, \mathcal{A}^*(D)\rangle_{\mathcal{H}} \qquad \textit{for all } \Sigma \in \mathcal{H},\ D \in \mathcal{K},$$

and it has the explicit block form

$$[\mathcal{A}^*(D)]_g = \sum_{h,h'=1}^{G} M_{gh}\, D_{hh'}\, M_{h'g}. \tag{18}$$

Lemma 4.2 allows us to transfer the Riesz representer from the latent-covariance space $\mathcal{H}$ to the residual-moment space $\mathcal{K}$.

4.3 Residual-moment representer and the Riesz variance estimator

We seek weights $D = \{D_{hh'}\} \in \mathcal{K}$ such that the residual-moment inner product $\langle C, D\rangle_{\mathcal{K}}$ reproduces the target variance functional $\Phi(\Sigma)$ for all admissible covariance arrays Σ satisfying $C = \mathcal{A}(\Sigma)$. By Proposition 4.1 and Lemma 4.2,

$$\langle C, D\rangle_{\mathcal{K}} = \langle \mathcal{A}(\Sigma), D\rangle_{\mathcal{K}} = \langle \Sigma, \mathcal{A}^*(D)\rangle_{\mathcal{H}}.$$

[4] Taking symmetric matrices loses nothing: transposition maps the solution set of (19) in $\mathbb{R}^{n\times n}$ onto itself and preserves the Frobenius norm, so the minimum-norm solution over all $n \times n$ matrices is symmetric.

[5] In general, the adjoint of a bounded operator $\mathcal{A}\colon \mathcal{H} \to \mathcal{K}$ between Banach spaces acts on their duals, sending each $\phi \in \mathcal{K}^*$ to $\phi \circ \mathcal{A} \in \mathcal{H}^*$. Here $\mathcal{H}$ and $\mathcal{K}$ are finite-dimensional inner-product spaces, hence Hilbert spaces, and the Riesz representation theorem identifies each isometrically with its own dual through $D \mapsto \langle\cdot, D\rangle$. Through these identifications the adjoint becomes the map $\mathcal{K} \to \mathcal{H}$ below, the transpose of $\mathcal{A}$ under the Frobenius inner products. Finite dimension also suffices for the asymptotics, which concern a sequence of such problems whose growing dimension enters only through inequalities with constants free of n and rate conditions such as Assumption A5(ii).

Hence, any $D \in \mathcal{K}$ satisfying

$$\mathcal{A}^*(D) = R^\star = \{\widetilde{X}_g Q \widetilde{X}_g'\}_{g=1}^G \tag{19}$$

ensures

$$\Phi(\Sigma) = \langle \Sigma, R^\star \rangle_{\mathcal{H}} = \langle C, D \rangle_{\mathcal{K}}.$$

Equation (19) is the central identification equation of the paper: it characterizes all residual-moment weighting schemes that exactly recover the target variance functional. Classical HC and CR estimators correspond to particular restricted solutions to this equation (see Section 7.2).

Because the residual-moment space $\mathcal{K}$ collects all within- and cross-cluster residual moments, it has higher dimension than the latent covariance space $\mathcal{H}$, the operator $\mathcal{A} : \mathcal{H} \to \mathcal{K}$ is generically non-square and tall. Its adjoint $\mathcal{A}^* : \mathcal{K} \to \mathcal{H}$ is therefore fat, so the adjoint equation (19) need not admit a unique solution. Moreover, the projection geometry encoded in $\{M_{gh}\}$ may induce linear dependencies among the residual moment equations, so $\mathcal{A}$ need not have full column rank. If $\mathrm{Ker}(\mathcal{A}^*) \neq \{0\}$, the solution set of (19) is an affine subspace: for any $Z \in \mathrm{Ker}(\mathcal{A}^*)$, we have $\langle C, Z \rangle_{\mathcal{K}} = 0$ for all $C \in \mathrm{range}(\mathcal{A})$, so that $D + Z$ induces the same functional on $\mathrm{range}(\mathcal{A})$.

The variance recovery problem is therefore a linear inverse problem: observable residual moments overidentify the latent covariance blocks, so the adjoint equation admits multiple solutions unless a normalization is imposed. We select the representer of smallest Frobenius norm.

Definition 4.3 (Minimum-Frobenius-norm representer and Riesz variance estimator)**.** Let $\|\cdot\|_F$ denote the Frobenius norm induced by $\langle \cdot, \cdot \rangle_{\mathcal{K}}$. Define

$$D^\star := \left(\mathcal{A}^*\right)^+ R^\star, \tag{20}$$

where $(\mathcal{A}^*)^+$ denotes the Moore–Penrose pseudoinverse under the chosen inner products (Golub and Kahan, 1965); equivalently, $D^\star$ is the minimum-Frobenius-norm minimizer of $\|\mathcal{A}^*(D) - R^\star\|_{\mathcal{H}}$ over $D \in \mathcal{K}$. Whenever (19) is solvable, that is whenever $R^\star \in \mathrm{Range}(\mathcal{A}^*)$, this coincides with

$$D^\star = \arg\min_{D \in \mathcal{K}} \left\{ \|D\|_F : \ \mathcal{A}^*(D) = R^\star \right\}. \tag{21}$$

Given sample residual moments $\widehat{C}_{hh'} = \widehat{U}_h \widehat{U}_{h'}'$, the Riesz variance estimator is

$$\boxed{\widehat{V}_{\mathrm{Riesz}} := \langle \widehat{C}, D^\star \rangle_{\mathcal{K}} = \frac{1}{n^2} \sum_{h,h'=1}^{G} \widehat{U}_h' D^\star_{hh'} \widehat{U}_{h'}.} \tag{22}$$

Definition (20) does not require (19) to be solvable.[6] The minimum-norm choice selects the unique representer orthogonal to $\mathrm{Ker}(\mathcal{A}^*)$. Adding $Z \in \mathrm{Ker}(\mathcal{A}^*)$ leaves $\langle C, D \rangle_{\mathcal{K}}$ unchanged for every *population* moment array $C \in \mathrm{Range}(\mathcal{A})$, so all solutions represent the same functional. The map $\mathcal{A}(\Sigma) \mapsto \Phi(\Sigma)$ is well defined on the space $\mathrm{Range}(\mathcal{A})$ of attainable moments exactly when $R^\star \in \mathrm{Range}(\mathcal{A}^*)$, and $D^\star \in \mathrm{Range}(\mathcal{A})$ is then its unique representer there; this is the sense in which the estimator is a Riesz estimator. The sample array does not lie in $\mathrm{Range}(\mathcal{A})$ — $\widehat{U}\widehat{U}' =$

[6] Solvability is the case Assumption A5(i) maintains. We define $D^\star$ through the pseudoinverse because Proposition 6.4 dispenses with solvability, requiring only that the unattainable component of $R^\star$ be negligible at the scale of the target; Proposition 4.5 makes the distinction exact.

$M(UU')M$ and UU' is not block diagonal — so different solutions give different estimators, and the normalization is a genuine choice. Proposition 4.4 shows it is the right one. Because M is symmetric and idempotent, the estimator reduces to a weighting $n^{-2}\sum_g \widehat{U}_g' H_g^\star \widehat{U}_g$ of the within-cluster outer products, the functional form of CR0–CR3 with different weights (Corollary 7.2); under weighted least squares it does not (Section 9).

Proposition 4.4 (Variance optimality of the minimum-norm representer)**.** *Let $D = D^\star + Z$ with $Z \in \mathrm{Ker}(\mathcal{A}^*)$. Then $MD^\star M = D^\star$, and*

$$\|MDM\|_F^2 \;=\; \|D^\star\|_F^2 + \|MZM\|_F^2 \;\geq\; \|D^\star\|_F^2.$$

Consequently, over all solutions of (19), $D^\star$ minimizes $\mathrm{Var}\left(\langle \widehat{C}, D\rangle_{\mathcal{K}} \mid X, W\right)$ when the errors are independent with common variance σ^2 and fourth moment $3\sigma^4$, as when $U \sim \mathcal{N}(0, \sigma^2 I_n)$. For any errors satisfying Assumptions A1 and A2, it minimizes the bound $\mathrm{Var}(\widehat{V} \mid X, W) \leq C(\sup_g N_g)^2 n^{-2}\|MDM\|_{\mathcal{K}}^2$ implied by Lemma 8.1, with C free of D.

Unlike the classical weights, $D^\star$ need not be positive semidefinite, so $\widehat{V}_{\mathrm{Riesz}}$ can be negative in a given sample, although $\Pr(\widehat{V}_{\mathrm{Riesz}} > 0) \to 1$ by Theorem 6.1.[7]

When (19) is not exactly solvable, a representer still delivers an exact identity rather than a bound, and it is the object around which the asymptotic theory of Section 6 is organized.

Proposition 4.5 (Exact bias of a representer)**.** *Fix n, let $\Sigma \in \mathcal{H}$ and $C = \mathcal{A}(\Sigma)$. For* every *$D \in \mathcal{K}$,*

$$\langle C, D\rangle_{\mathcal{K}} - \Phi(\Sigma) \;=\; \left\langle \Sigma,\; \mathcal{A}^*(D) - R^\star \right\rangle_{\mathcal{H}}, \tag{23}$$

so that $|\langle C, D\rangle_{\mathcal{K}} - \Phi(\Sigma)| \leq \|\Sigma\|_{\mathcal{H}}\|\mathcal{A}^*(D) - R^\star\|_{\mathcal{H}}$. The representer of Definition 4.3 satisfies $\mathcal{A}^*(D^\star) = (I - \Pi_{\mathcal{N}})R^\star$ and hence

$$\langle C, D^\star\rangle_{\mathcal{K}} \;=\; \Phi\big((I - \Pi_{\mathcal{N}})\Sigma\big) \;=\; \Phi(\Sigma) - \Phi\big(\Pi_{\mathcal{N}}\Sigma\big). \tag{24}$$

Neither identity requires $R^\star \in \mathrm{Range}(\mathcal{A}^*)$.

Theorem 4.6 (Exact unbiasedness)**.** *Fix n and the design. The Riesz variance estimator satisfies $\mathbb{E}[\widehat{V}_{\mathrm{Riesz}} \mid X, W] = V$ for every $\Sigma \in \mathcal{H}$ if and only if $R^\star \in \mathrm{Range}(\mathcal{A}^*)$. When the condition fails, $\mathbb{E}[\widehat{V}_{\mathrm{Riesz}} \mid X, W] = \Phi\big((I - \Pi_{\mathcal{N}})\Sigma\big)$, the variance of the identified part of the covariance, and no estimator of the form $\langle \widehat{C}, D\rangle_{\mathcal{K}}$ is unbiased for every $\Sigma \in \mathcal{H}$.*

Under Assumption A5(i) the right-hand side of (23) vanishes at $D = D^\star$ and (24) reduces to $\langle C, D^\star\rangle_{\mathcal{K}} = \Phi(\Sigma)$, the identity behind Theorem 4.6. Without it, $D^\star$ targets the *identified component* of the variance, the functional evaluated at the part of Σ that residualization preserves.[8] The

[7] Because $\widehat{U}$ ranges over $\mathrm{Range}(M)$ and $D^\star = MD^\star M$, the estimator is non-negative in every realization if and only if $D^\star \succeq 0$. Since $D^\star \in \mathrm{Range}(\mathcal{A})$, $D^\star = MH^\star M$ for a block-diagonal $H^\star$ (Corollary 7.2), so $H_g^\star \succeq 0$ for every g is sufficient, though not necessary. In a given sample we recommend reporting $\max\{\widehat{V}_{\mathrm{Riesz}}, \widehat{V}_{\mathrm{CR2}}\}$ together with $\min_g \lambda_{\min}(H_g^\star)$, available at no extra cost: a non-negative value certifies $\widehat{V}_{\mathrm{Riesz}} \geq 0$, and a negative value is inconclusive; the exact check $\lambda_{\min}(D^\star)$ can be computed matrix-free by Lanczos iteration on $v \mapsto MH^\star Mv$. Imposing $H_g \succeq 0$ in the minimum-norm problem gives a semidefinite program and forfeits the matrix-free structure of Section 5.

[8] The shortfall $\Phi(\Pi_{\mathcal{N}}\Sigma)$ is not signed, since $\Pi_{\mathcal{N}}\Sigma$ need not be positive semidefinite, so failure of the range condition does not make the estimator conservative. Exact solvability requires the shortfall to vanish for *every* admissible Σ, which

identity prices the three departures from exactness in this paper: an iterative solver stopped early (Theorem 6.2), a design on which the range condition fails (Proposition 6.4), and a classical weighting scheme that approximates the inverse rather than computing it (Proposition E.5).

The characterization above reduces variance recovery to the linear inverse problem (19). Because $\mathcal{K}$ can be high-dimensional and the operator is available only through evaluations of $(\mathcal{A}, \mathcal{A}^*)$, we compute the minimum-norm solution by an iterative matrix-free Krylov method (Section 5).

5 Feasible Estimation and Numerical Implementation

This section describes how to compute the minimum-norm representer $D^\star$ of (20), and hence the feasible variance estimator (22). Direct pseudoinversion by QR or SVD would require constructing the projection-induced operators explicitly, which is infeasible at the cluster scales considered here. We instead use the LSQR algorithm of Paige and Saunders (1982),[9] a matrix-free Krylov method that needs only repeated evaluations of $\mathcal{A}$ and $\mathcal{A}^*$. Neither operator is ever formed; Section 5.1 gives the algorithm and its stopping rule.

Two facts make $D^\star$ computable. First, $\mathcal{A}$ and $\mathcal{A}^*$ act through the projection blocks alone,

$$\begin{aligned} [\mathcal{A}(H)]_{hh'} &= \sum_{g=1}^{G} M_{hg} H_g M_{gh'}, \qquad H \in \mathcal{H}, \\ [\mathcal{A}^*(D)]_g &= \sum_{h,h'=1}^{G} M_{gh} D_{hh'} M_{h'g}, \quad D \in \mathcal{K}, \end{aligned} \tag{25}$$

so each evaluation is a sequence of products with the blocks $\{M_{gh}\}$. Second, a Krylov method started at $D^{(0)} = 0$ builds its iterates from repeated applications of $\mathcal{A}$ and $\mathcal{A}^*$, so they stay in $\mathrm{Range}(\mathcal{A}) = \mathrm{Ker}(\mathcal{A}^*)^\perp$ and converge to the minimum-norm solution $D^\star$ rather than to another solution of (19) (Paige and Saunders, 1982; Section C.1). LSQR has both properties and works with $\mathcal{A}$ and $\mathcal{A}^*$ directly. Solving the restricted system $\mathcal{A}^*\mathcal{A}(H) = R^\star$ and setting $D^\star = \mathcal{A}(H^\star)$ gives the same estimator (Corollary 7.2) but squares the condition number (Proposition E.1); LSQR forms neither $\mathcal{A}^*\mathcal{A}$ nor $\mathcal{A}\mathcal{A}^*$ and so avoids this.

5.1 The solver

We compute $D^\star$ with the LSQR algorithm of Paige and Saunders (1982), which applies Golub–Kahan bidiagonalization to the pair $(\mathcal{A}, \mathcal{A}^*)$: it builds orthonormal bases $\{u_j\} \subset \mathcal{H}$ and $\{v_j\} \subset \mathcal{K}$ of the Krylov subspaces $\mathscr{K}_t(\mathcal{A}^*\mathcal{A}, R^\star)$ and $\mathscr{K}_t(\mathcal{A}\mathcal{A}^*, \mathcal{A}R^\star)$, where $\mathscr{K}_t(\mathcal{T}, r) := \mathrm{span}\{r, \mathcal{T}r, \ldots, \mathcal{T}^{t-1}r\}$, by applying the two operators in turn, and takes as iterate the best approximation available in the subspace generated so far, which is (26) in step (2) of Algorithm 1. In these bases that problem is a

is why it is a condition on the design alone (Proposition 6.3). And the defect $\delta := \|\Pi_{\mathcal{N}} R^\star\|_{\mathcal{H}}$ is a floor rather than a tolerance: $\inf_{D \in \mathcal{K}} \|\mathcal{A}^*(D) - R^\star\|_{\mathcal{H}} = \delta$, so no representer, and by Proposition 7.1 no restricted system, reduces the bias below $\|\Sigma\|_{\mathcal{H}}\delta$ uniformly in Σ.

[9] Conjugate gradients on the normal equations produce the same iterates in exact arithmetic but square the condition number in floating point. LSMR (Fong and Saunders, 2011), which minimizes the second residual in (27) rather than the first over the same subspace, would make the second criterion of (28) monotone and is a natural alternative when the range condition is expected to fail.

bidiagonal least-squares problem of size t, and the algorithm solves it by updating a QR factorization with one Givens rotation per step; Section C.1 gives the details.

Algorithm 1: LSQR for the minimum-norm Riesz representer

Input: Actions of $\mathcal{A}$ and $\mathcal{A}^*$; target $R^\star \in \mathcal{H}$; tolerance τ_n

Output: $D_t \approx D^\star$; variance estimate $\widehat{V} = \langle \widehat{C}, D_t \rangle_{\mathcal{K}}$

(0) Initialization:

$b_1 \leftarrow \|R^\star\|_{\mathcal{H}}, \quad u_1 \leftarrow R^\star / b_1, \quad D_0 \leftarrow 0$

$a_1 \leftarrow \|\mathcal{A} u_1\|_{\mathcal{K}}, \quad v_1 \leftarrow \mathcal{A} u_1 / a_1$

$w_1 \leftarrow v_1, \quad \bar{\rho}_1 \leftarrow a_1, \quad \bar{\zeta}_1 \leftarrow b_1$

for $t = 1, 2, \ldots$ **do**

(1) Bidiagonalization (one $\mathcal{A}^*$, then one $\mathcal{A}$):

$\tilde{u} \leftarrow \mathcal{A}^* v_t - a_t u_t, \quad b_{t+1} \leftarrow \|\tilde{u}\|_{\mathcal{H}}, \quad u_{t+1} \leftarrow \tilde{u} / b_{t+1}$

$\tilde{v} \leftarrow \mathcal{A} u_{t+1} - b_{t+1} v_t, \quad a_{t+1} \leftarrow \|\tilde{v}\|_{\mathcal{K}}, \quad v_{t+1} \leftarrow \tilde{v} / a_{t+1}$

(2) Projected least-squares solve:

the iterate is the best approximation in the subspace generated so far,

$$D_t \;=\; \arg\min \big\{ \|\mathcal{A}^* D - R^\star\|_{\mathcal{H}} \;:\; D \in \mathscr{K}_t(\mathcal{A}\mathcal{A}^*, \mathcal{A} R^\star) \big\}, \tag{26}$$

computed by one Givens rotation:

$\rho_t \leftarrow (\bar{\rho}_t^2 + b_{t+1}^2)^{1/2}, \quad c_t \leftarrow \bar{\rho}_t / \rho_t, \quad s_t \leftarrow b_{t+1} / \rho_t$

$\theta_{t+1} \leftarrow s_t \, a_{t+1}, \quad \bar{\rho}_{t+1} \leftarrow -c_t \, a_{t+1}$

$\zeta_t \leftarrow c_t \, \bar{\zeta}_t, \quad \bar{\zeta}_{t+1} \leftarrow s_t \, \bar{\zeta}_t$

$D_t \leftarrow D_{t-1} + (\zeta_t / \rho_t) \, w_t, \quad w_{t+1} \leftarrow v_{t+1} - (\theta_{t+1} / \rho_t) \, w_t$

(3) Stopping rule (28) (scalars only):

$\phi_t \leftarrow |\bar{\zeta}_{t+1}|, \quad \psi_t \leftarrow |\bar{\zeta}_{t+1}| \, a_{t+1} \, |c_t|$

if $\phi_t / b_1 \le \tau_n$ ***or*** $\psi_t / \phi_t \le \tau_n$ **then**

break

return $D_t, \quad \widehat{V} \leftarrow \langle \widehat{C}, D_t \rangle_{\mathcal{K}}$

Every iterate lies in $\mathrm{Range}(\mathcal{A})$, so $D_t = \mathcal{A}(H_t)$ for block weights $H_t \in \mathcal{H}$, and since $M\widehat{U} = \widehat{U}$ the estimate is $\langle \widehat{C}, D_t \rangle_{\mathcal{K}} = n^{-2} \sum_g \widehat{U}_g' H_{t,g} \widehat{U}_g$, the functional form of Corollary 7.2. The implementation therefore carries H_t rather than D_t and stores neither the representer nor a basis of $\mathcal{K}$; each iteration applies $\mathcal{A}$ and $\mathcal{A}^*$ once (Section C.1).

It remains to say when to stop. We monitor two residuals,

$$\phi_t := \big\| R^\star - \mathcal{A}^* D_t \big\|_{\mathcal{H}}, \qquad \psi_t := \big\| \mathcal{A}\big(R^\star - \mathcal{A}^* D_t\big) \big\|_{\mathcal{K}}, \tag{27}$$

both available from the bidiagonalization scalars at no additional operator applications,[10] and terminate when

$$\frac{\phi_t}{\|R^\star\|_{\mathcal{H}}} \le \tau_n \qquad \text{or} \qquad \frac{\psi_t}{\phi_t} \le \tau_n. \tag{28}$$

[10] In the notation of Algorithm 1, $\phi_t = |\bar{\zeta}_{t+1}|$ and $\psi_t = |\bar{\zeta}_{t+1}| \, a_{t+1} |c_t|$ (Lemma C.1). The usual normalization of the second criterion divides by $\|\mathcal{A}\|_{\mathrm{op}}$ as well; we omit it, since $\|\mathcal{A}\|_{\mathrm{op}} \le 1$ by Lemma A.1 and the operator is never formed, and the omission only tightens the rule.

The first is the primal residual. By Proposition 4.5 it measures directly the quantity that governs the bias of the resulting estimator, and under Assumption A5(i) it is attainable. The second is the normal-equation residual, and it is the operative criterion when exact solvability fails: ϕ_t then decreases only to $\delta_n := \|\Pi_{\mathcal{N}_n} R^\star\|_{\mathcal{H}} > 0$ and the first criterion may never be met, while $\psi_t \to 0$, so the ratio ψ_t/ϕ_t does and the iteration terminates, returning the minimum-norm *least-squares* solution. Proposition 6.4 shows what that solution delivers, which is why the second criterion is not optional even though Assumption A5(i) is maintained. The two are complementary rather than redundant: when $\delta_n = 0$ the residual lies in $\mathrm{Ker}(\mathcal{A}_n)^\perp$, so $\psi_t \geq s_n^{-1}\phi_t$ and the second criterion cannot fire once $\tau_n < s_n^{-1}$ — which the tolerance condition (35) of Theorem 6.2 guarantees for large n — while the first does. Running both is what guarantees termination on any design. Lemma 5.1 collects what the two residuals control.

Lemma 5.1 (LSQR error and residual control)**.** *Let $\mathcal{T} := \mathcal{A}\mathcal{A}^*$, with norm $\|x\|_{\mathcal{T}} := \langle x, \mathcal{T}x\rangle_{\mathcal{K}}^{1/2} = \|\mathcal{A}^* x\|_{\mathcal{H}}$, and $s_n := 1/\sigma^+_{\min}(\mathcal{A}_n)$, so that $\lambda_{\min}(\mathcal{T}_n) = s_n^{-2}$ on $\mathrm{Range}(\mathcal{A}_n)$, and let $D^\star$ be the representer of Definition 4.3. Write $\Pi_{\mathcal{N}} := \Pi_{\mathcal{N}_n}$, $\delta_n := \|\Pi_{\mathcal{N}} R^\star\|_{\mathcal{H}}$ and*

$$r_t := R^\star - \mathcal{A}^* D_t, \qquad r_t^{\mathrm{red}} := (I - \Pi_{\mathcal{N}})R^\star - \mathcal{A}^* D_t.$$

Then, for every iteration t:

(i) (Decomposition.) $r_t = \Pi_{\mathcal{N}} R^\star + r_t^{\mathrm{red}}$, the two terms orthogonal with $r_t^{\mathrm{red}} \in \mathrm{Range}(\mathcal{A}^*)$, so that

$$\phi_t^2 \;=\; \delta_n^2 + \big\|r_t^{\mathrm{red}}\big\|_{\mathcal{H}}^2, \qquad \psi_t \;=\; \big\|\mathcal{A}\, r_t^{\mathrm{red}}\big\|_{\mathcal{K}}. \tag{29}$$

Under Assumption A5(i), $\delta_n = 0$ and $\phi_t = \|r_t^{\mathrm{red}}\|_{\mathcal{H}}$.

(ii) (Deterministic bounds.)

$$\big\|r_t^{\mathrm{red}}\big\|_{\mathcal{H}} \leq s_n\, \psi_t, \qquad \|D_t - D^\star\|_{\mathcal{T}} \leq s_n\, \psi_t, \qquad \|D_t - D^\star\|_{\mathcal{K}} \leq s_n^2\, \psi_t. \tag{30}$$

(iii) (Monotonicity.) ϕ_t is nonincreasing with $\phi_t \leq \|R^\star\|_{\mathcal{H}}$, immediately from (26), and $\|D_t\|_{\mathcal{K}}$ is nondecreasing with $\|D_t\|_{\mathcal{K}} \leq \|D^\star\|_{\mathcal{K}}$.

(iv) (Bracket for the defect.) $\sqrt{\max\{0,\ \phi_t^2 - s_n^2\psi_t^2\}} \;\leq\; \delta_n \;\leq\; \phi_t$.

Since $\psi_t \leq \|\mathcal{A}\|_{\mathrm{op}}\phi_t \leq \phi_t$, the bounds in (30) are no weaker than their counterparts in terms of ϕ_t, and unlike the latter they remain informative when $R^\star \notin \mathrm{Range}(\mathcal{A}^*)$.

Parts (i) and (ii) isolate the *reducible* residual, the only portion of ϕ_t that further iteration can remove, and bound it by the computable ψ_t; with Proposition 4.5 they control the bias of the feasible estimator. Part (iii) makes a separate condition on the iterate unnecessary, since $\|D_t\|_{\mathcal{K}}$ never exceeds $\|D^\star\|_{\mathcal{K}}$. Part (iv) is a diagnostic: a ϕ_t that stalls at a level not explained by $s_n\psi_t$ is evidence against Assumption A5(i), and an estimate of δ_n when it fails.

The conditioning sequence enters (30) linearly in the residual channel and quadratically in the coefficient channel, but the tolerance conditions of Theorem 6.2 and Proposition 6.4 involve it at most linearly, because the proof bounds the bias directly through Proposition 4.5 and never has

to bound $\|D_t - D^\star\|_{\mathcal{K}}$, the quantity that costs s_n^2. The asymmetry is the numerical counterpart of Section 2.2: the tolerance can always be tightened at the cost of more iterations, while the sampling noise cannot. When s_n is bounded, any $\tau_n = n^{-\alpha}$ with $\alpha \geq 1/3$ satisfies both (35) and (39), since s_n then contributes only a constant; when it is not, τ_n must absorb s_n, and Section 7.5 gives the practical reading.[11] In ill-conditioned settings, numerical stability may be improved by standard techniques such as ridge (Tikhonov) regularization or preconditioning; see, e.g., Saad (2003); Hansen (1998).

5.2 Degrees-of-freedom correction

The Riesz estimator is unbiased, but with few clusters it is a noisy estimate of V, and the studentized statistic then has heavier tails than the normal. We adopt the Satterthwaite approximation behind the Bell–McCaffrey adjustment for CR2 (Satterthwaite, 1946; Bell and McCaffrey, 2002; Imbens and Kolesar, 2016). Every estimator here is a quadratic form $\widehat{V} = U'KU$, with $K = n^{-2}MD^\star M$ for $\widehat{V}_{\mathrm{Riesz}}$. Under Gaussian errors with working covariance Ω_w, it is a weighted sum $\sum_j \lambda_j \xi_j^2$ of independent χ_1^2 variables, with λ_j the eigenvalues of $K\Omega_w$. Matching its mean $\sum_j \lambda_j$ and variance $2\sum_j \lambda_j^2$ to those of a scaled χ_ν^2 gives

$$\nu_{\mathrm{df}} \;=\; \frac{\left(\operatorname{tr} K\Omega_w\right)^2}{\operatorname{tr}\left(K\Omega_w K\Omega_w\right)} \;=\; \frac{\left(\sum_j \lambda_j\right)^2}{\sum_j \lambda_j^2}, \tag{31}$$

and the interval uses the $t_{\nu_{\mathrm{df}}}$ critical value; $\widehat{V}$ itself is unchanged. We take $\Omega_w = I_n$, as Bell and McCaffrey do, and write Riesz-t_ν for the resulting interval and Riesz-t_{G-1} for the common alternative with $G-1$ degrees of freedom.

ν_{df} acts as an effective sample size: $\widehat{V}$ is as precise as an average of ν_{df} independent squared normal errors, since $\operatorname{Var}(\widehat{V})/(\mathbb{E}\widehat{V})^2 = 2/\nu_{\mathrm{df}}$. Like an inverse Herfindahl index, it equals the number of terms when the weights λ_j are equal, and falls below it when a few weights dominate or when weights of opposite signs cancel in the mean but not in the variance. The rule t_{G-1} is accurate when $\widehat{V}$ behaves like a sum of G comparable cluster terms. The fully robust representer, being unbiased for every admissible covariance, concentrates on a few directions with weights of both signs, so its ν_{df} falls below $G-1$. The correction needs no block inverse, so it exists wherever $\widehat{V}_{\mathrm{Riesz}}$ does, including designs in which CR2 and its adjustment do not. With very few clusters ν_{df} is small and the t_ν interval, though long, still undercovers; a restricted covariance class (Section 8.3) has a smaller representer and a larger ν_{df}, and is the better choice there (Section 10).

6 Asymptotic Theory

6.1 Assumptions

Let $\|\cdot\|_F$ denote the Frobenius norm and $\|\cdot\|_{\mathrm{op}}$ the operator (spectral) norm.

Assumption A1 (Sampling)**.** The sample consists of $G = G_n$ clusters with $\sum_{g=1}^G N_g = n$ and $G \to \infty$. Conditional on (X, W), errors $\{U_g\}$ are independent across clusters and satisfy

$$\mathbb{E}[U_g \mid X, W] = 0, \qquad \mathbb{E}[U_g U_g' \mid X, W] = \Sigma_g,$$

[11] The first part of (35) holds as soon as $\tau_n \mu_n^{\lambda/(2+2\lambda)} \to 0$, since Assumption A3 gives $\sup_g N_g/n = o(\mu_n^{-(2+\lambda)/(2+2\lambda)})$ and hence $\mu_n \sup_g N_g/n = o(\mu_n^{\lambda/(2+2\lambda)})$. The exponent is at most $1/3$ over $\lambda \in (0, 2]$ and $\mu_n \leq n$.

where Σ_g is positive semi-definite.

Assumption A2 (Moment bounds)**.** For some $0 < \lambda \leq 2$,

$$\sup_{g,i} \mathbb{E}\Big[|u_{gi}|^{4+2\lambda} \mid X, W\Big] < \infty, \qquad \sup_{g} \frac{1}{N_g} \sum_{i=1}^{N_g} \mathbb{E}\|\tilde{x}_{gi}\|^{4+2\lambda} < \infty.$$

The first supremum is required to be bounded by a finite constant almost surely.

Assumption A3 (Cluster sizes)**.** There exists a deterministic sequence $\mu_n \to \infty$ and a finite constant $\nu_a > 0$ such that

$$\mu_n V \xrightarrow{p} \nu_a,$$

and the maximal cluster size obeys

$$\mu_n^{\frac{2+\lambda}{2+2\lambda}} \sup_g \frac{N_g}{n} \to 0.$$

Assumption A4 (Design regularity)**.** The residualized design satisfies

$$\Gamma_n := \frac{1}{n} \sum_{g=1}^{G} \widetilde{X}_g' \widetilde{X}_g \xrightarrow{p} \Gamma_0 \succ 0,$$

and $\operatorname{rank}([X\ W]) < n$ with probability approaching one.[12]

Assumption A5 (Riesz identification and stability)**.** Let $R_n^\star := \{\widetilde{X}_g Q_n \widetilde{X}_g'\}_{g=1}^G$ with $Q_n := \Gamma_n^{-1} a a' \Gamma_n^{-1}$, let $\mathcal{A}_n^*$ be the adjoint operator in (18), and set

$$s_n := \frac{1}{\sigma_{\min}^+(\mathcal{A}_n)} = \left\| (\mathcal{A}_n^*)^+ |_{\mathrm{Range}(\mathcal{A}_n^*)} \right\|_{\mathrm{op}},$$

where $\sigma_{\min}^+$ denotes the smallest nonzero singular value. Since $R_n^\star$, $\mathcal{A}_n$ and s_n are functions of (X, W):

(i) (Exact solvability.) With probability approaching one,

$$R_n^\star \in \mathrm{Range}(\mathcal{A}_n^*) = \mathrm{Ker}(\mathcal{A}_n)^\perp.$$

(ii) (Stability.) The conditioning sequence satisfies

$$s_n \mu_n \Big(\frac{\sup_g N_g}{n} \Big)^{3/2} \xrightarrow{p} 0.$$

Assumptions A1 to A5 jointly ensure that the variance functional $\Phi(\Sigma) = \mathrm{Var}(a'\widehat{\beta})$ is identifiable from the observable residual moments and that the resulting Riesz-based t-statistic is asymptotically normal. Assumptions A1 to A4 provide standard conditions for cluster asymptotics, following Djogbenou et al. (2019) and related work, and guarantee that the target representer $R_n^\star$ is well defined

[12]The theory uses W only through $M_W = I_n - \Pi_W$, which is defined for any W, so neither invertibility of $W'W$ nor a lower bound on $\lambda_{\min}(n^{-1}W'W)$ is required; both fail with many or collinear controls. The rank condition excludes only $M = 0$, in which no residual information is available.

with $\|R_n^\star\|_{\mathcal{H}_n}$ stochastically bounded. While Assumptions A1 to A4 restrict the data-generating process, Assumption A5 restricts the operator $\mathcal{A}_n^*$ induced by the projection geometry of the design, and is therefore a statement about (X, W) rather than about the errors.

The two parts do different work. Part (i) is an *identification* condition, not a regularity condition. Because $\mathrm{col}(\widetilde{X}) \subseteq \mathrm{col}([X\ W])$, the functional $\Phi(\Sigma)$ depends on Σ only through $\Pi_{[X\ W]}\Sigma\Pi_{[X\ W]}$, where $\Pi_{[X\ W]} = I_n - M$, while the observable moments reveal Σ only through $M\Sigma M$: the sampling uncertainty in $\widehat{\beta}$ originates in $\mathrm{col}([X\ W])$ and the evidence lives in its orthogonal complement, the tension described at the start of Section 2. Assumption A5(i) assumes the first is nonetheless recoverable from the second along the direction $R_n^\star$. Section 6.3 gives its exact form and weakens it to a rate; in the heteroskedastic case $N_g = 1$ it reduces to $R_n^\star \perp \mathrm{Ker}(M \odot M)$ and is implied by nonsingularity of $M \odot M$, so Assumption A5 is the block generalization of the invertibility-with-bounded-inverse requirement of Cattaneo et al. (2018).

Part (ii) is a well-posedness condition, stated as a rate rather than as a uniform bound. Since $\|\mathcal{A}_n\|_{\mathrm{op}} \le 1$ by Lemma A.1, requiring $\sup_n s_n \le c < \infty$ would bound the condition number $\kappa(\mathcal{A}_n) = \|\mathcal{A}_n\|_{\mathrm{op}}\, s_n \le s_n$ uniformly, and is equivalent to doing so whenever $\|\mathcal{A}_n\|_{\mathrm{op}}$ is bounded below — as it is when $\mathrm{rank}([X\ W])/n$ stays bounded away from one, since $\mathcal{A}_n(I_n) = M$ gives $\|\mathcal{A}_n\|_{\mathrm{op}} \ge \|M\|_F/\|I_n\|_F = (1 - \mathrm{rank}([X\ W])/n)^{1/2}$. A uniform bound would assume away precisely the ill-conditioned designs that motivate the estimator in Section 2.2. The rate form permits $s_n \to \infty$: under Assumption A3 the slack is $\mu_n^{(2-\lambda)/(4+4\lambda)}$, so s_n may diverge at any rate $o\big(\mu_n^{(2-\lambda)/(4+4\lambda)}\big)$, which approaches $\mu_n^{1/2}$ as $\lambda \to 0$ and closes only in the limiting case $\lambda = 2$, where the moment condition is strongest and the cluster-size restriction of Assumption A3 is weakest. (The slack is computed at the boundary of Assumption A3; designs with smaller clusters leave more.) The role of s_n is transparent: $\|D_n^\star\|_{\mathcal{K}_n} \le s_n \|R_n^\star\|_{\mathcal{H}_n}$ by Lemma A.5, and the sampling variance of $\widehat{V}_{\mathrm{Riesz}}$ is controlled by $\|D_n^\star\|_{\mathcal{K}_n}$.

6.2 Main Results

Theorem 6.1 (Asymptotic normality)**.** *Suppose Assumptions A1 to A5 hold. Then,*

$$V^{-1/2}\, a'(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, 1), \tag{32}$$

$$\widehat{V}_{\mathrm{Riesz}}/V \xrightarrow{p} 1, \tag{33}$$

$$\widehat{V}_{\mathrm{Riesz}}^{-1/2}\, a'(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, 1). \tag{34}$$

Theorem 4.6 gives exactness at every n; Theorem 6.1 adds the asymptotics. Its proof verifies the two conditions of Proposition D.1, a bias condition and a norm condition: Assumption A5(i) makes the first hold exactly and Assumption A5(ii) delivers the second.

Theorem 6.2 (Feasible implementation via LSQR)**.** *Let $\widehat{D}_n = D_{t_n}$ be the LSQR approximation to $D_n^\star$, computed as in Algorithm 1 and terminated by either criterion of (28), with a tolerance sequence τ_n satisfying*

$$\mu_n\, \tau_n\, \frac{\sup_g N_g}{n} \longrightarrow 0 \qquad \textit{and} \qquad \tau_n\, s_n \xrightarrow{p} 0. \tag{35}$$

Define $\widehat{V}^{\mathrm{LSQR}}_{\mathrm{Riesz}}$ *as in* (22) *with* $D^\star_n$ *replaced by* $\widehat{D}_n$. *Then, under Assumptions A1 to A5,*

$$\widehat{V}^{\mathrm{LSQR}}_{\mathrm{Riesz}}/V \xrightarrow{p} 1, \tag{36}$$

$$\big(\widehat{V}^{\mathrm{LSQR}}_{\mathrm{Riesz}}\big)^{-1/2} a'(\widehat{\beta}-\beta) \xrightarrow{d} \mathcal{N}(0,1). \tag{37}$$

Theorem 6.2 shows that the feasible estimator inherits the conclusions of Theorem 6.1. The first part of (35) makes the primal residual, and with it the bias through Proposition 4.5, negligible; the second keeps the dual criterion inactive under Assumption A5(i), so that the iteration is terminated by the primal one (Section 5.1). Because the bias is bounded directly rather than through $\|\widehat{D}_n - D^\star_n\|_{\mathcal{K}_n}$, the conditioning sequence enters only linearly. When exact solvability is in doubt, the dual criterion becomes the operative one and the tolerance must absorb s_n in the bias as well (Proposition 6.4). Section 7 applies the same approximation logic to classical estimators, whose representers are diagonal or block-diagonal weighting schemes rather than LSQR iterates.

Theorems 6.1 and 6.2 are the one-way case of Theorem 8.2 (Corollary 8.3), which extends them to estimators linear in the errors and to dependence that does not partition the observations, such as two-way clustering. The degrees-of-freedom correction of Section 5.2 does not affect these first-order results: the t_ν quantile exceeds the normal one and converges to it as $\nu \to \infty$, so the corrected test is asymptotically exact when ν_{df} diverges and conservative otherwise.

6.3 The range condition

Assumption A5(i) is easiest to read through the influence weights of the contrast, $z_g := \widetilde{X}_g \Gamma_n^{-1} a$, for which $R^\star_{n,g} = z_g z_g'$ and $V = n^{-2}\sum_g z_g' \Sigma_g z_g$.

Proposition 6.3 (Range condition as an orthogonality criterion). $R^\star_n \in \mathrm{Range}(\mathcal{A}^*_n)$ *if and only if*

$$\sum_{g=1}^{G} z_g' \, \Delta_g \, z_g = 0 \qquad \textit{for every } \Delta = \{\Delta_g\} \in \mathrm{Ker}(\mathcal{A}_n). \tag{38}$$

What residualization hides must not load on the influence weights. The kernel $\mathcal{N}_n$ is typically nonzero, since $\mathcal{A}_n$ is built from a rank-deficient projection, and it contains a whole within-cluster direction whenever M_{gg} is singular (Proposition 7.4), and whether such a direction matters depends on the contrast (Section 7.4). By Proposition 7.1 the criterion is the same for the full and restricted systems, so its failure is not one the cross-cluster moments can repair. It is measured continuously by the defect $\delta_n := \|\Pi_{\mathcal{N}_n} R^\star_n\|_{\mathcal{H}_n}$, a by-product of the solver, with $\delta_n = 0$ exactly under Assumption A5(i).

Weakening to a rate. Neither theorem needs exact solvability, although the feasible estimator then requires a stronger tolerance rate. Whatever the defect, $\langle C_n, D^\star_n\rangle_{\mathcal{K}_n} = \Phi_n\big((I - \Pi_{\mathcal{N}_n})\Sigma_n\big)$ by Proposition 4.5: the estimator targets the part of the variance that residualization preserves.

Proposition 6.4 (Approximate solvability). *Suppose Assumptions A1 to A4 and Assumption A5(ii) hold, and let* $\widehat{D}_n$ *be the LSQR iterate terminated by either criterion of* (28) *with a tolerance sequence* τ_n *satisfying*

$$\mu_n \, \tau_n \, s_n \, \frac{\sup_g N_g}{n} \longrightarrow 0. \tag{39}$$

Write $\widehat{V}$ for either $\widehat{V}_{\mathrm{Riesz}}$ or $\widehat{V}_{\mathrm{Riesz}}^{\mathrm{LSQR}}$. Then:

(i) whatever the defect,

$$\widehat{V} \;-\; \Phi_n\big((I - \Pi_{\mathcal{N}_n})\Sigma_n\big) \;=\; o_p(\mu_n^{-1});$$

(ii) if in addition

$$\mu_n\,\delta_n\,\sqrt{\frac{\sup_g N_g}{n}} \;\longrightarrow\; 0, \tag{40}$$

the conclusions (33)*,* (34)*,* (36) *and* (37) *of Theorems 6.1 and 6.2 hold.*

Part (i) says the solver reaches the target it can reach; part (ii) makes that target V to first order by requiring the unidentified component to be a vanishing fraction of V. A tighter tolerance cannot substitute, since iterating shrinks only the reducible residual (Lemma 5.1). Termination by the primal criterion certifies $\delta_n \le \tau_n \|R_n^\star\|_{\mathcal{H}_n}$ and hence (40); termination by the dual criterion does not, and (40) must then be assumed or assessed (Section 7.5). We nonetheless maintain the exact form: it is a property of a given design, which Section 7 characterizes, and it makes the estimator exactly unbiased at every n rather than only to first order.[13]

7 What the Classical Restrictions Cost

Classical HC and CR estimators are obtained from the framework of Section 4 by two operations: *restricting* the residual moment array to its within-cluster blocks, and *approximating* the inverse of the resulting system by one that factors across blocks. The first is free under orthogonal projection: the restricted system identifies everything the full array does (Chen, 2026), and its exact solution is the Riesz estimator. The second is not: it biases the estimator when the projection couples clusters, and it fails to exist when a within-cluster block is singular. The classical estimators are therefore restricted-*inversion* estimators, not restricted-*information* estimators. Further results and all proofs are in Section E.

7.1 The Restricted System and the Full System

Define the diagonal moment space and restriction

$$\mathcal{K}_n^{\mathrm{diag}} := \bigoplus_{h=1}^{G} \mathbb{S}^{N_h}, \qquad [P_{\mathrm{diag}}(C)]_{hh'} := \begin{cases} C_{hh}, & h = h', \\ 0, & h \neq h', \end{cases}$$

and the *partial Riesz operator* $\mathcal{A}_n^{\mathrm{diag}} := P_{\mathrm{diag}} \circ \mathcal{A}_n$, which maps $\{\Sigma_g\}$ to the induced within-cluster residual moments $\{C_{hh}\}$, the information set underlying HC and CR. Treating $\mathcal{A}_n^{\mathrm{diag}}$ as an operator on $\mathcal{H}_n$, its adjoint is $[(\mathcal{A}_n^{\mathrm{diag}})^*(\{D_{hh}\})]_g = \sum_h M_{gh} D_{hh} M_{hg}$, and a partial Riesz representer is any $\{D_{hh}\}$ solving

$$(\mathcal{A}_n^{\mathrm{diag}})^*(\{D_{hh}\}) = R_n^\star, \tag{41}$$

[13] Condition (40) also allows a representer that misses part of the target on purpose. The exact solution divides by the singular values of $\mathcal{A}_n$, so weakly identified directions make $\|D_n^\star\|_{\mathcal{K}_n}$, and with it the sampling variance of the estimator, large. A regularized representer keeps the weights small at the cost of a small bias: truncated SVD drops the directions with small singular values, and ridge (Tikhonov) regularization shrinks them. On badly conditioned designs the trade can be worth making, provided the bias stays within (40).

which vectorizes to the block–Khatri–Rao system

$$\underbrace{\begin{pmatrix} M_{11}\otimes M_{11} & \cdots & M_{1G}\otimes M_{1G} \\ \vdots & \ddots & \vdots \\ M_{G1}\otimes M_{G1} & \cdots & M_{GG}\otimes M_{GG} \end{pmatrix}}_{B} \underbrace{\begin{pmatrix} \operatorname{vec}(D_{11}) \\ \vdots \\ \operatorname{vec}(D_{GG}) \end{pmatrix}}_{d} = \underbrace{\begin{pmatrix} \operatorname{vec}(\widetilde{X}_1 Q_n \widetilde{X}_1') \\ \vdots \\ \operatorname{vec}(\widetilde{X}_G Q_n \widetilde{X}_G') \end{pmatrix}}_{r}, \tag{42}$$

since $\operatorname{vec}(M_{gh}D_{hh}M_{hg}) = (M_{gh}\otimes M_{gh})\operatorname{vec}(D_{hh})$. Block (g,h) of B is $N_g^2 \times N_h^2$. With $N_g = 1$, B is the Hadamard square $M \odot M$ and this is the inverse-Hadamard system of Cattaneo et al. (2018).

Proposition 7.1 (The restricted system is the Gram system). *For every design and every n, $\mathcal{A}_n^{\mathrm{diag}} = \mathcal{A}_n^*\mathcal{A}_n$ as operators on $\mathcal{H}_n$, both acting as*

$$\big[\mathcal{A}_n^*\mathcal{A}_n(H)\big]_g \;=\; \sum_{k=1}^{G} M_{gk}H_kM_{kg}, \qquad H \in \mathcal{H}_n; \tag{43}$$

in particular $\mathcal{A}_n^{\mathrm{diag}}$ is self-adjoint and positive semidefinite. Consequently $\operatorname{Ker}(\mathcal{A}_n^{\mathrm{diag}}) = \operatorname{Ker}(\mathcal{A}_n)$ and $\operatorname{Range}((\mathcal{A}_n^{\mathrm{diag}})^*) = \operatorname{Range}(\mathcal{A}_n^*)$, so (41) is solvable if and only if the full adjoint system (19) is. Whenever either holds, for any covariance array Σ_n, any solution $D_n^\star$ of (19) and any solution D_n^{diag} of (41),

$$\big\langle \mathcal{A}_n(\Sigma_n), D_n^\star\big\rangle_{\mathcal{K}_n} = \big\langle \mathcal{A}_n^{\mathrm{diag}}(\Sigma_n), D_n^{\mathrm{diag}}\big\rangle_{\mathcal{K}_n^{\mathrm{diag}}} = \operatorname{Var}(a'\widehat{\beta} \mid X, W).$$

Cross-cluster residual moments respond to the latent covariance, yet they distinguish no pair of covariance arrays the within-cluster moments cannot already distinguish.[14] The equivalence is exact at the level of the estimator, not only in population.

Corollary 7.2 (The restricted and full systems give the same estimator). *Let $H_n^\star := (\mathcal{A}_n^{\mathrm{diag}})^+ R_n^\star$ be the minimum-norm solution of (41), which block by block reads*

$$\sum_{k=1}^{G} M_{gk}H_kM_{kg} \;=\; \widetilde{X}_g Q_n \widetilde{X}_g', \qquad g = 1,\dots,G. \tag{44}$$

Then $D_n^\star = MH_n^\star M$, and for every realization of the residuals

$$\widehat{V}_{\mathrm{Riesz}} \;=\; \big\langle \widehat{C}_n, D_n^\star\big\rangle_{\mathcal{K}_n} \;=\; \frac{1}{n^2}\sum_{g=1}^{G} \widehat{U}_g' H_{n,g}^\star \widehat{U}_g \;=\; \big\langle P_{\mathrm{diag}}\widehat{C}_n, H_n^\star\big\rangle_{\mathcal{K}_n^{\mathrm{diag}}}. \tag{45}$$

Every other solution of (41), if one exists, gives the same value.[15]

Equation (45) is the functional form of CR0–CR3: a block-diagonal weighting of the within-cluster residual outer products. The cross-cluster moments $\widehat{U}_h\widehat{U}_{h'}'$ never have to be formed.

The proof uses only $\widehat{U} = M\widehat{U}$ and $M' = M$, so the corollary holds whether or not the maintained partition is correct, and for any error distribution. The two formulations differ only in conditioning:

[14] This proposition, with Propositions E.1 and 6.3, is the block specialization of a result established in general form in the companion paper (Chen, 2026, Theorem 5.2, Corollary 3.4 and Theorem 5.6).

[15] Neither statement requires (41) to be solvable: $(\mathcal{A}_n^{\mathrm{diag}})^+ R_n^\star$ is then the minimum-norm least-squares solution, and $D_n^\star = (\mathcal{A}_n^*)^+ R_n^\star$ still equals $MH_n^\star M$.

the restricted system is the normal-equation form of the full one, with condition number $\kappa(\mathcal{A}_n)^2$ (Proposition E.1), which costs iterations but never changes the estimator. With weights, where the residual maker is not an orthogonal projection, they do differ (Section 9).

7.2 Classical Estimators as Restriction plus Approximate Inversion

Classical HC and CR do not solve (41). They replace the exact inverse of the coupled restricted system by a block-wise leverage correction, treating $\widehat{U}_g\widehat{U}_g'$ as informative about Σ_g alone. Each restricts the moment set by P_{diag} and then approximates $((\mathcal{A}_n^{\text{diag}})^*)^+$ by an operator that factors across blocks (table 1).

Table 1: Classical estimators as restriction plus approximate inversion

Estimator	Restriction	Approximate inverse	Adequate when
CJN high-dim. HC	P_{diag}, $N_g=1$	$(M_W\odot M_W)^{-1}$	$M\odot M$ invertible, $s_n^4/\sqrt{n}\to 0$
HC2, HC3	P_{diag}, $N_g=1$	$\{(1-h_{ii})^{-1}\}$, $\{(1-h_{ii})^{-2}\}$	$\max_i h_{ii}=O_p(n^{-1})$
CR0–CR3	P_{diag}, general N_g	$M_{gg}^{-\delta}$ blockwise; $\delta=1$ is B_0^{-1}	(Proj-Sparsity), M_{gg} invertible

Every classical weight thus solves a system in which each M_{gk}, $k\neq g$, has been set to zero. With scalar clusters, the exact solution of the restricted system is, to first order, the estimator of Cattaneo et al. (2018), and HC2 and HC3 approximate its leading term when leverage is small. With general clusters, CR3 solves the decoupled system obtained by replacing B in (42) by B_0, with $(B_0)_{gg}=M_{gg}\otimes M_{gg}$, and CR0–CR2 further approximate CR3. Each corrects the local term $M_{gg}\Sigma_g M_{gg}$ block by block and leaves the projection spillover of Section 2.1 untouched (Section E).

Leaving the spillover out is harmless when it is small. The block-wise inverse then delivers an estimator first-order equivalent to the Riesz estimator (Proposition E.5), and a computable sufficient condition is that the normalized block row sum be small,

$$\varrho_n^{\text{sp}} := \max_g \sum_{h\neq g} \left\|M_{gg}^{-1}M_{gh}\right\|_{\text{op}}^2 \;=\; o_p\Big(\min\Big\{1,\;\frac{n}{\mu_n \sup_g N_g}\Big\}\Big). \tag{Proj-Sparsity}$$

Under (Proj-Sparsity), within-cluster leverage is negligible at the μ_n^{-1} scale, and CR0–CR3 are first-order equivalent to one another and to the Riesz estimator. When leakage is not negligible, they can be biased at first order.

7.3 The Bias the Block-wise Inverse Leaves

What the block-wise inverse leaves can be written down exactly. When M_{gg} is nonsingular, conjugating the within-cluster moment (7) by $M_{gg}^{-1/2}$ splits the weight into M_{gg} and $H_{gg}=I_{N_g}-M_{gg}$, the matrix counterparts of $1-h_{ii}$ and h_{ii}:

$$M_{gg}^{-1/2}\,\mathbb{E}[\widehat{U}_g\widehat{U}_g'\mid X,W]\,M_{gg}^{-1/2}-\Sigma_g \;=\; H_{gg}^{1/2}\big(\tilde{\Sigma}_g-\Sigma_g\big)H_{gg}^{1/2} \;+\; R_g, \tag{46}$$

where $\tilde{\Sigma}_g$ is a congruence average of the neighbouring $\{\Sigma_k\}$ and R_g is a remainder that vanishes when Σ_g commutes with H_{gg}, in particular when $N_g=1$ (Lemma E.7). The first term is (5) with

matrices in place of numbers, mass times contrast, and CR2 leaves both terms (Bell and McCaffrey, 2002; Imbens and Kolesar, 2016); the contrast can rotate as well as reweight (Remarks E.8 and E.9).

The level of the mass, for the contrast of interest, is the block leverage κ_n of (8). With controls it is computed from the influence weights of the contrast, $z_g := \widetilde{X}_g c$ with $c := \Gamma^{-1}a$, in place of $X_g c$, so that $\kappa_n = \sum_g z_g' H_{gg} z_g / \sum_g \|z_g\|^2$ and $V = n^{-2}\sum_g z_g' \Sigma_g z_g$.

Proposition 7.3 (The level factor). *Let CR2 and CR3 use the block weights of table 1, CR3 without the jackknife factor $(G-1)/G$.*

(i) Let $\Sigma = \sigma^2 I_n$. Then

$$\mathbb{E}\big[\widehat{V}_{\mathrm{CR0}} \mid X, W\big] \;=\; (1-\kappa_n)\, V, \qquad \mathbb{E}\big[\widehat{V}_{\mathrm{CR1}} \mid X, W\big] \;=\; c_n(1-\kappa_n)\, V,$$

with $c_n = \frac{G}{G-1}\frac{n-1}{n-p}$. If in addition every M_{gg} is nonsingular, then $\mathbb{E}[\widehat{V}_{\mathrm{CR2}} \mid X, W] = V$ and $\mathbb{E}[\widehat{V}_{\mathrm{CR3}} \mid X, W] = (1+\tilde{\kappa}_n)V$, where $\tilde{\kappa}_n := \sum_g z_g' H_{gg} M_{gg}^{-1} z_g / \sum_g \|z_g\|^2$ satisfies

$$\frac{\kappa_n}{1-\kappa_n} \;\le\; \tilde{\kappa}_n \;\le\; \frac{\kappa_n}{1-\bar{\eta}}, \qquad \bar{\eta} := \max_g \|H_{gg}\|_{\mathrm{op}}.$$

(ii) For every $\Sigma \in \mathcal{H}$, if every M_{gg} is nonsingular, $\mathbb{E}[\widehat{V}_{\mathrm{CR3}} \mid X, W] \ge V$.

(iii) If every column of $[X\ W]$ is constant within clusters, with value r_g' in cluster g, then $H_{gg} = h_g^c\, \iota_g \iota_g'/N_g$ with $h_g^c := N_g\, r_g' \big(\sum_k N_k r_k r_k'\big)^{+} r_g$, $\sum_g h_g^c = p$, and κ_n is the weighted average $\sum_g N_g \zeta_g^2 h_g^c / \sum_g N_g \zeta_g^2$ of $\{h_g^c\}$, where $z_g = \zeta_g \iota_g$.

With the decomposition, the proposition separates what sets the level of the bias from what sets its sign: under homoskedasticity CR0 is biased by exactly $-\kappa_n$, CR2 is exact and CR3 over-corrects by at least $\kappa_n/(1-\kappa_n)$, and for any covariance CR3 errs upward, by the spillover it leaves. These are the patterns of figure 8. Part (iii) makes the few-cluster regime of Section 2 exact: κ_n is then of order p/G even if $p/n \to 0$. Section E.3 notes when the off-block mass vanishes. The same number explains CR1: its factor is approximately $1 + p/n + 1/G$, so under homoskedasticity it removes CR0's bias when $\kappa_n \approx p/n + 1/G$, that is, when the regressor varies at the observation level, and misses it by the difference otherwise. Being a scalar, κ_n measures the level of the mass, not its shape, and no scalar reweighting of the clusters removes it.

The sufficient condition (Proj-Sparsity) of Section 7.2 is the uniform, contrast-free version of $\kappa_n \to 0$: since $\sum_{h \ne g} M_{gh} M_{hg} = M_{gg} H_{gg}$,

$$\varrho_n^{\mathrm{sp}} \;\ge\; \max_g \big\| H_{gg} M_{gg}^{-1} \big\|_{\mathrm{op}} \;\ge\; \max_g \|H_{gg}\|_{\mathrm{op}} \;\ge\; \kappa_n \tag{47}$$

(Section E.3), so (Proj-Sparsity) implies $\kappa_n \to 0$ but not conversely: it takes the worst cluster in the worst direction, whereas κ_n weights the block leverage by where the contrast loads.

7.4 When the Block-wise Inverse Does Not Exist

Both $M_{gg}^{-1/2}$ and M_{gg}^{-1} presuppose that $M_{gg} = I_{N_g} - [\Pi_{[X\ W]}]_{gg}$ is nonsingular.

Proposition 7.4 (Singularity of within-cluster blocks). *Let $\mathcal{W}_g := \{w \in \mathbb{R}^n : w_i = 0 \text{ for } i \notin C_g\}$. Then M_{gg} is singular if and only if $\mathcal{W}_g \cap \mathrm{col}([X\ W]) \ne \{0\}$.*

The condition holds whenever the design contains a regressor supported on a single cluster. Cluster fixed effects, the modal specification with clustered data, make every M_{gg} singular at once, and cluster-specific covariates, trends or interactions with a cluster indicator do the same. CR2 and CR3 then cannot be evaluated, and the leave-cluster-out estimator of Anatolyev and Ng (2026) inherits the requirement. The remedies in the literature replace the block inverse by a Moore–Penrose inverse (Pustejovsky and Tipton, 2018). The Riesz estimator forms no block inverse, so singularity does not prevent its computation. It matters only for identification: the residuals carry no information about the errors along an absorbed direction, so the variance must not depend on them, which is the range condition. A confined direction lying in $\mathrm{col}(W)$, such as a cluster fixed effect, satisfies it for every contrast, because the influence weights are built from $\widetilde{X} = M_W X$ and are orthogonal to $\mathrm{col}(W)$ (Proposition E.10).

Proposition 7.5 (Existence hierarchy)**.** *Consider sequences $\{(X_n, W_n, \{N_g\}, a)\}_n$ of designs and contrasts satisfying Assumptions A1 to A4, and let $\mathcal{D}_{\mathrm{Riesz}}$, $\mathcal{D}_{\mathrm{part}}$, $\mathcal{D}_{\mathrm{CR}}$ and $\mathcal{D}^1_{\mathrm{CR}}$ denote the sets of such sequences along which, for all sufficiently large n: (a) $R_n^\star \in \mathrm{Range}(\mathcal{A}_n^*)$, which is Assumption A5(i); (b) the exact restricted Riesz solution exists; (c) (a) holds and every M_{gg} is nonsingular, so that CR2 and CR3 can be evaluated; and, in addition to (c), (d) (Proj-Sparsity) holds, so that CR0–CR3 are first-order valid. Then*

$$\mathcal{D}^1_{\mathrm{CR}} \subsetneq \mathcal{D}_{\mathrm{CR}} \subsetneq \mathcal{D}_{\mathrm{part}} = \mathcal{D}_{\mathrm{Riesz}}.$$

The equality is Proposition 7.1. The second inclusion is strict because of cluster fixed effects, which make every M_{gg} singular while leaving the range condition intact (Proposition E.10); the first, because of the many-parameter and few-cluster regimes of Section 2.1, in which every M_{gg} is nonsingular but leakage is not negligible. Remark E.11 discusses what remains when condition (a) itself fails.

7.5 Practical Guidance

The results of this section and of Sections 5 and 6 translate into five recommendations.

1. *Compute the design diagnostics first.* Three come from the projection blocks before any variance is estimated: κ_n, which under homoskedasticity gives the bias of CR0 exactly, $-\kappa_n$, and that of CR1, $c_n(1-\kappa_n)-1$ (Proposition 7.3); ϱ_n^{sp}, whose smallness (Proj-Sparsity) makes CR0–CR3 first-order equivalent to the Riesz estimator; and the blocks M_{gg}, a singular one meaning that CR2 and CR3 do not exist (Proposition 7.4).

2. *Report several estimators:* CR0 and CR1, CR2 or CR3 where they exist, and the Riesz estimator. A wide spread locates the designs where the block-wise corrections misattribute variance; agreement is reassuring but not conclusive, since a biased estimator can match the Riesz estimate in a given sample (Section 11.2). Since the Riesz weights need not be positive semidefinite, report $\max\{\widehat{V}_{\mathrm{Riesz}}, \widehat{V}_{\mathrm{CR2}}\}$ and $\min_g \lambda_{\min}(H_g^\star)$, whose non-negativity certifies $\widehat{V}_{\mathrm{Riesz}} \geq 0$.

3. *Correct the reference distribution.* Report ν and use t_ν critical values (Section 5.2); when ν is small, use an estimator exact on a restricted class that the residuals do not reject (Section 8.3).

4. *Report which stopping rule fired.* Primal termination certifies $\delta_n \le \tau_n \|R_n^\star\|_{\mathcal{H}_n}$, so the target is recovered to the requested tolerance. Dual termination says only the identified component of V is recovered; Lemma 5.1(iv) then brackets the defect between $\sqrt{\max\{0, \phi_t^2 - s_n^2 \psi_t^2\}}$ and ϕ_t.

5. *Tighten the tolerance when conditioning is poor.* The error at a given tolerance grows roughly like $\tau_n \kappa(\mathcal{A}_n)$ (table 30), so a large $\kappa(\mathcal{A}_n)$ calls for a smaller τ_n, at a cost of iterations.

8 A General Theorem and Extensions

Nothing in Sections 4 to 6 is specific to OLS or to the maintained space $\mathcal{H}_n$. Three ingredients do all the work: an estimator that is *linear* in the errors, a residual vector that is a *known linear map* of the errors, and a maintained *linear class* of covariance arrays. The ingredients also set the scope: the estimation error must be linear in the errors, with weights fixed by a design on which the errors have mean zero; this excludes instrumental variables, nonlinear models, feasible GLS with estimated weights (Section 9.3) and lagged dependent variables, but not time-series regressions with strictly exogenous regressors (Remark 8.7). The dependence support $\mathcal{S}_n$ must also be specified rather than estimated. This section states the construction and its theory at that level once (Section 8.1), with the one-way model as its leading case, and reads off two extensions: dependence that does not partition the observations, including two-way clustering (Section 8.2), and classes of deviations from a working covariance model (Section 8.3); Section 9 takes up weighted and generalized least squares, which the same theorem covers.

8.1 The general construction and theorem

Write $Z = [X\ W]$ and $\theta = (\beta', \gamma')'$ for its coefficient vector, so that $Y = Z\theta + U$.

Assumption A6 (Linear estimator and invariant residuals)**.** The estimator of $a'\beta$ satisfies $a'\widehat{\beta} - a'\beta = \omega' U$ for a vector $\omega = \omega_n(X, W)$, and the observed residual vector is $\check{u} = TY$ for a matrix $T = T_n(X, W)$ with $TZ = 0$, so that $\check{u} = TU$ does not depend on θ.

OLS is $\omega = \widetilde{X}\Gamma_n^{-1} a/n$, the influence weights z of Section 7 divided by n, and $T = M$. Weighted and generalized least squares with a known weight matrix Υ are covered as well: ω is then the influence vector ω_Υ of the weighted estimator and T is either its residual maker $T_\Upsilon = I_n - Z(Z'\Upsilon Z)^{-1} Z'\Upsilon$, idempotent but not symmetric, or M, since the two maps need not come from the same fit (Section 9).

The dependence among the errors is described by a set of index pairs. Let $\mathcal{S}_n \subseteq \{1, \dots, n\}^2$ be symmetric and contain the diagonal, let $\mathcal{H}_n^{\mathcal{S}} \subseteq \mathcal{K}_n$ be the symmetric matrices supported on $\mathcal{S}_n$, and write $\mathcal{S}(i) := \{j : (i, j) \in \mathcal{S}_n\}$ and $d_n := \max_i |\mathcal{S}(i)|$ for the largest neighbourhood, counting i itself. We keep writing $\langle \cdot, \cdot \rangle_{\mathcal{H}_n}$ and $\| \cdot \|_{\mathcal{H}_n}$ for the scaled Frobenius inner product and norm of Section 4, now applied to arrays in $\mathcal{H}_n^{\mathcal{S}}$.

Assumption A7 (Dependence structure)**.** $\mathcal{S}_n$ is (X, W)-measurable. Conditionally on (X, W), $\mathbb{E}[u_i \mid X, W] = 0$ for every i. For all disjoint $I, J \subseteq \{1, \dots, n\}$ with $(I \times J) \cap \mathcal{S}_n = \emptyset$, the vectors U_I and U_J are independent. And $\sup_i \mathbb{E}[u_i^4 \mid X, W] \le m_4$ for a finite constant m_4, almost surely.

Assumption A7 says that $\mathcal{S}_n$ is a dependency graph for U (Chen and Shao, 2004; Ross, 2011). It implies $\Sigma_n \in \mathcal{H}_n^{\mathcal{S}}$. One-way clustering is $\mathcal{S}_{1W} = \bigcup_g C_g \times C_g$, with $d_n = \sup_g N_g$; there Assumption A7

is equivalent to the independence across clusters of Assumption A1, and $\mathcal{H}_n^{\mathcal{S}_{1W}} = \mathcal{H}_n$. Heteroskedasticity is the diagonal, with $d_n = 1$. Two-way clustering, spatial or network neighbourhoods and serial bands are the cases of Section 8.2 (Conley, 1999; Kojevnikov et al., 2021).

Assumption A8 (Maintained class)**.** $\Sigma_n \in \Omega_n + \mathcal{D}_n$, where $\Omega_n \in \mathcal{H}_n^{\mathcal{S}}$ is (X, W)-measurable, possibly zero, and $\mathcal{D}_n \subseteq \mathcal{H}_n^{\mathcal{S}}$ is a linear space.

Assumption A8 splits the covariance into a known part Ω_n and an unknown deviation $\Delta := \Sigma_n - \Omega_n \in \mathcal{D}_n$, supported on the dependency graph. Usually $\Omega_n = 0$. The model of Section 3 is $\mathcal{D}_n = \mathcal{H}_n$, and smaller classes impose structure: equicorrelation is $\mathcal{D}_n = \operatorname{span}\{I_n, J - I_n\}$, with J the block-diagonal matrix of within-cluster ones, and a common within-cluster covariance level is the diagonal arrays plus $\operatorname{span}\{J - I_n\}$ (Section 8.3). A nonzero Ω_n is a part of the covariance taken as known, such as the reported sampling variances $\operatorname{diag}(s_i^2)$ of estimated outcomes in meta-regression, with $\mathcal{D}_n = \operatorname{span}\{I_n\}$ for the unknown heterogeneity. Only the component of Ω_n orthogonal to $\mathcal{D}_n$ matters; the rest can be absorbed into Δ and drops out of the estimator.

Let $\Pi_{\mathcal{D}}$ be the orthogonal projection of $\mathcal{K}_n$ onto $\mathcal{D}_n$ and define

$$\mathcal{A}_{\mathcal{D}}^T : \mathcal{D}_n \to \mathcal{K}_n, \quad \Delta \mapsto T\Delta T', \qquad (\mathcal{A}_{\mathcal{D}}^T)^*(D) = \Pi_{\mathcal{D}}\big(T'DT\big), \qquad R_{\mathcal{D}}^T := n^2\,\Pi_{\mathcal{D}}(\omega\omega'),$$

so that $V = \omega'\Sigma_n\omega = \omega'\Omega_n\omega + \langle \Delta, R_{\mathcal{D}}^T\rangle_{\mathcal{H}_n}$ for $\Sigma_n = \Omega_n + \Delta$. The *class-restricted Riesz estimator* is

$$\widehat{V}_{\mathcal{D}} \;:=\; \omega'\Omega_n\omega \;+\; \big\langle \check{u}\check{u}' - T\Omega_n T',\, D_{\mathcal{D}}\big\rangle_{\mathcal{K}_n}, \qquad D_{\mathcal{D}} := \big((\mathcal{A}_{\mathcal{D}}^T)^*\big)^+ R_{\mathcal{D}}^T. \tag{48}$$

With $T = M$, $\mathcal{D}_n = \mathcal{H}_n$ and $\Omega_n = 0$ it is $\widehat{V}_{\mathrm{Riesz}}$ of Definition 4.3. It is computed by Algorithm 1 with $(\mathcal{A}_{\mathcal{D}}^T, (\mathcal{A}_{\mathcal{D}}^T)^*)$ in place of $(\mathcal{A}, \mathcal{A}^*)$.

Assumption A9 (Identification, stability, weights and scale)**.** With probability approaching one:

(i) $R_{\mathcal{D}}^T \in \operatorname{Range}\big((\mathcal{A}_{\mathcal{D}}^T)^*\big)$;

(ii) $s_n^{T,\mathcal{D}} := 1/\sigma_{\min}^+(\mathcal{A}_{\mathcal{D}}^T)$ satisfies $s_n^{T,\mathcal{D}}\,\mu_n(d_n/n)^{3/2} \to_p 0$, and $\|T\|_{\mathrm{op}} = O_p(1)$;

(iii) $n^3 \sum_{(i,j)\in\mathcal{S}_n} \omega_i^2\omega_j^2 = O_p(d_n)$;

(iv) $\mu_n V \to_p \nu_a \in (0, \infty)$ for a deterministic $\mu_n \to \infty$.

Condition 1 (Central limit theorem)**.** $V^{-1/2}\omega'U \to_d \mathcal{N}(0, 1)$.

Part (i) is the range condition of Assumption A5(i) for the pair $(T, \mathcal{D}_n)$. Part (ii) is the stability condition of Assumption A5(ii) with $\sup_g N_g$ replaced by d_n, together with a bound on the residual map. Part (iii) says the influence weights are of order n^{-1} per observation; it holds whenever $n^3 \sum_i \omega_i^4 = O_p(1)$ (Lemma F.1). Part (iv) is the first half of Assumption A3. The partition enters only through d_n. The central limit theorem is left as a condition because the sharpest available argument differs across dependence structures: a cluster-level Lyapunov argument for one-way clustering, and Stein's method for dependency graphs (Lemma F.2). The variance of the estimator, by contrast, needs no case-specific argument: parts (i) and (ii) of Theorem 8.2 below hold without Condition 1. When the condition fails, normal and t critical values no longer apply, although the variance estimator remains exact and consistent.

Lemma 8.1 (Quadratic forms under a dependency graph). *Under Assumption A7, for every (X, W)-measurable $K \in \mathbb{R}^{n\times n}$,*

$$\operatorname{Var}\left(U'KU \mid X, W\right) \;\le\; C_0\, m_4\, d_n^2\, \|K\|_F^2,$$

where C_0 is a universal constant; one can take $C_0 = 66$.

Theorem 8.2 (General Riesz estimator). *Suppose Assumptions A6 and A8 hold.*

(i) If $R_{\mathcal{D}}^T \in \operatorname{Range}((\mathcal{A}_{\mathcal{D}}^T)^*)$, then $\mathbb{E}[\widehat{V}_{\mathcal{D}} \mid X, W] = V$ for every $\Sigma_n \in \Omega_n + \mathcal{D}_n$.

(ii) If in addition Assumptions A7 and A9 hold, then $\mu_n(\widehat{V}_{\mathcal{D}} - V) \to_p 0$ and $\widehat{V}_{\mathcal{D}}/V \to_p 1$.

(iii) If in addition Condition 1 holds, then $\widehat{V}_{\mathcal{D}}^{-1/2}(a'\widehat{\beta} - a'\beta) \to_d \mathcal{N}(0,1)$.

(iv) If $\|\Omega_n\|_{\mathcal{H}_n} = O_p(\sqrt{d_n/n})$, (ii) and (iii) hold for the LSQR iterate terminated by either criterion of (28), with a tolerance satisfying $\mu_n\tau_n d_n/n \to 0$ and $\tau_n s_n^{T,\mathcal{D}} \to_p 0$.

Exactness uses only linearity; consistency uses Lemma 8.1 and the bound $\|D_{\mathcal{D}}\|_{\mathcal{K}_n} \le s_n^{T,\mathcal{D}} \|R_{\mathcal{D}}^T\|_{\mathcal{H}_n}$.

Corollary 8.3 (One-way clustered OLS). *Under Assumptions A1 to A5, the hypotheses of Theorem 8.2, including Condition 1, hold with $\mathcal{S}_n = \mathcal{S}_{1W}$, $d_n = \sup_g N_g$, $\omega = \widetilde{X}\Gamma_n^{-1}a/n$, $T = M$, $\Omega_n = 0$, $\mathcal{D}_n = \mathcal{H}_n$ and $s_n^{M,\mathcal{H}} = s_n$. Theorems 6.1 and 6.2 are therefore parts (ii)–(iv) of Theorem 8.2.*[16]

Remark 8.4 (Robustness is bought monotonically). If $\mathcal{D}'_n \subseteq \mathcal{D}_n$ and Assumption A9(i) holds for $\mathcal{D}_n$, then $D_{\mathcal{D}}$ satisfies the constraints defining $D_{\mathcal{D}'}$, since $\Pi_{\mathcal{D}'}\Pi_{\mathcal{D}} = \Pi_{\mathcal{D}'}$, so $\|D_{\mathcal{D}'}\|_{\mathcal{K}_n} \le \|D_{\mathcal{D}}\|_{\mathcal{K}_n}$. With $T = M$ this norm governs the Gaussian variance under a working identity covariance (Proposition 4.4), so every enlargement of the class buys exactness on more covariance arrays at a weakly larger variance, from the model-based $\omega'\Omega_n\omega$ ($\mathcal{D}_n = \{0\}$) to $\widehat{V}_{\mathrm{Riesz}}$ ($\mathcal{D}_n = \mathcal{H}_n$).

8.2 Dependence that does not partition the observations

Two-way clustering along $\{C_g^1\}_{g=1}^{G_1}$ and $\{C_h^2\}_{h=1}^{G_2}$ gives $\mathcal{S}_{2W} = \bigcup_g (C_g^1 \times C_g^1) \cup \bigcup_h (C_h^2 \times C_h^2)$, which is not generated by a partition of the observations; spatial, network and banded serial supports fail to partition for the same reason. The construction (48) applies verbatim with $\mathcal{D}_n = \mathcal{H}_n^{\mathcal{S}}$, and so do Propositions 7.1 and E.1, which use only linearity of $\mathcal{A}_n$, orthogonality of $\Pi_{\mathcal{H}_n^{\mathcal{S}}}$ and $M'M = M$. For two-way clustering, write $g(i)$ and $h(i)$ for the clusters of observation i and $d_n = \max_i |C_{g(i)}^1 \cup C_{h(i)}^2| \le \sup_g N_g^1 + \sup_h N_h^2$.

Proposition 8.5 (Two-way clustering). *Let $\omega = \widetilde{X}\Gamma_n^{-1}a/n$ and $T = M$, and suppose:*

(a) conditionally on (X, W), $\mathcal{S}_{2W}$ satisfies Assumption A7; its independence requirement holds, for instance, when $u_i = \varphi_i(\alpha_{g(i)}, \beta_{h(i)}, \varepsilon_i)$ with $\{\alpha_g\}$, $\{\beta_h\}$ and $\{\varepsilon_i\}$ mutually independent given (X, W);

(b) Assumption A4 holds and $n^{-1}\sum_i \|\tilde{x}_i\|^4 = O_p(1)$;

[16] The one-way case verifies Condition 1 by a cluster-level Lyapunov argument, Step 1 of the proof of Theorem 6.1. The Stein bound of Lemma F.2 applies as well, requiring only fourth moments and $\mu_n^{3/4} \sup_g N_g/n \to 0$. Since $(2+\lambda)/(2+2\lambda) \ge 3/4$ exactly when $\lambda \le 1$, the Stein route is weaker for $\lambda \le 1$; for $\lambda > 1$ the Lyapunov route allows larger clusters at the cost of more moments.

(c) $\mu_n V \to_p \nu_a \in (0,\infty)$ *for a deterministic* $\mu_n \to \infty$*, and* $\mu_n^{3/4} d_n/n \to 0$*.*

Let $\mathcal{D}_n \subseteq \mathcal{H}_n^{\mathcal{S}_{2W}}$ *be linear and* $\Omega_n \in \mathcal{H}_n^{\mathcal{S}_{2W}}$ *be* (X,W)*-measurable, with* $\Sigma_n \in \Omega_n + \mathcal{D}_n$*, and let Assumption A9(i)–(ii) hold for* $(M,\mathcal{D}_n)$*. Then Condition 1 holds, and so do conclusions (i)–(iii) of Theorem 8.2; conclusion (iv) holds if in addition* $\|\Omega_n\|_{\mathcal{H}_n} = O_p(\sqrt{d_n/n})$*. This covers the fully robust two-way estimator,* $\mathcal{D}_n = \mathcal{H}_n^{\mathcal{S}_{2W}}$ *and* $\Omega_n = 0$*.*

Remark 8.6 (Reading the two-way rates). The rate μ_n of V enters both requirements of Proposition 8.5: stability, $s_n \mu_n (d_n/n)^{3/2} \to 0$ with $s_n := s_n^{M,\mathcal{D}}$, for consistency, and $\mu_n^{3/4} d_n/n \to 0$ for Condition 1. Suppose the clusters are balanced, $\sup_g N_g^1 \lesssim n/G_1$ and $\sup_h N_h^2 \lesssim n/G_2$, and let $G := \min(G_1, G_2)$, so that $d_n/n \lesssim 1/G_1 + 1/G_2 \leq 2/G$. As in the asymptotic theory of two-way clustering (Cameron et al., 2011; Davezies et al., 2021; MacKinnon et al., 2021), μ_n depends on which dimensions carry a common component.

- *Both dimensions carry a common shock.* Then $V \asymp G^{-1}$ and $\mu_n \asymp G$. The central limit condition, $\mu_n^{3/4} d_n/n \lesssim 2G^{-1/4}$, holds automatically; this is the non-degenerate case, in which the literature uses $t_{\min(G_1,G_2)-1}$ critical values (Cameron et al., 2011) or the wild bootstrap (MacKinnon et al., 2021). Stability requires $s_n = o(G^{1/2})$, the one-way requirement with G clusters. When the second dimension is time and its common shocks are serially correlated, the support widens beyond $\mathcal{S}_{2W}$ to pairs in nearby periods, and d_n grows with the lag length (Chiang et al., 2024). With two-way fixed effects at the clustering levels, additive shocks lie in $\mathrm{col}(W)$ and contribute nothing to V, so the design can be degenerate although both dimensions carry a common shock.

- *Only the first dimension carries a common shock.* Then $\mu_n \asymp G_1$. The central limit condition becomes $G_1^{-1/4} + G_1^{3/4}/G_2 \to 0$, so the other dimension must have many clusters, and with $G_2 \gtrsim G_1$ stability again requires $s_n = o(G_1^{1/2})$.

- *No common shock.* Then $\mu_n = n$. The central limit condition asks for $d_n = o(n^{1/4})$, and stability for $s_n d_n^{3/2} = o(n^{1/2})$, which is weaker when s_n is bounded. Both are conservative, since the dependency graph declares dependence the errors do not use.

- *Product errors,* $u_i = \alpha_{g(i)}\beta_{h(i)}$ with $G_1 = G_2 = G$. Then $V \asymp G^{-2}$, so $\mu_n \asymp G^2$, and both requirements fail: $\mu_n^{3/4} d_n/n \asymp G^{1/2}$ and $s_n \mu_n (d_n/n)^{3/2} \asymp s_n G^{1/2}$ diverge. The central limit theorem should fail, since $\omega' U$ is then a bilinear form in the two sets of shocks, with the non-Gaussian limit found by Menzel (2021). The estimator remains exactly unbiased, but Theorem 8.2 does not establish its consistency.

Remark 8.7 (Serial and spatial bands). With observations ordered in time and $\mathcal{S}_n$ the pairs at most m periods apart, Theorem 8.2 covers m-dependent errors with $d_n = 2m+1$, the setting of kernel and HAC variance estimators (Newey and West, 1986; Andrews, 1991). Under the weight and scale conditions of Lemma F.2, that lemma gives Condition 1 when $\mu_n^{3/4}(2m+1)/n \to 0$; with $\mu_n \asymp n$, this is $m = o(n^{1/4})$. Spatial distance bands are analogous (Conley, 1999; Kelejian and Prucha, 2007), as is the panel support of Driscoll and Kraay (1998), all pairs at most m periods apart whatever their units. Two cautions apply.

- *Exactness holds only on the band.* For errors that are weakly dependent but not m-dependent, the estimator is unbiased for the part of Σ_n inside the band. The truncation bias $\omega'(\Sigma_n - \Pi_{\mathcal{H}^{\mathcal{S}}}\Sigma_n)\omega$ of any kernel-type estimator remains, together with the spillover of that out-of-band covariance through residualization; what the construction removes is the residualization bias of the part inside the band. The truncation bias is the object of the bandwidth literature, and the variability of the variance estimator that of the fixed-b literature (Kiefer and Vogelsang, 2005; Lazarus et al., 2018), both of which the construction complements rather than replaces.

- *The range condition caps the bandwidth.* In the location model, $Z = \iota$, the kernel of $\mathcal{A}_n$ consists of the matrices $\iota b' + b\iota'$, and none of them except zero is banded once $m \leq (n-3)/2$, so the condition holds for every practical bandwidth. Additional regressors enlarge the kernel and can lower the cap; the dual stopping rule reports a failure.

Two features are specific to supports that do not partition. First, no single block-local correction exists: no one partition supplies blocks M_{gg} whose inversion accounts for all the dependence, and the scalar degrees-of-freedom combinations in common use for two-way clustering (Cameron et al., 2011; Thompson, 2011) address none of the projection bias. Second, the range condition can bind.

Remark 8.8 (Counting bound on a general support; Chen, 2026, Proposition 7.1). Since $\mathcal{S}$ is symmetric and contains the diagonal, $\dim(\mathcal{H}_n^{\mathcal{S}}) = \frac{1}{2}(|\mathcal{S}| + n)$. The map $\Delta \mapsto M\Delta M$ takes values in the symmetric matrices with row and column spaces in $\mathrm{col}(M)$, of dimension $\frac{1}{2}r(r+1)$ with $r = \mathrm{rank}(M) = n - \mathrm{rank}([X\ W])$, and an injective linear map cannot increase dimension. Hence $\mathrm{Ker}(\mathcal{A}_{\mathcal{D}}^M) = \{0\}$ requires $\dim(\mathcal{D}_n) \leq \frac{1}{2}r(r+1)$, which for the fully robust class $\mathcal{D}_n = \mathcal{H}_n^{\mathcal{S}}$ reads $\frac{1}{2}(|\mathcal{S}| + n) \leq \frac{1}{2}r(r+1)$.

For one-way clustering $|\mathcal{S}| = \sum_g N_g^2$ and the bound is slack by a wide margin. For two-way clustering $|\mathcal{S}|$ is of order $n^2/G_1 + n^2/G_2$, and the bound tightens as the number of controls grows, so the range condition must be checked rather than assumed. When the bound fails, $\mathrm{Ker}(\mathcal{A}_{\mathcal{D}}^M) \neq \{0\}$, and the fully robust target is recoverable only if it is orthogonal to the kernel. The dual stopping rule of (28) checks this at no cost; in the two-way experiment of Section 10 it classified every design correctly. When the target is not orthogonal to the kernel, no unbiased estimator based on residual second moments exists, from the full or the restricted system; the responses are to shrink the maintained class, as in Section 8.3, or to use a procedure that does not estimate the variance.

8.3 Deviation classes: from model-based to robust

The fully robust estimator $\widehat{V}_{\mathrm{Riesz}}$ is exactly unbiased for every $\Sigma_n \in \mathcal{H}_n$ (Theorem 4.6), and it pays for that generality in sampling variance: in the simulations its relative root mean squared error exceeds that of CR2, whose bias buys it a smaller variance. Theorem 8.2 with a smaller class resolves the trade without giving up exactness on the class. Three classes, all containing arbitrary heteroskedasticity, illustrate the range:

$$\mathcal{D}^{\mathrm{het}} = \{\text{diagonal}\}, \qquad \mathcal{D}^{\mathrm{het}+\rho} = \mathcal{D}^{\mathrm{het}} \oplus \mathrm{span}\{J - I_n\}, \qquad \mathcal{D}^{\mathrm{het}+\rho_g} = \mathcal{D}^{\mathrm{het}} \oplus \mathrm{span}\{J_g - I_g\}_{g=1}^{G},$$

with J as above, J_g the $n \times n$ indicator of pairs within cluster g and I_g the diagonal indicator of its members, so that $J = \sum_g J_g$ and $I_n = \sum_g I_g$. The second is exact whenever the errors are

heteroskedastic with a common within-cluster covariance level, the third whenever that level varies across clusters, and neither requires the level to be known: the level is part of the deviation Δ, and any Ω_n in the class cancels from (48), so the estimator is the one with $\Omega_n = 0$. By Remark 8.4 each has a weakly smaller representer, and hence a weakly smaller variance under homoskedastic Gaussian errors, than $\widehat{V}_{\mathrm{Riesz}}$, and in the simulations $\mathcal{D}^{\mathrm{het}+\rho}$ has a smaller relative root mean squared error than CR2 while remaining unbiased on its class. The reduction matters most with few clusters, where the fully robust representer is too variable for the degrees-of-freedom correction of Section 5.2 to work well and the restricted one is not.

The same construction applies on supports that do not partition the observations. For two-way clustering, write S_k for the $n \times n$ indicator of sharing a cluster on dimension k, with ones on the diagonal. The analogue of $\mathcal{D}^{\mathrm{het}+\rho}$ is

$$\mathcal{D}_{2W}^{\mathrm{het}+\rho} = \mathcal{D}^{\mathrm{het}} \oplus \operatorname{span}\{S_1 - I_n,\ S_2 - I_n\} \subseteq \mathcal{H}_n^{S_{2W}}, \tag{49}$$

heteroskedasticity with one covariance level per dimension, a pair sharing both clusters carrying the sum of the two. It is exact for the two-way random-effects structure $u_i = \sqrt{\rho_1}\, e_{g(i)} + \sqrt{\rho_2}\, e'_{h(i)} + \varepsilon_i$ with heteroskedastic ε_i, without knowledge of ρ_1 or ρ_2, and it is covered by Proposition 8.5, which allows any linear $\mathcal{D}_n \subseteq \mathcal{H}_n^{S_{2W}}$. Its dimension is $n+2$, so the counting bound of Remark 8.8 becomes $n+2 \le \frac{1}{2}r(r+1)$, which holds whenever $r \ge \sqrt{2n}+1$. The bound is therefore rarely binding for this class. It is the natural response when the stopping rule reports that the fully robust two-way target is not recoverable, and in Section 10 it remains identified where that target is not.

Two practical points. First, the covariance level should be estimated *inside* the class, as part of the same representer system, and not plugged in from a separate moment of the OLS residuals: those residuals carry the projection spillover, and a plug-in level reintroduces the bias the construction removes (the class $\mathcal{D}^{\mathrm{het}}$ with a plug-in equicorrelation as Ω_n is biased by -13% to -16% in Experiment I). Second, a class is a maintained assumption: the estimator is exact only when Σ_n lies in it. Outside its class the bias of $\mathcal{D}^{\mathrm{het}+\rho}$ is small in the median but reaches -12% and -21% in the worst designs of Experiment I (Section H.2.4), so a restricted class trades robustness for precision, and the fully robust estimator remains the benchmark against which to check it.

9 Weighted and Generalized Least Squares

With known weights a weighted fit satisfies Assumption A6, so Theorem 8.2 applies unchanged. What changes is the residual maker, which now projects obliquely, so the full and restricted systems written in the weighted residuals no longer coincide. They differ in two ways. Typically both identify the target, but they are different estimators, with different *noise*: the oblique map spreads the identifying content and the noise differently across the residual products, and each system weights them differently. With special weights that couple clusters, the restricted system can also lose *information* that only the cross-cluster products carry (Section 9.1). The remedy is a change of coordinates. The weighted and OLS residuals determine each other, and in OLS coordinates the two systems coincide again. Riesz-M writes the full weighted system in these coordinates, so it uses all the information while keeping the noise under control (Section 9.2). When the weights model the covariance, whitened coordinates are more efficient (Section 9.3).

9.1 Two systems in the weighted residuals

Let Υ be an (X, W)-measurable positive definite weight matrix and consider $\widehat{\theta}_\Upsilon = (Z'\Upsilon Z)^{-1}Z'\Upsilon Y$. This covers weighted least squares (diagonal Υ: sampling, design or analytic weights; DuMouchel and Duncan, 1983) and GLS with a known or working covariance model ($\Upsilon = \Omega^{-1}$, for example random-effects weights; Liang and Zeger, 1986; Wooldridge, 2003). Its influence weights are $\omega_\Upsilon = \Upsilon\widetilde{X}_\Upsilon(\widetilde{X}'_\Upsilon\Upsilon\widetilde{X}_\Upsilon)^{-1}a$, with $\widetilde{X}_\Upsilon$ the Υ-orthogonal residual of X on W; its residual maker is $T_\Upsilon = I_n - Z(Z'\Upsilon Z)^{-1}Z'\Upsilon$, idempotent but not symmetric; and the target is $V_\Upsilon = \omega'_\Upsilon\Sigma\omega_\Upsilon$. The weights are given; the question is how to estimate the variance of the estimator they define.

As in Section 7, estimating V_Υ from the weighted residuals leads to two systems. Every quadratic form $n^{-2}\check{u}'D\check{u}$ equals $U'KU$ with $K = n^{-2}T'_\Upsilon DT_\Upsilon$, and it is unbiased for V_Υ at every $\Sigma \in \mathcal{H}_n$ if and only if D solves the *full weighted system*

$$\Pi_{\mathcal{H}}\big(T'_\Upsilon DT_\Upsilon\big) = n^2\,\Pi_{\mathcal{H}}\big(\omega_\Upsilon\omega'_\Upsilon\big). \tag{50}$$

Its minimum-norm solution, the analogue of Definition 4.3, gives Riesz-F(T). The *restricted* system keeps only the within-cluster blocks of $\mathbb{E}[\check{u}\check{u}' \mid X, W] = T_\Upsilon\Sigma T'_\Upsilon$. Its adjoint maps $\{d_h\}$ to $\{\sum_h T'_{\Upsilon,hg}d_hT_{\Upsilon,hg}\}_g$, with the transpose on $T_{\Upsilon,hg}$: the OLS form $\sum_h M_{gh}d_hM_{hg}$ of (44) relies on $M_{gh} = M'_{hg}$, which fails for T_Υ. Its minimum-norm solution gives Riesz-P(T). Under OLS the two solutions give the same estimator in every realization (Corollary 7.2), because $M'\widehat{U} = \widehat{U}$ collapses the full estimator onto the within-cluster products. With $T'_\Upsilon\check{u} \neq \check{u}$ that collapse fails, and Riesz-F(T) and Riesz-P(T) are different estimators with the same expectation.

The restricted system can also lose *information*, though only with special weights. Under OLS the restricted operator is the Gram operator of the full one (Proposition 7.1), which uses $M'M = M$; T_Υ is orthogonal in the Υ-norm $\langle a, b\rangle_\Upsilon = a'\Upsilon b$ instead, so $T'_\Upsilon T_\Upsilon \neq T_\Upsilon$. There can then be directions Δ with $[T_\Upsilon\Delta T'_\Upsilon]_{gg} = 0$ for all g but $[T_\Upsilon\Delta T'_\Upsilon]_{gh} \neq 0$ for some $g \neq h$. If the target loads on such a direction, no weighting of the within-cluster weighted moments, including the cluster-robust estimators computed from weighted residuals, is unbiased for every $\Sigma \in \mathcal{H}_n$, while the full array identifies it whenever the OLS range condition holds (Proposition 9.1). This requires weights that couple clusters: when the weights respect the partition, whitening turns the Υ-norm into the Euclidean one within each cluster and nothing is lost (Proposition 9.4). Even then it needs special weights: in Example G.1 the loss occurs on a line in the space of weights, and in Experiment J it never occurs.

9.2 Changing coordinates: Riesz-M

Whether the within-cluster moments carry all the information depends on the coordinates in which the residuals are written. The next result shows that the weighted and OLS residuals are two coordinate systems for the same information.

Proposition 9.1 (The OLS residuals are sufficient)**.** *Let T be any residual map with $TZ = 0$. Then $T = TM$, so the T-residuals are a function of the OLS residuals, $\check{u} = T\widehat{U}$. Every statistic of $\check{u}$ is therefore a statistic of $\widehat{U}$, and* $\mathrm{Range}\big((\mathcal{A}^T)^*\big) \subseteq \mathrm{Range}(\mathcal{A}_n^*)$*, with equality when* $\mathrm{Ker}(T) = \mathrm{col}(Z)$*. For the weighted fit, $T_\Upsilon M = T_\Upsilon$ and $MT_\Upsilon = M$ for every positive definite Υ. The two residual vectors then*

determine each other, $\check{u} = T_\Upsilon \widehat{U}$ *and* $\widehat{U} = M\check{u}$*, and* $\mathrm{Range}\left((\mathcal{A}^{T_\Upsilon})^*\right) = \mathrm{Range}(\mathcal{A}_n^*)$*. By Proposition 7.1, a covariance functional recoverable from the residuals of any such fit is recoverable from the within-cluster moments of the OLS residuals.*

Proposition 9.2 (The OLS organization of the full weighted system)**.** *Among the solutions of* (50)*, the one whose kernel* $K = n^{-2} T_\Upsilon' D T_\Upsilon$ *has minimum Frobenius norm gives the estimator*

$$\widehat{V}_\Upsilon^M := \frac{1}{n^2}\sum_{g=1}^G \widehat{U}_g' H_g^M \widehat{U}_g = \frac{1}{n^2}\sum_{g=1}^G (M\check{u})_g' H_g^M (M\check{u})_g, \qquad \sum_{k=1}^G M_{gk} H_k^M M_{kg} = n^2 \omega_{\Upsilon,g}\omega_{\Upsilon,g}', \tag{51}$$

the OLS Riesz estimator with the weighted target ω_Υ*, which we call Riesz-M. It is* (48) *with* $T = M$*,* $\omega = \omega_\Upsilon$ *and* $\mathcal{D}_n = \mathcal{H}_n$*. It is exactly unbiased for* V_Υ *under the range condition of Assumption A5(i) with target* $n^2\{\omega_{\Upsilon,g}\omega_{\Upsilon,g}'\}$*. Under homoskedastic Gaussian errors it is the most precise solution of* (50)*.*

In OLS coordinates the results of Sections 4 to 7 — Gram equivalence, the cluster-robust form of Corollary 7.2, the conditioning of Proposition E.1 — hold as stated: the operator, and hence its conditioning, does not depend on Υ; only the target does. Riesz-M solves the same system as Riesz-F(T) and uses the same cross-cluster products of $\check{u}$, through $M\check{u}$; the two differ only in the coordinates in which the norm is minimized, as the following remark makes precise.

Remark 9.3 (Riesz-F(T) is a working-covariance estimator). Write every unbiased estimator as $U'KU$ with $K = n^{-2}T_\Upsilon' D T_\Upsilon$. The kernel K_M of Riesz-M is the minimum-Frobenius-norm element of the affine set of unbiased kernels, so the kernel of Riesz-F(T) is $K_F = K_M + (K_F - K_M)$ with $\langle K_M, K_F - K_M\rangle = 0$: Riesz-F$(T)$ is Riesz-M plus a mean-zero term, uncorrelated with it under homoskedastic Gaussian errors, and never more efficient there. More precisely, K_F is the kernel of the efficient estimator of Proposition 9.5 below for the working covariance $\Omega^* := (T_\Upsilon' T_\Upsilon)^+ + \Pi_Z$, which matches neither what the weights are for nor any model of Σ.

We therefore propose Riesz-M when the weights define the estimand rather than model the covariance (Solon et al., 2015): its efficiency does not rest on the weights describing the errors. Leading examples are sampling and survey weights (DuMouchel and Duncan, 1983), inverse-propensity and balancing weights (Hirano et al., 2003), and kernel weights in regression discontinuity designs (Calonico et al., 2014).

9.3 Weights that model the covariance

Riesz-M is efficient near homoskedasticity. When the weights model Σ and are known — analytic weights for group means, inverse-variance weights with known sampling variances, or a random-effects or GLS model fixed in advance — another set of orthogonal coordinates is available: the whitened ones.

Proposition 9.4 (Whitening)**.** *Let* $\Upsilon = \mathrm{diag}(\Upsilon_1, \ldots, \Upsilon_G)$ *be block-diagonal and conformable, and* $L = \Upsilon^{1/2}$ *blockwise. Then* $L T_\Upsilon L^{-1} = \widetilde{M} := I_n - \Pi_{LZ}$*, the orthogonal residual maker of the regression of* LY *on* LZ*. Consequently the estimator*

$$\widehat{V}_\Upsilon^\Upsilon := \frac{1}{n^2}\sum_{g=1}^G \check{u}_g' d_g \check{u}_g, \qquad \sum_h T_{\Upsilon,hg}' \, d_h \, T_{\Upsilon,hg} = n^2 \omega_{\Upsilon,g}\omega_{\Upsilon,g}',$$

built from the within-cluster moments of the weighted residuals, equals the OLS Riesz estimator of the whitened model — residual maker $\widetilde{M}$, residuals $\widetilde{u} = L\check{u}$, target $\widetilde{\omega} = L^{-1}\omega_\Upsilon$ — in every realization, and its range condition is that of $\mathcal{A}_n$ with target $n^2\{\omega_{\Upsilon,g}\omega'_{\Upsilon,g}\}$. In particular the within-cluster weighted moments lose no identifying direction.

With weights that respect the partition, including all diagonal weights, Riesz-P(T) is therefore the whitened estimator Riesz-Υ. It estimates the same V_Υ as Riesz-M; the two differ only in precision.

Which coordinates to choose is a choice of *working covariance*: a guess Ω_w at Σ used only to choose among unbiased estimators, which affects their precision but never their validity.

Proposition 9.5 (Working-covariance optimality)**.** *Let Ω_w be a positive definite working covariance and $W_\Omega := (M\Omega_w M)^+$. Among all quadratic forms in the residuals of any fit satisfying Assumption A6 that are unbiased for V_Υ at every $\Sigma \in \mathcal{H}_n$, when such forms exist, that is, under the range condition of Assumption A5(i) with target $n^2\{\omega_{\Upsilon,g}\omega'_{\Upsilon,g}\}$, the one with the smallest variance under Gaussian errors with covariance Ω_w is*

$$\widehat{V}^{\Omega}_{\Upsilon} := (W_\Omega \widehat{U})'\Lambda_\Omega(W_\Omega \widehat{U}) \qquad \Lambda_\Omega \in \mathcal{H}_n, \quad \Pi_{\mathcal{H}}(W_\Omega \Lambda_\Omega W_\Omega) = \Pi_{\mathcal{H}}(\omega_\Upsilon \omega'_\Upsilon).$$

If Ω_w is conformable, $\widehat{V}^{\Omega}_{\Upsilon}$ is the OLS Riesz estimator of the $\Omega_w^{-1/2}$-whitened model: $\widehat{V}^{M}_{\Upsilon}$ is the case $\Omega_w = I_n$, and $\widehat{V}^{\Upsilon}_{\Upsilon}$ the case $\Omega_w = \Upsilon^{-1}$, for which $W_\Omega = T'_\Upsilon \Upsilon T_\Upsilon$. Under infeasible GLS, $\Upsilon = \Sigma^{-1}$, the latter is the minimum-variance invariant quadratic unbiased estimator under Gaussian errors.

This is the MINQUE principle of Rao (1970, 1971) applied to the variance functional. It shares the working-model logic of GEE (Liang and Zeger, 1986) and CR2 (Bell and McCaffrey, 2002; Pustejovsky and Tipton, 2018), with one difference: CR2 is unbiased only when its working model is correct, whereas $\widehat{V}^{\Omega}_{\Upsilon}$ is unbiased for every $\Sigma \in \mathcal{H}_n$. Riesz-Υ remains exactly unbiased when the weights do not model Σ, the property that motivates pairing weighted least squares with robust inference (Romano and Wolf, 2017). When such weights couple clusters, as spatial or serial GLS weights do, the efficient estimator is $\widehat{V}^{\Omega}_{\Upsilon}$ with $\Omega_w = \Upsilon^{-1}$ (Experiment J, Section H.5.5).

The working covariance also explains the difference in noise between the full and restricted systems of Section 9.1. Each minimizes a norm, and each norm is efficient for an implicit working covariance: with conformable weights, Υ^{-1} for Riesz-P(T), and for Riesz-F(T) the covariance Ω^* of Remark 9.3, which equals I_n when the weights are constant and stays close to it when they vary little. Riesz-F(T) thus treats the residual products as roughly equally noisy. They are not when variances differ: the variance of $\check{u}_i\check{u}_j$ scales with the variances of the two residuals, so the products of high-variance observations are far noisier, and weighting all products alike lets them dominate. When the weights model Σ, Riesz-P(T) down-weights these products by the right amount and Riesz-F(T) does not. When the weights are unrelated to Σ and Σ is close to homoskedastic, the products are about equally noisy and the ranking reverses.

Proposition 9.6 (Full against restricted, by the role of the weights)**.** *Let Υ be conformable, let the errors be Gaussian, and suppose $\mathcal{A}_n$ is injective. Write V_F and V_P for the variances of Riesz-F(T) and Riesz-P(T).*

(i) (Weights that model Σ.) If $\Sigma \propto \Upsilon^{-1}$, then $V_P \leq V_F$.

(ii) (Weights unrelated to Σ.) If $\Sigma = \sigma^2 I_n$ and $\Upsilon = \exp(\varepsilon E)$ for a fixed symmetric conformable E, then

$$V_P - V_F \;=\; 2\sigma^4\varepsilon^2 \,\|\dot{K}[E]\|_F^2 + O(\varepsilon^3),$$

where $\dot{K}[E]$ is the derivative at $\Omega = I_n$, in the direction $-E$, of the kernel of the efficient estimator for the working covariance Ω (Proposition 9.5), evaluated at the OLS target ω. In particular $V_F < V_P$ for all small $\varepsilon > 0$ whenever $\dot{K}[E] \neq 0$.

Part (i) is exact and part (ii) is local in the dispersion of the weights. The simulations agree in both directions, well beyond small dispersion. With weights that do not model Σ, Riesz-F(T) covers better in 79% of the cells of Experiment K, by up to 27 percentage points; with correct GLS weights, Riesz-P(T) keeps intervals near the oracle length while those of Riesz-F(T) reach 32 times it (table 8).

With conformable weights, Riesz-M, Riesz-Υ and Riesz-F(T) share one limit theory: under the conditions of Section 6, restated for the weighted design, each satisfies the conclusions of Theorem 8.2 (Corollary G.3 in Section G), and by Propositions 9.1 and 9.4 their range conditions coincide.

Estimated weights. With estimated weights $\widehat{\Upsilon} = \Upsilon(\widehat{\vartheta})$, the residual maker $T_{\widehat{\Upsilon}}$ and the influence weights $\omega_{\widehat{\Upsilon}}$ are themselves estimated, and Assumption A6 fails: the point estimator is no longer linear in U, and the projection geometry that the representer inverts depends on the errors whose variance it estimates. Two routes suggest themselves: a uniform version of Assumption A9(ii) near a pseudo-true ϑ^*, with the representer Lipschitz in ϑ, as in the theory of feasible weighted least squares (Carroll, 1982; Robinson, 1987); or cross-fitting the weights across clusters, which fixes the geometry conditionally and returns the problem to Corollary G.3. Feasible GLS with cluster-robust inference is recommended in panel and difference-in-differences applications (Hansen, 2007b; Brewer et al., 2018), and feasible GLS with estimated cross-sectional and serial covariances has been developed for panels (Bai et al., 2021), so the question matters in practice. We leave it, together with the comparison of Proposition 9.6(i) under estimated weights, to future work.

10 Simulations

Each theoretical result states a condition; what it does not state is a magnitude. This section measures the magnitudes that matter for practice, in five exhibits. Table 2 lists them together with the claim each tests. Further experiments are listed at the end and reported in Section H.

10.1 Design

Base model. The designs follow Cattaneo et al. (2018) and Djogbenou et al. (2019). For unit i in cluster g,

$$y_{gi} = x_{gi}\beta + w_{gi}'\gamma + u_{gi}, \qquad \beta = 1, \quad \gamma = 0, \tag{52}$$

where x_{gi} is the scalar regressor of interest and $w_{gi} \in \mathbb{R}^\ell$ includes an intercept; the target (15) does not depend on (β, γ). Errors combine a cluster shock with an idiosyncratic component,

$$u_{gi} = \sqrt{\rho_u}\,\varepsilon_g + \sqrt{1-\rho_u}\,\varepsilon_{gi}, \qquad \varepsilon_g \overset{iid}{\sim} \mathcal{N}(0,1), \quad \varepsilon_{gi} \mid X, W \sim \mathcal{N}\big(0, \sigma_\varepsilon^2(x_{gi}, w_{gi})\big),$$

with

$$\sigma_\varepsilon^2(x,w) = c_\varepsilon\big[1 + \big(\operatorname{clip}(x) + s(w)\big)^2\big]^\vartheta, \qquad s(w) = \operatorname{clip}\Big(\frac{w^{(2)} + \cdots + w^{(\ell)}}{\sqrt{\ell - 1}}\Big),$$

where $\operatorname{clip}(a) = \max\{-2, \min\{a, 2\}\}$. The switch $\vartheta \in \{0, 1\}$ turns heteroskedasticity off and on, and c_ε normalizes the average of σ_ε^2 over the design to one. The variance is a function of the regressors, as in Cattaneo et al. (2018), so with $\vartheta = 1$ it varies with the leverage, which makes the contrast term of (5) nonzero. The control index $s(w)$ excludes the intercept and is scaled by the number of controls, so its variance does not grow with ℓ; without the scaling, ℓ/n would move the strength of the heteroskedasticity together with the number of controls, and the two could not be told apart. Both x and the index are capped, so σ_ε^2 is bounded above and away from zero and the conditional moment bound of Assumption A2 holds; with normal controls an uncapped index would be unbounded, and so would the error variance.

Controls and regressor. Controls are $w_{gi}^{(j)} = \sqrt{\rho_w}\, z_g^{(j)} + \sqrt{1 - \rho_w}\, z_{gi}^{(j)}$ for $j \geq 2$, where the draws z are i.i.d. and follow one of three laws: (i) standard normal; (ii) uniform on $[-1, 1]$; or (iii) rare indicators, $z = \mathbf{1}\{\tilde{z} \geq 2.5\}$ with $\tilde{z}$ standard normal, as in Exhibit 2.[17] The regressor of interest is $x_{gi} = \sqrt{\rho_x}\, v_g + \sqrt{1 - \rho_x}\, v_{gi}$ with $v_{gi} \mid W \sim \mathcal{N}(0, c_v[1 + s(w_{gi})^2]^\vartheta)$. The share ℓ/n fixes how much leverage there is, and the cluster components ρ_w and ρ_x affect where it sits. They give the regressors a part that is constant within a cluster, which concentrates within-cluster leverage along ι_g and raises the level factor κ_n of Section 2 above p/n, toward p/G (Exhibit 5). Exhibit 1 has three settings, with $\ell/n \in \{0.01, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50\}$ and $\vartheta \in \{0, 1\}$ in each:

- *Scalar:* $G = n = 1000$.
- *Clusters for errors only:* $n = 1000$, $G = 200$, $\rho_u \in \{0.3, 0.6\}$ and $\rho_w = \rho_x = 0$.
- *Clusters for errors and regressors:* as before with $\rho_u \in \{0.3, 0.6\}$ and $\rho_w = \rho_x = 0.6$. Here the within-cluster block of Σ meets within-cluster leverage concentrated along ι_g, the clustered case of (46). This setting is also run at $\rho_u = 0$, where the errors are independent and the scalar estimators ($G = n$) apply.

Cluster sizes. Cluster sizes follow Djogbenou et al. (2019):

$$N_g \;\propto\; \exp(\eta_N\, g/G), \qquad g = 1, \ldots, G,$$

rounded by largest remainders so that they sum to n. The parameter η_N controls the imbalance: $\eta_N = 0$ gives equal sizes, and the largest cluster is about e^{η_N} times the smallest. The exhibits use $\eta_N = 1$, so that with $n = 1000$ and $G = 200$ the sizes range from 3 to 8, with $G = 50$ from 12 to 31, and with $G = 10$ from 61 to 151. Unequal sizes can matter in two ways. They make the clusters' shares of the variation in the direction of interest unequal, which lowers the effective number of clusters and can raise κ_n (Section 2); and, with heteroskedasticity, they can place the high-variance

[17] Rare indicators follow Cattaneo et al. (2018). Each equals one with probability about 0.6%, so it is nonzero for only a handful of observations. An indicator that is nonzero for a single observation fits it exactly, $h_{ii} = 1$; one that is nonzero within a single cluster puts a vector supported on that cluster in $\operatorname{col}([X\ W])$, so M_{gg} is singular (Proposition 7.4). These designs are therefore where the leverage corrections cannot be evaluated.

observations where the leverage is, which raises the contrast term of (5). With $\eta_N = 1$ and many clusters both effects are small. Experiment L isolates them at $(n, G) = (1000, 100)$ with $\eta_N \in \{0, 2, 4\}$, which gives equal sizes of 10, sizes from 3 to 23, and sizes from 1 to 40 (Section 10.2).

Estimators. In the scalar setting we compare HC1, which rescales HC0 by $n/(n-p)$, HC2, HC3, the exact inversion HCK of Cattaneo et al. (2018), and the Riesz estimator. In the clustered settings we compare their cluster counterparts CR1, CR2 and CR3; and the leave-cluster-out estimator LCO of Anatolyev and Ng (2026). CR2 is paired with the Bell and McCaffrey (2002) degrees of freedom (CR2-BM), as is standard. HC0 and CR0 are reported in table 3 only, since they fall well below the others. A suffix names the reference distribution of the interval and never changes $\widehat{V}$: without a suffix the critical value is normal; -t_ν uses the Satterthwaite degrees of freedom of (31), computed from the estimator's own representer; and -t_{G-1} uses $G-1$ degrees of freedom ($\min(G_1, G_2)-1$ with two-way clustering), the usual practice with few clusters (Bester et al., 2011; Colin Cameron and Miller, 2015). This convention holds in every table and figure. The t_ν correction is the construction behind CR2-BM, a moment-matching approximation derived under Gaussian errors with working covariance I_n. It has no asymptotic justification of its own, so the evidence for it, as for CR2-BM, is the simulations. Table 14 in Section H lists every estimator and what it requires.

Replication scheme. The target (15) is a variance conditional on (X, W), so the simulation has three levels. A *configuration* is one setting of the parameters, such as $\ell/n = 0.2$, $\rho_w = \rho_x = 0.6$ and $\vartheta = 1$. A *design* is one draw of (X, W) and Σ from a configuration; it is the unit of the simulation, and within it V is computed exactly. Each configuration is drawn as S independent designs, and each design receives R error vectors. In some experiments the parameters are themselves drawn afresh for every design, so that each design is its own configuration. Coverage is computed within each design and averaged over designs, which gives the unconditional coverage that Cattaneo et al. (2018) report. It is the same quantity they estimate by drawing (X, W, U) afresh in every replication; fixing the design in addition gives the exact conditional bias and the variation across designs, and lets the Riesz system be solved once per design. Bias and standard deviation are conditional on the design, and their distribution over designs can be heavy-tailed, since a few nearly singular or ill-conditioned designs produce extreme values; we therefore report medians over designs, which describe a typical design rather than an average dominated by outliers, and report the worst design where the tails matter; table 17 gives the 10th and 90th percentiles for Exhibit 1. Table 16 in Section H.1.1 lists S and R for every experiment.

Metrics. Several metrics are computed for each design, conditional on (X, W), and the estimators are compared on bias, variability and coverage.

- *Bias.* Every estimator except LCO and those with weights estimated in Experiment H is a quadratic form $\widehat{V} = U'KU$ in the errors, with K fixed given the design. Its conditional mean is therefore known exactly, $\mathbb{E}[\widehat{V} \mid X, W] = \operatorname{tr}(K\Sigma)$, and we report the exact mean bias $\operatorname{tr}(K\Sigma)/V - 1$, which carries no Monte Carlo noise. For LCO and Experiment H the mean bias is the Monte Carlo mean.
- *Variability.* The standard deviation of $\widehat{V}/V$ is computed over the Monte Carlo draws.

- *Coverage.* Coverage is that of the nominal 95% interval. A draw with $\widehat{V} \leq 0$ gives no interval and counts as a failure to cover; excluding such draws would overstate the coverage of an estimator that is sometimes negative. Because an upward-biased estimator buys coverage with length, coverage is reported together with the interval length relative to the oracle's.

The oracle interval $\widehat{\beta} \pm z_{0.975} V^{1/2}$ uses the true V, which depends only on Σ and is therefore known exactly. It is the benchmark against which the estimators' coverage is read. With Gaussian errors, $\widehat{\beta} - \beta$ is exactly normal given the design, so the oracle's coverage probability is exactly 95% in every design; its simulated coverage varies across designs due to the Monte Carlo noise alone, and only approximately so with the non-Gaussian errors of Section H.2.5. Finally, an estimator is *undefined* on a design if it cannot be computed, for example CR2 when some M_{gg} is singular; Exhibit 2 (table 5) studies such designs. Unless a note says otherwise, tables and figures report coverage of nominal 95% intervals pooled over designs, and the exact mean bias (Monte Carlo for LCO), the standard deviation of $\widehat{V}/V$ and the interval length relative to the oracle's as medians over designs.

10.2 Results

The simulations support every claim in table 2. The classical corrections are biased where the theory predicts, and the size of their bias is governed by the level factor κ_n. The Riesz estimator removes this bias in every design in which it is identified, including those in which the classical estimators do not exist, and for weighted as well as unweighted fits. What exactness costs is variance. With many clusters the degrees-of-freedom correction absorbs it and restores nominal coverage; with few clusters the variance must first be reduced by a restricted class.

Table 2: Exhibits, experiments and claims

	Exhibit	Experiments	Claim
1	Coverage as ℓ/n grows	A, B, I	Classical estimators are biased as leverage grows; Riesz is unbiased and covers nominally
2	Existence	C	Riesz exists where CR2, CR3 and LCO do not
3	Two-way clustering	E	Riesz and the class cover nominally where CGM does not; the stopping rule flags when the variance is not identified
4	Weighted fits	F, G, H, K	Riesz-M and Riesz-Υ are unbiased; which is more precise depends on the weights
5	Level factor	A, B, L, N	Classical bias follows κ_n, not p/n

Experiments: lettered designs, described in Section H, with the number of designs and draws in table 16. ℓ/n: number of controls relative to the sample size. Riesz: the fully robust Riesz estimator. CR2, CR3: the cluster leverage corrections; LCO: the leave-cluster-out estimator of Anatolyev and Ng (2026). CGM: the two-way estimator of Cameron et al. (2011); the class: the estimator exact on $\mathcal{D}_{2W}^{\text{het}+\rho}$ (49). Riesz-M and Riesz-Υ: the weighted estimators in OLS and in whitened coordinates (Sections 9.2 and 9.3). κ_n: the level factor (8).

Exhibit 1: coverage as ℓ/n grows. Figures 2 to 4 and table 3 report coverage, bias and standard deviation. First, the Riesz estimator attains nominal coverage in all three settings, whereas the classical estimators do not when growing leverage meets heteroskedasticity or within-cluster covariance (figure 2). In the scalar setting, Riesz and HCK, which are first-order equivalent (Lemma E.2), cover 0.947 and 0.946. In the clustered settings, Riesz covers slightly better than LCO (0.944 against 0.941 with clustered regressors). The classical estimators behave as documented in the literature. Under

homoskedasticity, HC1 and HC2 remain accurate even as ℓ/n grows. Under heteroskedasticity, their coverage falls with ℓ/n, to 0.927 for HC2 at $\ell/n = 0.5$. When errors and regressors are both clustered, the distortion appears at smaller ℓ/n: with $\rho_u = 0.3$, CR1 and CR2-BM cover 0.895 and 0.928 at $\ell/n = 0.2$. HC1/CR1 and HC2/CR2-BM undercover, whereas HC3/CR3 over-cover.

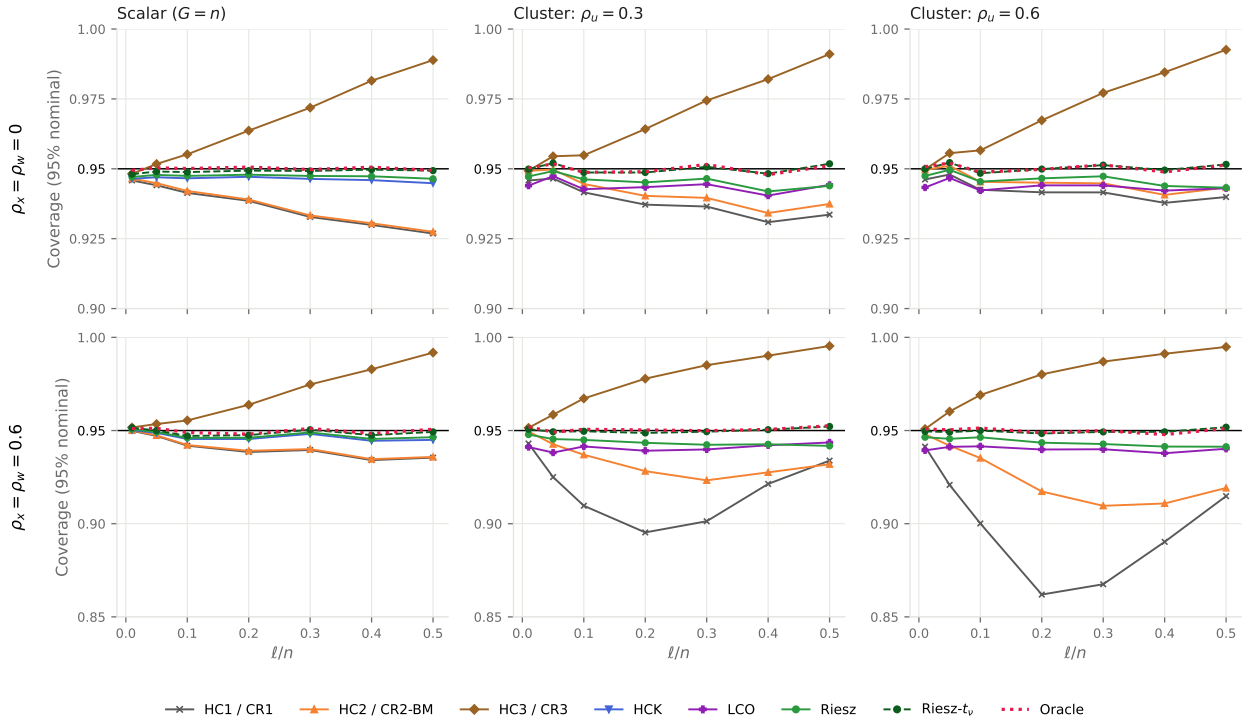


Figure 2: Exhibit 1: Coverage against ℓ/n, heteroskedastic designs ($\vartheta = 1$), $n = 1000$. Rows: regressors without ($\rho_x = \rho_w = 0$) and with ($\rho_x = \rho_w = 0.6$) a cluster component. Columns: independent errors ($G = n$), and errors clustered in $G = 200$ clusters with $\rho_u = 0.3$ and 0.6. HCK is shown only where the errors are independent. Coverage is pooled over designs; the horizontal line is 95%. HC0 and CR0 are in table 3.

Second, the Riesz estimator is exactly unbiased, whereas the classical estimators are not (figure 3). HCK is slightly biased in the scalar setting, and LCO is close to unbiased in the clustered settings. HC2 is unbiased under homoskedasticity but biased downward under heteroskedasticity, reaching -14.8% at $\ell/n = 0.5$. HC3/CR3 are biased upward, and the bias grows rapidly with ℓ/n. When errors and regressors are both clustered, the downward bias of CR1 and CR2-BM is largest at moderate ℓ/n.

Third, exactness costs variance (figure 4). Riesz has a larger standard deviation than HC1/CR1 and HC2/CR2-BM, and the gap widens with ℓ/n. Its standard deviation stays below that of HC3/CR3 throughout, and below that of LCO except at the largest ℓ/n. With many clusters, the degrees-of-freedom correction absorbs this cost: Riesz-t_ν covers 0.950 at $G = 200$ and 0.950 at $G = 50$ (table 4). With very few clusters, it's not enough: at $G = 10$, the degrees of freedom of Riesz fall to 3.2, and Riesz-t_ν covers only 0.909; the restricted class $\mathcal{D}^{\text{het}+\rho}$ can restore coverage to 0.948.

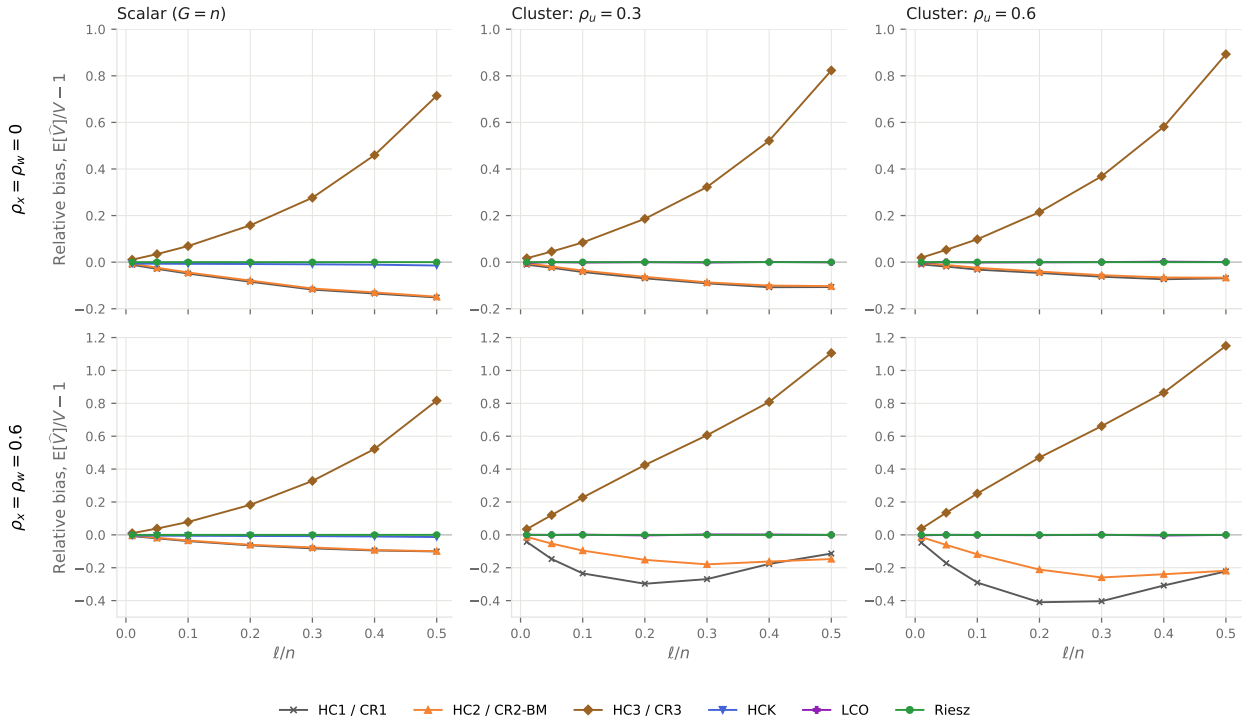


Figure 3: Exhibit 1: Exact relative bias $\mathbb{E}[\widehat{V} \mid X, W]/V - 1$, median over designs. Panels as in figure 2.

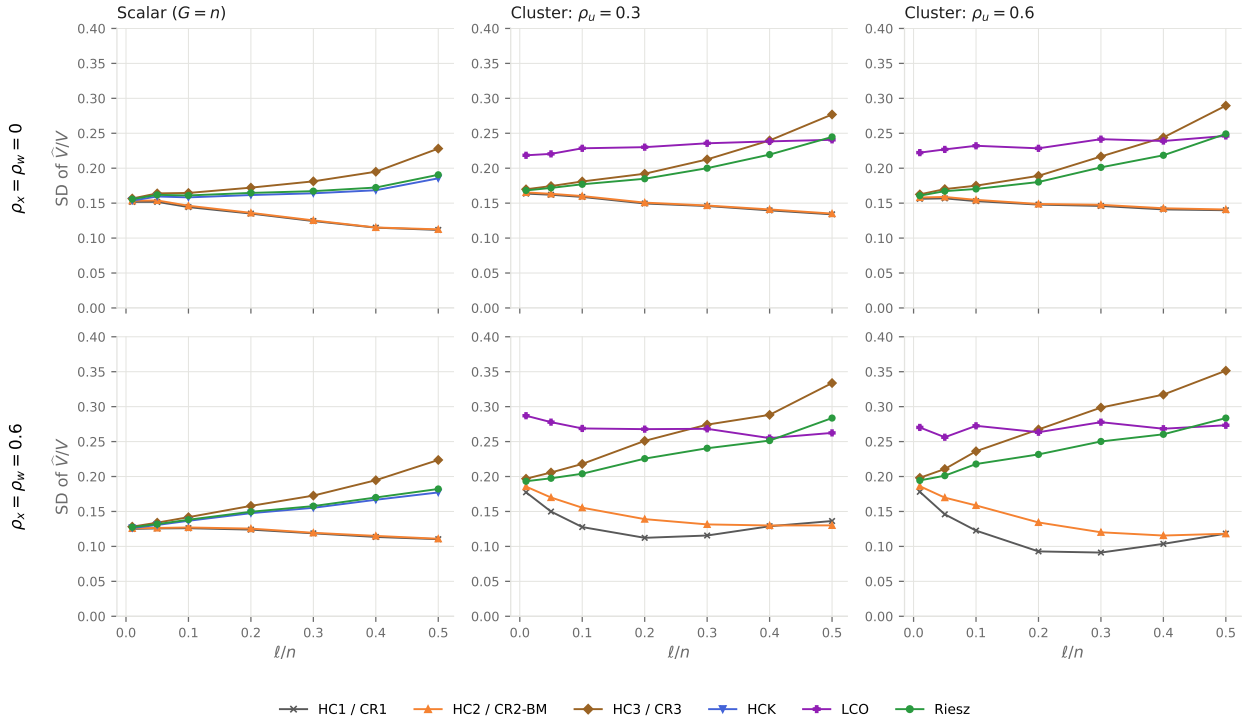


Figure 4: Exhibit 1: Standard deviation of $\widehat{V}/V$, median over designs. Panels as in figure 2.

Table 3: Exhibit 1: Coverage and bias

	$\vartheta = 0$						$\vartheta = 1$					
	$\ell/n = 0.01$		$\ell/n = 0.5$		All ℓ/n		$\ell/n = 0.01$		$\ell/n = 0.5$		All ℓ/n	
	Cov.	Bias	Cov.	Bias	Cov.	Bias	Cov.	Bias	Cov.	Bias	Cov.	Bias
Scalar ($G = n$)												
HC0	0.946	−1.3%	0.831	−50.2%	0.906	−20.3%	0.944	−2.2%	0.795	−57.7%	0.889	−26.8%
HC1	0.948	−0.2%	0.948	−0.2%	0.948	−0.2%	0.946	−1.2%	0.927	−15.2%	0.937	−8.4%
HC2	0.948	0.0%	0.948	0.0%	0.949	0.0%	0.946	−0.6%	0.927	−14.8%	0.938	−7.9%
HC3	0.949	1.3%	0.994	101.1%	0.971	25.5%	0.948	1.1%	0.989	71.4%	0.966	15.8%
HCK	0.947	−0.3%	0.945	−1.0%	0.947	−0.4%	0.946	−0.6%	0.945	−1.4%	0.946	−0.9%
Riesz	0.948	0.0%	0.946	0.0%	0.948	0.0%	0.947	0.0%	0.946	0.0%	0.947	0.0%
Riesz-t_ν	0.949	0.0%	0.949	0.0%	0.950	0.0%	0.948	0.0%	0.949	0.0%	0.949	0.0%
Oracle	0.949	0.0%	0.949	0.0%	0.949	0.0%	0.949	0.0%	0.949	0.0%	0.950	0.0%
Clusters, errors only ($\rho_w = \rho_x = 0$, $\rho_u = 0.3$)												
CR0	0.945	−1.9%	0.832	−50.5%	0.904	−20.5%	0.944	−2.5%	0.804	−55.6%	0.891	−26.0%
CR1	0.947	−0.4%	0.947	−0.4%	0.948	−0.4%	0.946	−1.1%	0.934	−10.7%	0.939	−6.9%
CR2-BM	0.949	0.0%	0.950	−0.1%	0.950	0.0%	0.949	−0.5%	0.937	−10.3%	0.942	−6.3%
CR3	0.949	1.9%	0.994	103.1%	0.971	26.5%	0.949	1.7%	0.991	82.3%	0.967	18.6%
HCK	0.947	−0.2%	0.946	−1.2%	0.948	−0.3%	0.947	−0.7%	0.946	−1.4%	0.947	−0.7%
LCO	0.944	−0.1%	0.943	0.1%	0.945	0.0%	0.944	−0.3%	0.944	−0.1%	0.944	−0.1%
Riesz	0.947	0.0%	0.942	0.0%	0.946	0.0%	0.947	0.0%	0.944	0.0%	0.946	0.0%
Riesz-t_ν	0.949	0.0%	0.950	0.0%	0.950	0.0%	0.950	0.0%	0.952	0.0%	0.950	0.0%
Oracle	0.949	0.0%	0.949	0.0%	0.950	0.0%	0.950	0.0%	0.951	0.0%	0.950	0.0%
Clusters, errors and regressors ($\rho_w = \rho_x = 0.6$, $\rho_u = 0.3$)												
CR0	0.939	−5.4%	0.825	−51.6%	0.875	−42.7%	0.941	−5.7%	0.807	−56.0%	0.867	−43.8%
CR1	0.941	−3.9%	0.944	−2.5%	0.924	−14.8%	0.943	−4.3%	0.934	−11.4%	0.919	−17.7%
CR2-BM	0.947	−0.9%	0.942	−7.0%	0.940	−7.8%	0.950	−1.3%	0.932	−14.7%	0.934	−13.8%
CR3	0.949	3.9%	0.996	130.0%	0.977	52.0%	0.951	3.5%	0.995	110.6%	0.975	42.5%
HCK	0.859	−43.8%	0.936	−9.4%	0.897	−30.5%	0.883	−36.0%	0.937	−9.2%	0.908	−24.7%
LCO	0.937	−0.4%	0.942	0.0%	0.939	0.0%	0.941	0.0%	0.944	−0.1%	0.941	0.0%
Riesz	0.945	0.0%	0.940	0.0%	0.944	0.0%	0.948	0.0%	0.942	0.0%	0.944	0.0%
Riesz-t_ν	0.948	0.0%	0.950	0.0%	0.949	0.0%	0.951	0.0%	0.952	0.0%	0.950	0.0%
Oracle	0.950	0.0%	0.951	0.0%	0.950	0.0%	0.952	0.0%	0.953	0.0%	0.951	0.0%

Cov.: coverage of the nominal 95% interval, pooled over designs. Bias: exact relative bias $\mathbb{E}[\widehat{V} \mid X, W]/V - 1$, median over designs (Monte Carlo for LCO). $\vartheta = 0$: homoskedastic; $\vartheta = 1$: heteroskedastic. All ℓ/n: pooled over $\ell/n \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$. $n = 1000$; clustered settings: $G = 200$.

Table 4: Exhibit 1: Many and few clusters

	Coverage			Length relative to oracle			Degrees of freedom ν		
	$G = 200$	$G = 50$	$G = 10$	$G = 200$	$G = 50$	$G = 10$	$G = 200$	$G = 50$	$G = 10$
CR1-t_{G-1}	0.920	0.906	0.901	0.91	0.92	1.03	199	49	9
CR3-t_{G-1}	0.976	0.977	0.978	1.19	1.21	1.33	199	49	9
CR2-BM	0.934	0.926	0.928	0.94	0.94	1.07	113.5	38.1	7.8
Riesz	0.944	0.931	0.834	0.99	0.98	0.95	∞	∞	∞
Riesz-t_{G-1}	0.945	0.937	0.864	1.00	1.01	1.10	199	49	9
Riesz-t_ν	0.950	0.950	0.909	1.02	1.06	1.50	51.9	16.7	3.2
Class-t_ν	0.951	0.949	0.948	1.01	1.01	1.00	168.7	125.6	132.8

Clustered errors and regressors, $n = 1000$, $\rho_u = 0.3$, heteroskedastic ($\vartheta = 1$), pooled over ℓ/n (by ℓ/n in table 19). Coverage of the nominal 95% interval, pooled over designs; length relative to the oracle interval and degrees of freedom ν, medians over designs (∞: normal critical value). Class-t_ν: $\mathcal{D}^{\text{het}+\rho}$ (Section 8.3).

Exhibit 2: existence. A block inverse fails whenever a regressor is supported on a single cluster (Proposition 7.4), as with cluster fixed effects or cluster-specific indicators. CR2/CR2-BM, CR3 and LCO are then undefined in every design with cluster fixed effects and in 60% of the designs with rare indicators (table 5). HCK fails in a different way in a two-period panel with unit fixed effects: the two residuals of each unit sum to zero, so the two rows of $M_W \odot M_W$ for the unit are equal, and HCK is undefined in every design. The Riesz estimator is defined in every design. With cluster fixed effects the singular direction lies in $\mathrm{col}(W)$, where the contrast does not load (Proposition E.10); in the panel the influence weights of a unit's two observations are equal and opposite, so V depends only on the sum of their variances, which the residuals identify. With t_ν critical values, Riesz covers 0.949 to 0.950 across the four designs. Among the estimators that remain defined, the generalized-inverse CR2$^+$ of Pustejovsky and Tipton (2018) covers 0.940 to 0.945, close to Riesz with normal critical values (0.942 to 0.945); CR3$^+$ over-covers (0.970 to 0.998), and the coverage of CR1 ranges from 0.942 to 0.966.

Table 5: Exhibit 2: Existence and coverage

	Cluster FE, all clusters		Cluster FE, some clusters		Rare indicators		Two-period panel, unit FE ($G = n$)	
	Undef.	Cov.	Undef.	Cov.	Undef.	Cov.	Undef.	Cov.
CR2, CR3, CR2-BM	1.00	–	1.00	–	0.60	0.950		
LCO	1.00	–	1.00	–	0.60	0.943		
HCK							1.00	–
CR2$^+$ / HC2	0.00	0.945	0.00	0.940	0.00	0.940	0.00	0.945
CR3$^+$ / HC3	0.00	0.982	0.00	0.977	0.00	0.970	0.00	0.998
CR1 / HC1	0.00	0.966	0.00	0.947	0.00	0.944	0.00	0.942
Riesz	0.00	0.942	0.00	0.943	0.00	0.945	0.00	0.945
Riesz-t_ν	0.00	0.950	0.00	0.950	0.00	0.950	0.00	0.949

Undefined: share of designs on which the estimator cannot be evaluated; coverage is over the others. Cluster fixed effects for every cluster or a share $\pi \sim \mathrm{U}[0.1, 0.9]$ of them, and rare-indicator controls: $n = 1000$, $G = 200$. Two-period panel: $n = G = 1000$, 500 units observed twice, with unit fixed effects and a period dummy. Details in Section H.3.

Exhibit 3: two-way clustering. Experiment E has two parts (Section H.4). The first concerns validity: $n = 1000$ observations clustered on two dimensions with $G_1, G_2 \in \{10, 30, 50, 100\}$ and $\ell = 10$ controls, so that the counting bound holds in every design. We compare the two-way estimator of Cameron et al. (2011) (CGM), $\widehat{V}_1 + \widehat{V}_2 - \widehat{V}_{12}$, which adds CR0 clustered on each dimension and subtracts CR0 clustered on their intersection, and CGM1, which scales each component by the small-sample factor $G_k/(G_k - 1) \cdot (n-1)/(n-p)$ that Cameron et al. (2011) recommend; t_{G-1} uses $\min(G_1, G_2) - 1$ degrees of freedom. Table 6 reports the results by the smaller number of clusters.

CGM1-t_{G-1} undercovers, most with few clusters: it covers 0.921 with ten clusters and 0.941 with a hundred. Two sources contribute. First, a clustered regressor biases CGM1 downward even with few controls: with ten clusters its median bias is -3.0% when the regressor has no cluster component and -20.7% when $\rho_x = 0.6$, through the level factor of Proposition 7.3. Second, t_{G-1} understates the variability of a two-way variance estimator even when it is unbiased: the exactly unbiased Riesz estimator covers only 0.931 with t_{G-1} and ten clusters, because its median degrees of freedom are indeed 6.0 rather than 9.

With its own t_ν, Riesz covers nominally at every number of clusters (0.951 with ten, 0.948 with a hundred), at the cost of intervals 1.20 times as long as the oracle's with ten clusters. The class $\mathcal{D}_{2W}^{\text{het}+\rho}$ of (49), which contains the error covariance of these designs, covers 0.947 with ten clusters and 0.948 with a hundred, with intervals as long as the oracle's.

Table 6: Exhibit 3: two-way clustering with few controls

	$\min(G_1, G_2) = 10$	$\min(G_1, G_2) = 30$	$\min(G_1, G_2) = 50$	$\min(G_1, G_2) = 100$
Coverage, normal critical values				
CGM	0.859	0.916	0.930	0.935
CGM1	0.882	0.925	0.935	0.938
Riesz	0.898	0.931	0.940	0.941
Coverage, t_{G-1} critical values				
CGM-t_{G-1}	0.902	0.928	0.936	0.938
CGM1-t_{G-1}	0.921	0.936	0.941	0.941
Riesz-t_{G-1}	0.931	0.941	0.945	0.943
Coverage, t_ν critical values				
Riesz-t_ν	0.951	0.950	0.951	0.948
Class-t_ν	0.947	0.949	0.950	0.948
Oracle	0.951	0.950	0.951	0.948
Interval length relative to the oracle's, median				
CGM1-t_{G-1}	1.02	1.00	0.99	0.99
Riesz-t_{G-1}	1.11	1.03	1.01	1.00
Riesz-t_ν	1.20	1.07	1.04	1.02
Class-t_ν	1.00	1.00	1.00	1.00
Degrees of freedom ν, median				
Riesz	6.0	15.6	22.9	39.4
Class	132.0	216.4	220.2	248.2
Exact relative bias of CGM1, median				
$\rho_x = 0$	-3.0%	-2.5%	-2.0%	-1.2%
$\rho_x = 0.3$	-18.4%	-7.2%	-5.1%	-2.6%
$\rho_x = 0.6$	-20.7%	-10.2%	-6.6%	-3.9%

Validity part of Experiment E: $n = 1000$, $\ell = 10$, $G_1, G_2 \in \{10, 30, 50, 100\}$, grouped by $\min(G_1, G_2)$ (140, 100, 60 and 20 designs). $\rho_w, \rho_x \in \{0, 0.3, 0.6\}$ are drawn for each design, and the bias rows pool over ρ_w (table 23). Coverage is pooled over designs. CGM: $\widehat{V}_1 + \widehat{V}_2 - \widehat{V}_{12}$ with CR0 components; CGM1: with the factor $G_k/(G_k - 1) \cdot (n-1)/(n-p)$ on each; t_{G-1}: $\min(G_1, G_2) - 1$ degrees of freedom. Riesz: fully robust, on $\mathcal{H}_n^{S_{2W}}$; Class: $\mathcal{D}_{2W}^{\text{het}+\rho}$.

The second part concerns identification. With many controls the fully robust two-way variance need not be identified: with $G_1 = G_2 = G$, the counting bound of Remark 8.8 requires approximately $(1 - p/n)^2 \geq 2/G - 1/G^2$, which fails above $\ell/n \approx 0.56$ with ten clusters and 0.86 with a hundred. With $n = 500$ and ℓ/n from 0.1 to 0.9, the stopping rule of LSQR switches from the primal to the dual criterion exactly where the bound fails, in all 360 designs: the residual at which it stops is about 10^{-10} where the bound holds and of order one where it fails (figure 5(a) and (b)). Riesz-t_ν undercovers as ℓ/n approaches the bound, where the variance rests on few residual dimensions, and fails beyond it (panel (c)). The class remains identified past the bound and covers nominally up to $\ell/n = 0.8$; at $\ell/n = 0.9$, with fewer than fifty residual dimensions, it too undercovers (panel (d)). Table 7 compares all the estimators on the two sides of the bound.

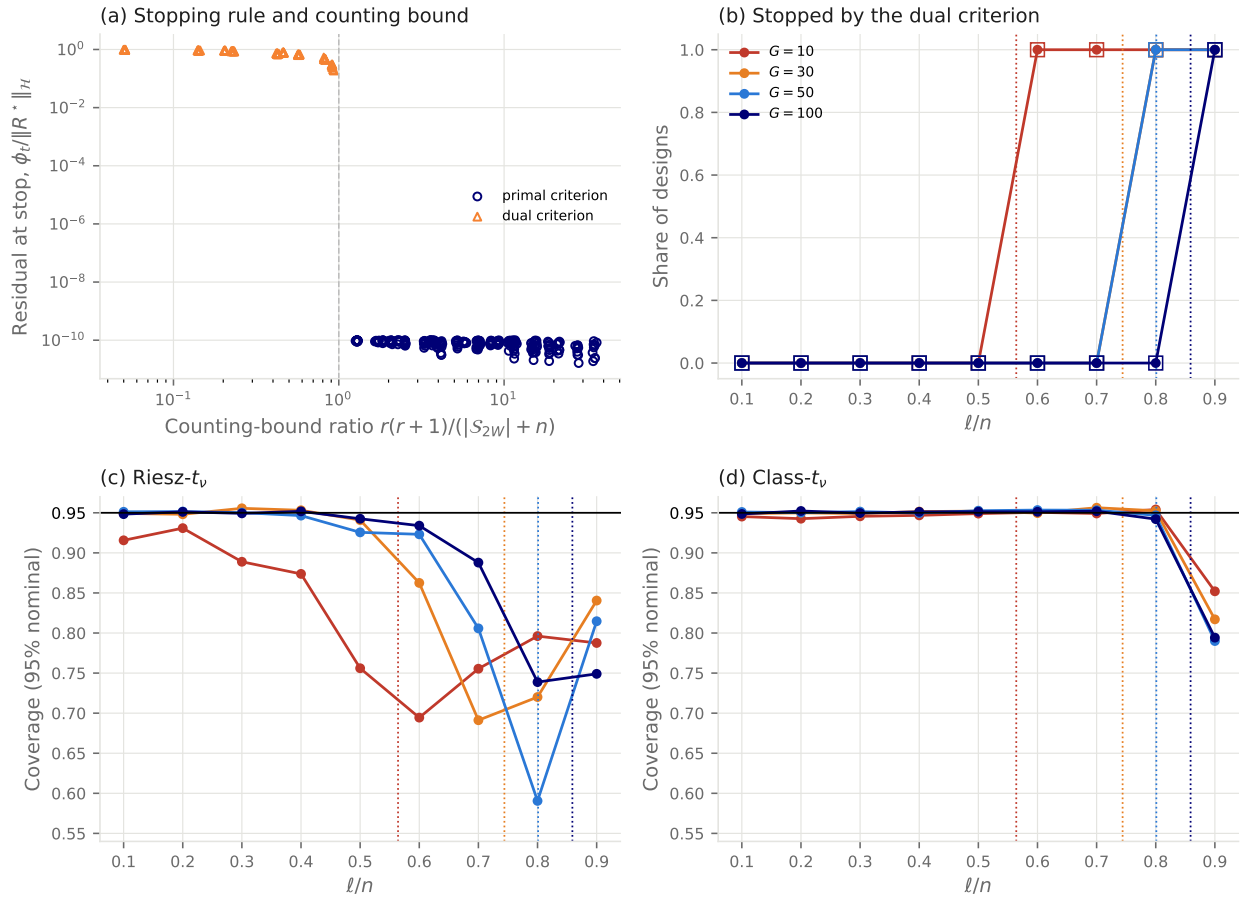


Figure 5: Exhibit 3: identification, $G_1 = G_2 = G$, $n = 500$. (a) Residual at which LSQR stops, by stopping criterion; the counting bound (Remark 8.8) holds right of the dashed line. (b) Share of designs stopped by the dual criterion (dots) and failing the bound (squares); the line for $G = 30$ lies under that for $G = 50$. (c), (d) Coverage, pooled over designs. Dotted lines: the approximate threshold $(1 - p/n)^2 = 2/G - 1/G^2$.

Table 7: Exhibit 3: identification, split by the counting bound

	Bound holds (270 designs)				Bound fails (90 designs)			
	Cov.	Bias	Neg.	Length	Cov.	Bias	Neg.	Length
One-way CR0	0.736	−60.8%	0.000	0.62	0.484	−89.7%	0.000	0.32
CGM	0.775	−53.8%	0.005	0.67	0.475	−89.4%	0.016	0.32
CGM-t_{G-1}	0.794	−53.8%	0.005	0.71	0.505	−89.4%	0.016	0.34
CGM1	0.900	−12.0%	0.003	0.93	0.871	−15.2%	0.009	0.90
CGM1-t_{G-1}	0.914	−12.0%	0.003	0.95	0.895	−15.2%	0.009	0.97
Thompson	0.775	−52.8%	0.006	0.67	0.473	−89.3%	0.021	0.32
Riesz	0.851	0.0%	0.066	0.98	0.632	−30.2%	0.233	0.90
Riesz-t_ν	0.905	0.0%	0.066	1.34	0.750	−30.2%	0.233	5.49
Class	0.941	0.0%	0.000	0.99	0.840	0.0%	0.082	0.99
Class-t_ν	0.950	0.0%	0.000	1.02	0.890	0.0%	0.082	1.19
Oracle	0.950	0.0%	0.000	1.00	0.950	0.0%	0.000	1.00

Identification part of Experiment E, as in figure 5. Cov.: coverage, pooled over designs. Bias: exact relative bias; Length: interval length relative to the oracle's; both medians over designs. Neg.: share of draws with $\widehat{V} \leq 0$. One-way CR0: CR0 clustered on the first dimension; Thompson: CGM with HC0 in place of the intersection term. Other estimators as in table 6.

Exhibit 4: weighted and generalized least squares. Experiments F and G refit the scalar and clustered designs by weighted least squares with known weights, either unrelated to Σ or modelling it (Section H.5). The weighted cluster sandwich carries the projection spillover and undercovers (down to 0.724; table 8, Panel A). Riesz-M and Riesz-Υ are exactly unbiased and, with t_ν critical values, cover close to nominal. Which is more precise follows the working covariance (Proposition 9.5): Riesz-M when the weights are unrelated to Σ, Riesz-Υ when they model it. Experiment H, with estimated weights, is descriptive only (Section 9.3).

Panel B (Experiment K) is consistent with Proposition 9.6 for the full and restricted systems in the weighted residuals. With weights unrelated to Σ, the full system covers better in 79% of the cells, by up to 27 percentage points, and Riesz-M, on the OLS residuals, covers at least as well as both; with correct GLS weights, the restricted system covers better, although at very dispersed weights every fully robust estimator is fragile. With weights constant within a few strata the two systems nearly coincide.

Table 8: Exhibit 4: weighted and generalized least squares

Panel A. Coverage and efficiency of the estimators of V_Υ

	Coverage					Efficiency	
Weights	WLS-CR0	Riesz-Υ	Riesz-M	Riesz-Υ-t_ν	Riesz-M-t_ν	Share Υ	SD Υ/M
F, unit, $\varsigma = 0.5$	0.869	0.920	0.922	0.954	0.955	0.07	1.02
F, unit, $\varsigma = 2$	0.724	0.855	0.921	0.940	0.969	0.00	1.28
G, unit	0.777	0.877	0.914	0.938	0.957	0.00	1.13
G, unit + block	0.784	0.869	0.906	0.934	0.947	0.22	1.09
G, block	0.885	0.921	0.914	0.947	0.937	0.84	0.96
G, infeasible GLS	0.897	0.923	0.906	0.951	0.931	1.00	0.86
H, random effects	0.871	0.920	0.914	0.949	0.943	–	–
H, heteroskedastic	0.866	0.918	0.917	0.947	0.944	–	–
H, both	0.891	0.920	0.908	0.945	0.934	–	–

Panel B. The full and the restricted system in the weighted residuals (Experiment K)

		Coverage			Length / oracle		Negative $\widehat{V}$		Ahead by 2+ pts	
Weights	ς	F(T)	P(T)	M	F(T)	P(T)	F(T)	P(T)	F(T)	P(T)
Independent	1	0.907	0.906	0.917	0.89	0.89	0.003	0.000	8%	8%
	3	0.834	0.743	0.919	0.85	0.88	0.094	0.151	83%	4%
Two strata	1	0.922	0.922	0.922	0.91	0.91	0.000	0.000	0%	0%
	3	0.922	0.919	0.922	0.92	0.91	0.000	0.000	8%	4%
Monotone in x	1	0.893	0.833	0.905	0.82	0.81	0.026	0.013	67%	21%
	3	0.858	0.660	0.929	0.80	0.90	0.086	0.245	100%	0%
Correct GLS	1	0.591	0.899	0.791	1.16	0.94	0.336	0.000	0%	100%
	3	0.426	0.691	0.426	18.29	1.00	0.573	0.214	0%	100%

Panel A: Share Υ: share of designs in which Riesz-Υ has the smaller exact Gaussian standard deviation; SD Υ/M: median ratio of the two; in H the weights are estimated on every draw. Panel B: $\Upsilon = \exp(\varsigma\xi)$ with $\xi \sim \mathcal{N}(0,1)$ (Independent) or the normal scores of x (Monotone in x), or weight $e^{-\varsigma}$ on an oversampled fifth (Two strata); Correct GLS: $\Upsilon = \Sigma^{-1}$. Medians over cells. Ahead by 2+ pts: share of a row's cells in which one system covers at least two points more.

Exhibit 5: the level factor across designs. Exhibit 5 asks what sets the level of the classical bias. It pools the $4,940$ designs of Exhibit 1, the ten-cluster designs of Section H.2.3, and Experiments L and N, which add unequal cluster sizes, treatments on few clusters and dyadic designs (Section H.6). Figure 6 shows what moves the level factor κ_n of Proposition 7.3, figures 7 and 8 how it drives the bias of the classical estimators, and table 9 how it drives their coverage.

First, κ_n is close to p/n in the scalar setting and when only the errors are clustered (figure 6), since the regressors then carry no cluster component. Clustered regressors, few clusters and few treated clusters raise it above p/n, toward p/G: with ten clusters it ranges from 0.11 to 0.61, and at $p/n = 0.05$ a treatment on few clusters gives κ_n between 0.11 and 0.28. What matters is where the regressor varies. Unequal cluster sizes move κ_n only through it, from 0.074 when the regressor loads on the small clusters to 0.107 when it loads on the large ones, and in the dyadic designs node fixed effects lower κ_n from 0.28 to 0.18 at the same p/n and the same off-block mass $\sum_g \sum_{h\neq g} \|M_{gh}\|_F^2 / \sum_g \|M_{gg}\|_F^2$.

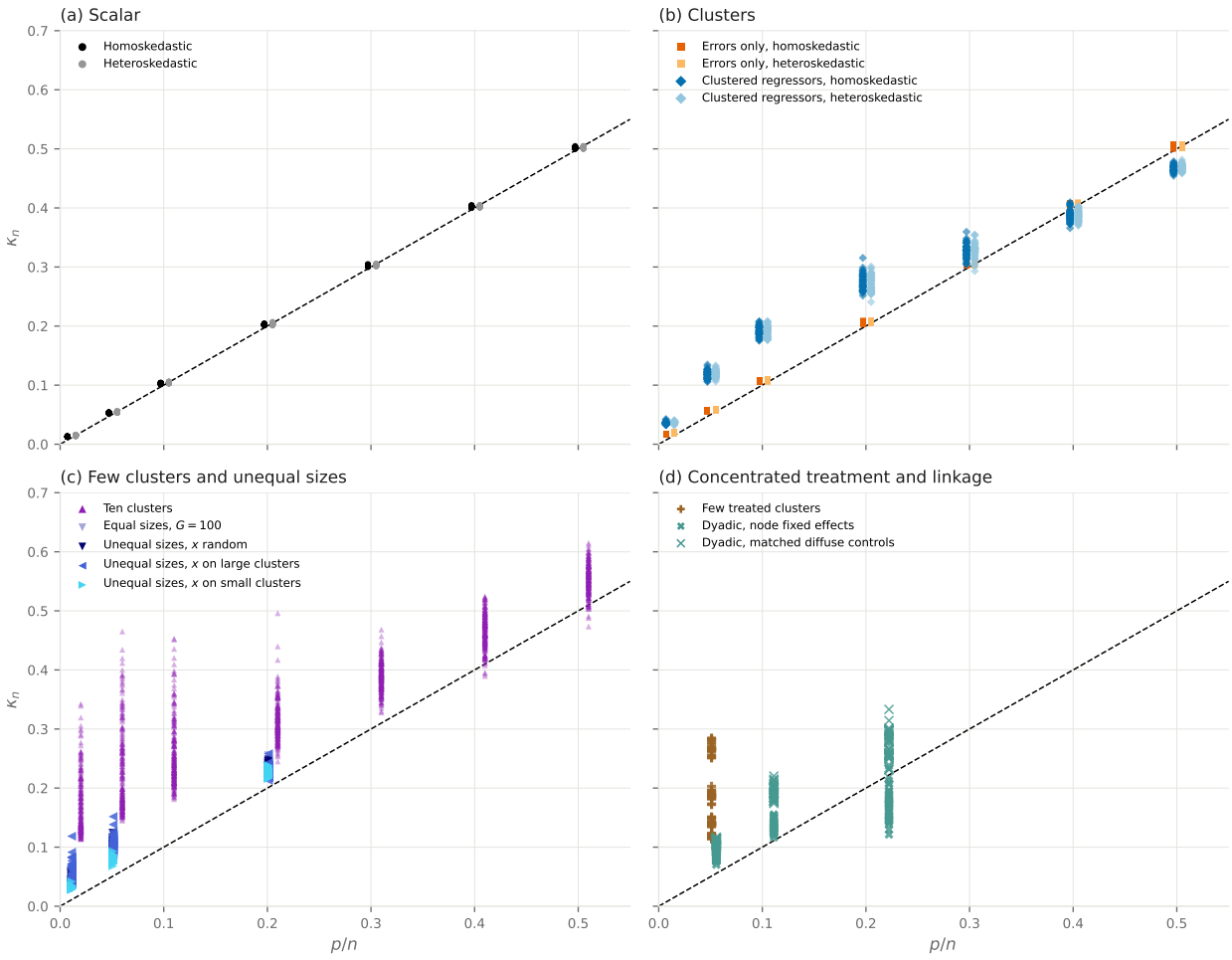


Figure 6: Exhibit 5: κ_n against p/n; dashed line $\kappa_n = p/n$. (a) Scalar and (b) clustered settings of Exhibit 1; (c) ten clusters with $n = 100$ (Section H.2.3) and equal and unequal sizes at $G = 100$ (Experiment L); (d) few treated clusters (L) and dyadic designs (N). Exhibit 1 settings: dark $\vartheta = 0$, light $\vartheta = 1$.

Second, the bias of CR0 follows κ_n (figure 7): across the designs its exact bias has correlation -0.94 with κ_n and -0.82 with p/n. Under $\Sigma = \sigma^2 I_n$ it lies on the line $-\kappa_n$; heteroskedasticity and within-cluster correlation move it below the line, through the contrast. The contrast also explains

why unequal sizes matter more than κ_n suggests: when the regressor loads on the large clusters, the bias of CR1 grows from -12.9% to -30.0% at $\ell/n = 0.05$, while κ_n rises only from 0.092 to 0.107, because the heteroskedasticity then sits where the leverage is. The leverage corrections remove the level but not the contrast (figure 8). HC2 is exact under homoskedasticity in the scalar setting, but once Σ departs from $\sigma^2 I_n$ the bias of CR2 fans out: with heteroskedasticity its median is -7.5% and its worst design -40%. CR3 over-corrects, by at least $\kappa_n/(1-\kappa_n)$ under $\Sigma = \sigma^2 I_n$, and by up to 283% across the designs.

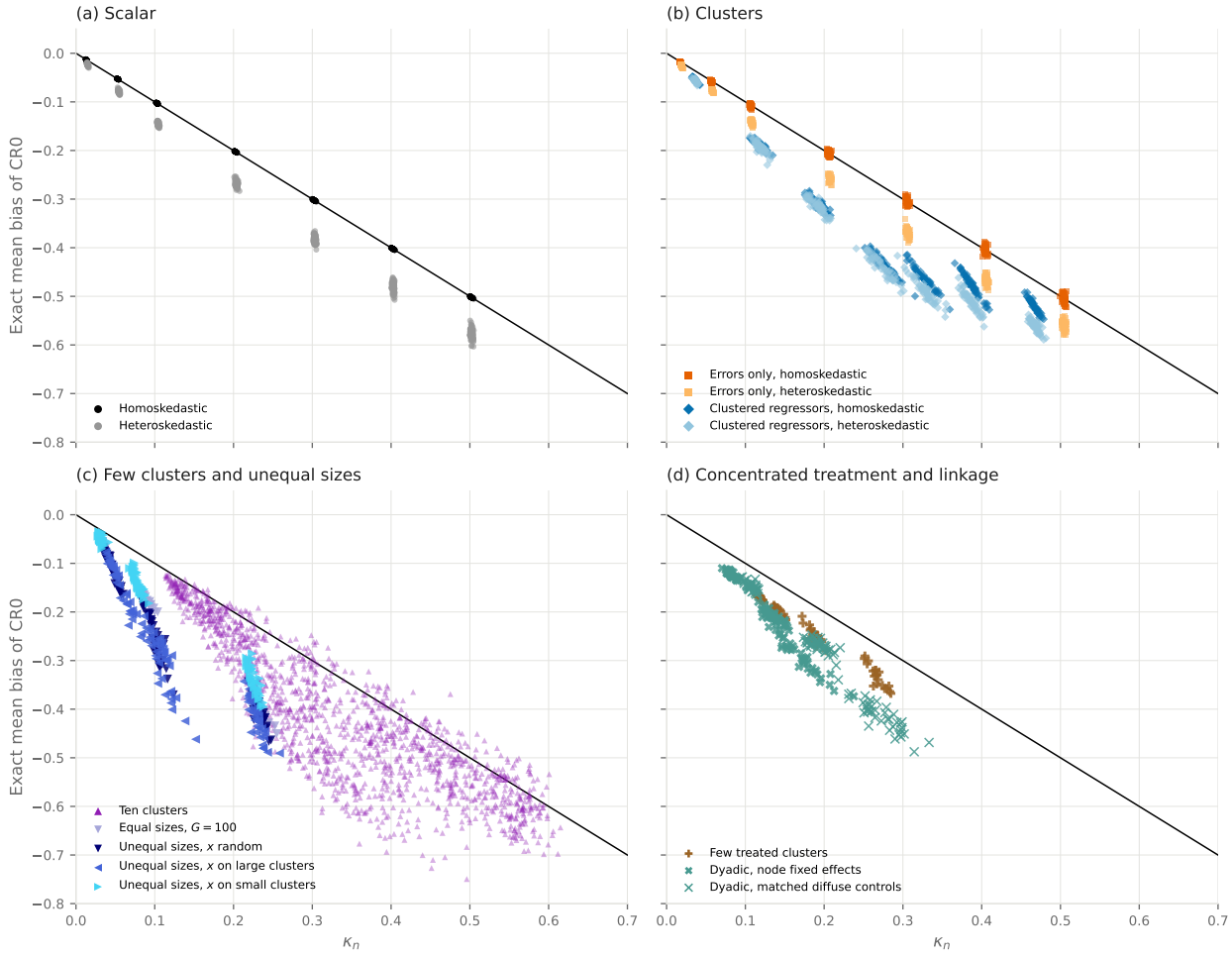


Figure 7: Exhibit 5: exact mean bias of CR0 (HC0 when $N_g = 1$) against κ_n; the line is $-\kappa_n$, exact under $\Sigma = \sigma^2 I_n$ (Proposition 7.3). Panels and shading as in figure 6.

Third, coverage follows κ_n as well (table 9). Within each block of the table, the coverage of CR1 falls as κ_n rises, even where p/n is fixed and small: with a treatment on fifty clusters $\kappa_n = 0.12$ and CR1 covers 0.928; with five, $\kappa_n = 0.27$, CR1 is biased by -29% and covers 0.839. In the dyadic designs, the diffuse controls, with the larger κ_n, give the larger bias of CR1 (-24.6% against -10.2%) and the lower coverage, so the bias follows κ_n, not how much the controls link the clusters (Section H.6). CR2-BM, which removes the level, covers between 0.93 and 0.95 throughout. The Riesz estimator is exact in every design, and with t_ν critical values it covers close to nominal in every row.

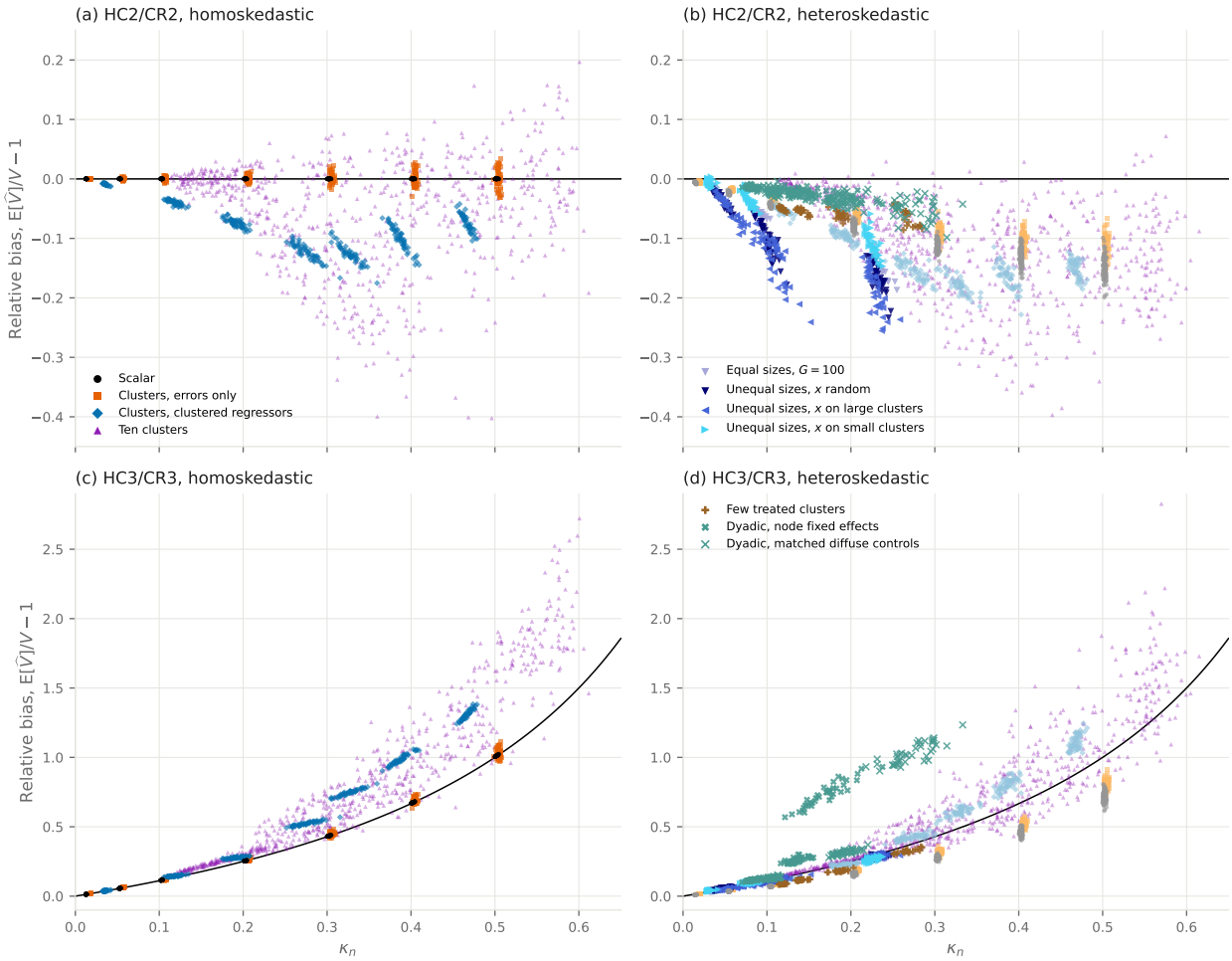


Figure 8: Exhibit 5: exact mean bias of HC2/CR2 (a, b) and HC3/CR3 (c, d) against κ_n, without (a, c) and with (b, d) heteroskedasticity. Lines: bias under $\Sigma = \sigma^2 I_n$, zero for HC2/CR2 and the lower bound $\kappa_n/(1-\kappa_n)$ for HC3/CR3.

Table 9: Experiments L and N: unequal sizes, few treated clusters, dyadic designs

				CR1		Coverage				
	p/n	p/G	κ_n	Bias	Cov.	CR2-BM	HCK	Riesz	-t_ν	Class-t_ν
Unequal cluster sizes (Experiment L), $G = 100$, $\ell/n = 0.05$										
$\eta_N = 0$, equal sizes	0.05	0.51	0.09	-12.9%	0.929	0.943	0.893	0.945	0.950	0.950
$\eta_N = 2$, random	0.05	0.51	0.09	-16.7%	0.922	0.941	0.886	0.945	0.951	0.952
$\eta_N = 2$, x on large clusters	0.05	0.51	0.11	-22.3%	0.907	0.931	0.861	0.938	0.946	0.948
$\eta_N = 2$, x on small clusters	0.05	0.51	0.08	-10.2%	0.930	0.942	0.910	0.943	0.948	0.948
$\eta_N = 4$, random	0.05	0.51	0.10	-24.0%	0.902	0.933	0.864	0.939	0.948	0.950
$\eta_N = 4$, x on large clusters	0.05	0.51	0.11	-30.0%	0.890	0.927	0.856	0.936	0.948	0.949
$\eta_N = 4$, x on small clusters	0.05	0.51	0.07	-6.3%	0.938	0.949	0.936	0.945	0.951	0.951
Few treated clusters (Experiment L), $G = 200$, $\ell/n = 0.05$										
$T = 5$ treated	0.05	0.26	0.27	-29.2%	0.839	0.944	0.809	0.875	0.953	0.957
$T = 10$ treated	0.05	0.26	0.19	-20.6%	0.893	0.943	0.822	0.919	0.951	0.953
$T = 20$ treated	0.05	0.26	0.14	-15.0%	0.918	0.943	0.823	0.936	0.950	0.951
$T = 50$ treated	0.05	0.26	0.12	-12.1%	0.928	0.942	0.812	0.944	0.949	0.950
Dyadic designs (Experiment N), $G = 240$ *dyads of three*										
$k = 3$, diffuse	0.22	0.67	0.28	-24.6%	0.907	0.944	0.918	0.941	0.948	0.950
$k = 3$, node FE	0.22	0.67	0.18	-10.2%	0.935	0.946	0.935	0.943	0.947	0.950
$k = 6$, diffuse	0.11	0.33	0.19	-17.4%	0.921	0.945	0.916	0.945	0.949	0.949
$k = 6$, node FE	0.11	0.33	0.14	-11.1%	0.933	0.947	0.934	0.947	0.949	0.949
$k = 12$, diffuse	0.06	0.17	0.11	-9.4%	0.934	0.947	0.919	0.947	0.949	0.950
$k = 12$, node FE	0.06	0.17	0.08	-6.5%	0.943	0.950	0.936	0.950	0.952	0.952

Riesz: the fully robust estimator. Class-t_ν: $\mathcal{D}^{\text{het}+\rho}$. $\vartheta = 1$, $\rho_u = 0.3$. Coverage with normal critical values, except CR2-BM (Bell–McCaffrey) and the t_ν columns.

Further results. Section H reports further experiments and results: the replication of Cattaneo et al. (2018), the deviation classes in detail and on covariances outside them, non-Gaussian errors, further weighted fits and feasible GLS, the conditioning of the restricted system, the factor design and the other design parameters, and a misspecified partition.

11 Empirical Applications

Two applications exercise the two sources of the classical bias. In the first, many controls, many of them constant within clusters, make κ_n large, and the variance estimator changes the conclusions. In the second, p/n is negligible, but household weights concentrate the variation in a few clusters. The level factor is then large and few effective clusters carry the regressor, so the variance estimator and the reference distribution both change the conclusions, and a restricted covariance class that the data do not reject restores all but one of them.

11.1 Many Controls: Connections and Colonial Governors

We revisit Xu (2018), who studies how social connections to the Secretary of State shaped the careers of colonial governors in the British Empire. The salary regressions of his Table 2 are

$$\log(w_{ist}) = \alpha + \beta\, conn_{it} + \gamma\, served_{it} + \theta_i + \tau_t + \mu_{it} + \varepsilon_{ist},$$

where w_{ist} is the salary of governor i in colony s in year t, and $conn_{it}$ indicates a connection to the Secretary of State: shared ancestry, both aristocrats, both Etonians, both at Oxford or Cambridge, or, in regression (6), any of these. $served_{it}$ is the number of colonies previously served, and θ_i, τ_t and μ_{it} are governor, year and tenure fixed effects. Standard errors are clustered on the governor–Secretary pair, and the published ones are CR1; our CR1 reproduces them to three decimals in every regression of both tables.

The sample has $n = 3{,}510$ observations in $G = 1{,}518$ pairs and 572 controls, so $p/n = 0.163$. Many controls, notably the governor fixed effects, are constant within pairs, and κ_n is between 0.28 and 0.39, well above p/n. For 5.1% of the pairs M_{gg} is singular, so CR2 and CR3 do not exist.[18] The Riesz estimator does, and the diagnostics of Section 7.5 confirm that the variance is identified: LSQR stops on the primal criterion within 46 iterations, $\delta_n = 0$ to machine precision, and $\kappa(\mathcal{A}_n) = 5.3$.

Table 10 shows that the variance estimator changes the conclusions for three of the five measures. Shared ancestry falls from the 5% to the 10% level ($p = 0.069$, against 0.028 under CR1). Aristocratic membership and Eton attendance, significant at 10% under CR1, are no longer significant ($p = 0.24$ and 0.19). Only the composite measure remains significant at 5% ($p = 0.019$). The changes follow κ_n: the aristocrat measure has the largest κ_n (0.39), the largest increase in the standard error (48%) and the largest rise in the p-value (from 0.083 to 0.24), and across the five regressions the change in the standard error has a correlation of 0.98 with κ_n. The degrees-of-freedom correction changes none of this: ν is between 25 and 134, except 4 in regression (2), where the Riesz standard error is itself imprecise and should be read with caution.

Table 10: The connection coefficient: variance estimators and the level factor

			Standard error			p-value			
Regression	Estimate	κ_n	CR1	Riesz	Change	CR1	Riesz	Riesz-t_ν	ν
(1) Shared ancestors	0.103	0.284	0.0470	0.0570	+21%	0.028	0.069	0.073	80.7
(2) Both aristocrats	0.215	0.386	0.1238	0.1838	+48%	0.083	0.242	0.307	4.0
(3) Both Eton	0.133	0.333	0.0772	0.1002	+30%	0.085	0.185	0.197	25.0
(4) Both Oxbridge	0.072	0.300	0.0468	0.0562	+20%	0.126	0.202	0.206	73.4
(6) Connected	0.097	0.285	0.0356	0.0414	+16%	0.006	0.019	0.020	133.5

Regressions of Table 2 of Xu (2018): log governor salary on the connection measure in the row label. $n = 3{,}510$ observations in $1{,}518$ governor–Secretary pairs, $\ell = 572$ controls, standard errors clustered by pair. κ_n: the level factor (8). CR1: the published standard error. Change: from the CR1 to the Riesz standard error. p-values use normal critical values, except Riesz-t_ν, which uses t_ν with ν of (31).

The regressions behind the paper's main finding, that connected governors raised less revenue and spent less (Table 5 of Xu, 2018), are

$$y_{ist} = \alpha + \beta\, conn_{it} + \lambda\, conn_{it}\mathbf{1}\{t > 1930\} + \gamma\, served_{it} + \theta_{is} + \tau_t + \mu_{it} + \varepsilon_{ist},$$

where y_{ist} is the log of a fiscal outcome of colony s, $conn_{it}$ is the composite connection measure, and θ_{is} are governor–colony fixed effects. The post-1930 interaction enters only regressions (2) and (6), where β is the pre-reform effect. These fixed effects raise p/n to between 0.24 and 0.27 and κ_n to between 0.37 and 0.42, and make 10% to 12% of the M_{gg} singular. Table 11 shows that three of the

[18] Their generalized-inverse versions exist (Pustejovsky and Tipton, 2018). For shared ancestry CR2$^+$ and CR3$^+$ disagree at the 5% level ($p = 0.047$ and 0.110), and the Riesz estimator gives 0.069 (Section I.1).

original results lose significance. Customs revenue falls from the 5% to the 10% level ($p = 0.084$, against 0.040 under CR1). Spending on revenue services and on public works, significant at 10% under CR1, is no longer significant ($p = 0.112$ and 0.103, against 0.094 and 0.085). Total revenue remains significant at 5% ($p = 0.041$, against 0.021). The Riesz standard errors are 5% to 19% larger than CR1's. The largest change is for customs revenue, the one result that loses 5% significance. Here κ_n does not order the regressions: revenue services has the largest κ_n (0.42) and the smallest change (5%). κ_n ignores the within-pair correlation, which differs across outcomes. The correction raises the p-values by at most 0.006 (ν between 50 and 101); only pre-1930 expenditure moves, from 0.096 to 0.100. In both sets of regressions the variance estimator, not the reference distribution, decides the conclusions.

Table 11: Fiscal performance: the connection coefficient

			Standard error			p-value			
Regression	Estimate	κ_n	CR1	Riesz	Change	CR1	Riesz	Riesz-t_ν	ν
(1) Revenue	-0.040	0.384	0.0174	0.0196	+13%	0.021	0.041	0.044	101.2
(2) Revenue, pre-1930	-0.055	0.366	0.0208	0.0236	+13%	0.008	0.020	0.023	77.0
(3) Customs revenue	-0.053	0.395	0.0256	0.0304	+19%	0.040	0.084	0.088	73.6
(4) Internal revenue	-0.043	0.395	0.0319	0.0370	+16%	0.175	0.243	0.246	72.6
(5) Expenditure	-0.029	0.384	0.0189	0.0206	+9%	0.120	0.154	0.157	101.2
(6) Expenditure, pre-1930	-0.042	0.366	0.0232	0.0254	+9%	0.069	0.096	0.100	77.0
(7) Revenue services	-0.089	0.423	0.0535	0.0563	+5%	0.094	0.112	0.118	50.2
(8) Public works	-0.107	0.397	0.0619	0.0654	+6%	0.085	0.103	0.107	79.0

Regressions of Table 5 of Xu (2018): the log fiscal outcome in the row label on the composite connection measure, with governor–colony, year and tenure fixed effects; rows (2) and (6) report the pre-1930 coefficient. $n = 3{,}510$, except (3) 2,670, (4) 2,657, (7) 1,742 and (8) 2,588. Columns as in table 10.

11.2 Weights and Few Effective Clusters: Housing Wealth and Employment

Mian and Sufi (2014) relate the 2007–2009 change in non-tradable employment across US counties to the 2006–2009 change in housing net worth. Using their public replication files, we revisit the seven weighted least-squares columns of their Table III,

$$\Delta E_c = \alpha + \beta\, \Delta HNW_c + W_c'\gamma + u_c,$$

where ΔE_c is non-tradable employment growth in county c, 2007–2009 (retail and restaurants, or industries with low geographic concentration), and ΔHNW_c is the 2006–2009 change in housing net worth. The controls W_c are absent in (1) and (2); they are the 2006 employment shares of 23 two-digit industries in (3) and (4), these and the 2007 construction share and its interaction with ΔHNW_c in (8) and (9), and the shares and the 2007–2009 change in construction employment in (10). Every regression is weighted by the number of households and clustered by state. There are $n = 944$ counties in $G = 51$ states and p between 2 and 26, so $p/n < 0.03$.[19] With weights Υ the target is $V_\Upsilon = \omega_\Upsilon'\Sigma\omega_\Upsilon$ (Section 9), and the published standard errors are the weighted cluster sandwich with Stata's factor, CR1.

[19] Columns (5) to (7) instrument the housing net worth shock with the housing supply elasticity of Saiz (2010). Their influence weights depend on the endogenous regressor, so they are not covered by Assumption A6 and are left aside. The 23 industry employment shares sum to one, and one is dropped, as Stata does.

The weights make the classical correction fail although p/n is negligible. A county's influence on the estimate grows with its number of households, and California carries two thirds of the variation in the direction of interest, $S_g = \|\omega_{\Upsilon,g}\|^2/\|\omega_\Upsilon\|^2$, so G^{eff} is 2.1 to 2.3 although $G = 51$; without the weights it would be 4.6 to 5.7. The weighted level factor κ_n is 0.49 to 0.57.[20] Under homoskedasticity the weighted cluster sandwich is therefore biased downward by about half, and the factor of CR1, between 1.02 and 1.05, does not offset it. No M_{gg} is singular, so CR3, the leave-one-state-out jackknife, exists. Every Riesz system is solved to a relative residual below 10^{-11} (Section I.2 reports the exactness checks).

Table 12 reports the standard errors and ν in Panel A and the p-values in Panel B. CR1 reproduces the published standard errors to three decimals in every column. The exact estimators are much larger. Riesz-M, the estimator proposed for weighted fits in Section 9, is 66% to 169% above CR1 in six of the seven regressions; Riesz-P(T), which for these diagonal weights is Riesz-Υ (Proposition 9.4), is 43% to 113% above it in all seven; and Riesz-F(T) is between 25% below and 204% above. CR3 is within 19% of Riesz-M in the same six regressions, but the agreement is not a confirmation. CR3 over-corrects (Proposition 7.3): its expected value is 1.24 to 1.87 times V_Υ under every covariance of Section I.2, an upward bias of 11% to 37% in the standard error. The two agree because both are driven by California's residuals, which carry 42% to 59% of each estimate. Regression (9) is the exception: California carries 95% of the Riesz-M estimate and 29% of CR3's, and Riesz-M is 21% below CR1 and half of CR3. With two effective clusters an unbiased estimator can fall far from its mean in a single realization.

The ν columns show how imprecise the fully robust estimators are: ν is between 1.3 and 1.8 for Riesz-M, 0.5 and 1.3 for Riesz-P(T), and 1.0 and 1.5 for Riesz-F(T). These ν use the identity working covariance of (31) and so favour Riesz-M. Which of the three is most precise depends on Σ (Proposition 9.5). Under covariances fitted to the residuals, with $\mathrm{Var}(u_c) \propto N_c^\eta$ for N_c the households in county c, $\eta \approx -0.44$ the slope of the log squared residuals on $\log N_c$, and a within-state correlation of 0 or 0.1, Riesz-M and Riesz-P(T) are nearly tied, and Riesz-F(T) is less precise than both in every regression, as Remark 9.3 predicts near homoskedasticity (Section I.2).

Panel B reads by critical value. With t_{G-1}, CR1 is significant at 5% in all seven regressions, and CR3 in one, (1). With normal critical values, Riesz-M is significant at 5% in (1) and (9) and at 10% in (2), (3) and (8). With t_ν, none of the three fully robust estimators is significant even at 10% in any regression (p between 0.11 and 0.47). The variance estimator and the reference distribution both change the conclusions.

The last columns use the restricted class $\mathcal{D}^{\text{het}+\rho}$ (Section 8.3), with heteroskedasticity and one within-state covariance common to all states. The data do not reject it: z_{class}, the difference between the Riesz-M and the class estimates, exactly centred under the class, over its standard deviation under the member of the class fitted to the residuals, is between -0.3 and 1.9. That member is $\hat\sigma^2\{(1-\hat\rho)I_n + \hat\rho J\}$, with J the block-diagonal matrix of within-state ones, and $\hat\sigma^2$ and $\hat\rho$ match $\widehat{U}'\widehat{U}$ and $\widehat{U}'(J - I_n)\widehat{U}$ to their exact expectations given the design; the standard deviation is the exact Gaussian one under it. The class standard errors are between 4% below and 39% above CR1's, and its ν is between 2.9 and 7.3. Class-t_ν keeps six of the seven results significant at 5% (p between

[20] For a weighted fit we define $\kappa_n := 1 - \sum_g \omega'_{\Upsilon,g}[T_\Upsilon T'_\Upsilon]_{gg}\,\omega_{\Upsilon,g}/\|\omega_\Upsilon\|^2$. It reduces to (8) when $\Upsilon = I_n$ and, by the computation of Proposition 7.3, the relative bias of the weighted CR0 under $\Sigma = \sigma^2 I_n$ is exactly $-\kappa_n$.

0.009 and 0.049); the exception is regression (9), the geographic-concentration measure with the construction interaction ($p = 0.196$). The conclusions of Mian and Sufi (2014) therefore survive, with that exception, under a restriction that the residuals do not reject, but not under fully robust inference, which two effective clusters cannot support.

Table 12: Housing net worth and employment: weights and few effective clusters

Panel A. Standard errors and degrees of freedom

		Standard error						ν			
Regression	Estimate	CR1	CR3	Riesz-M	Riesz-P(T)	Riesz-F(T)	Class	Riesz-M	Riesz-P(T)	Riesz-F(T)	Class
(1) Retail and restaurants	0.190	0.042	0.078	0.076	0.070	0.081	0.054	1.3	0.5	1.0	7.3
(2) Low geographic concentration	0.199	0.049	0.101	0.104	0.094	0.102	0.059	1.3	0.5	1.0	7.3
(3) (1) + industry shares	0.174	0.043	0.090	0.091	0.085	0.100	0.052	1.4	0.6	1.1	6.7
(4) (2) + industry shares	0.166	0.046	0.104	0.123	0.098	0.140	0.057	1.4	0.6	1.1	6.7
(8) (3) + construction share × shock	0.440	0.140	0.244	0.233	0.232	0.286	0.134	1.8	1.3	1.5	2.9
(9) (4) + construction share × shock	0.212	0.091	0.142	0.072	0.130	0.068	0.126	1.8	1.3	1.5	2.9
(10) (3) + construction growth	0.133	0.036	0.076	0.086	0.072	0.102	0.050	1.3	0.5	1.0	6.9

Panel B. p-values

	t_{G-1}		Normal			t_ν				
Regression	CR1	CR3	Riesz-M	Riesz-P(T)	Riesz-F(T)	Riesz-M	Riesz-P(T)	Riesz-F(T)	Class	z_{class}
(1) Retail and restaurants	<0.001	0.018	0.012	0.007	0.018	0.187	0.377	0.255	0.009	+0.40
(2) Low geographic concentration	<0.001	0.054	0.055	0.035	0.051	0.254	0.426	0.301	0.012	+1.10
(3) (1) + industry shares	<0.001	0.058	0.056	0.041	0.083	0.246	0.416	0.315	0.014	+0.77
(4) (2) + industry shares	<0.001	0.118	0.180	0.091	0.236	0.355	0.458	0.432	0.024	+1.89
(8) (3) + construction share × shock	0.003	0.078	0.059	0.058	0.123	0.213	0.262	0.299	0.049	+0.72
(9) (4) + construction share × shock	0.023	0.142	0.003	0.103	0.002	0.110	0.305	0.126	0.196	−0.30
(10) (3) + construction growth	<0.001	0.087	0.122	0.064	0.191	0.322	0.465	0.417	0.034	+0.67

Weighted least-squares regressions of Table III of Mian and Sufi (2014): non-tradable employment growth, 2007Q1–2009Q1, on the change in housing net worth, 2006–2009, across $n = 944$ counties, weighted by households and clustered by $G = 51$ states. Columns (1), (3), (8) and (10) use retail and restaurant employment, (2), (4) and (9) industries with low geographic concentration; the column numbers are those of the published table. CR1: the published standard error. CR3: the leave-one-state-out jackknife. Riesz-M: Proposition 9.2. Riesz-P(T) and Riesz-F(T): the minimum-norm solutions of the restricted and the full system (50) in the weighted residuals. Class: the class $\mathcal{D}^{\text{het}+\rho}$ in the OLS coordinates of Riesz-M, with target ω_Υ. ν: (31) with $\Omega_w = I_n$. z_{class}: the difference between the Riesz-M and class estimates over its standard deviation under the fitted class member $\hat{\sigma}^2\{(1-\hat{\rho})I_n + \hat{\rho}J\}$.

12 Conclusion

This paper treats robust variance estimation as a linear inverse problem. The variance of an estimator linear in the errors is a Riesz functional of the latent covariance, the residual moments are a known linear image of that covariance, and the minimum-norm representer turns the one into an estimator of the other. The resulting Riesz estimator is exactly unbiased, conditional on the design, for every covariance the maintained model admits whenever the residuals identify the variance, and otherwise for its identified part. It needs no within-cluster block to be invertible, and LSQR computes it matrix-free and reports, through its stopping rule, which case obtains.

The same system explains when the classical corrections fail. They invert it cluster by cluster and ignore how the projection couples clusters. The level of the resulting bias is the block leverage κ_n, computed from the design alone: it biases CR0 by exactly $-\kappa_n$ under homoskedasticity, and it is far from p/n when the regressors vary mainly across clusters or the regressor of interest is carried by a few of them. Heterogeneity aligned with the design biases CR2 beyond this level, and a singular M_{gg} leaves CR2 and CR3 undefined. Under least squares nothing is lost by using only the within-

cluster residual moments, since the restricted system is the Gram system of the full one; the cost is conditioning, which the solver pays in iterations.

Weights change this. The weighted fit is orthogonal in the Υ-norm rather than the Euclidean one, so the restricted system in the weighted residuals is no longer the Gram system of the full one: with special weights that couple clusters, its within-cluster products can lose information that the full array retains, and even when both identify the variance, their minimum-norm solutions differ in precision. A change of coordinates resolves both. Riesz-M works in the OLS residuals with the weighted target, and Riesz-Υ, for weights that respect the partition, in the regression transformed by the weights; both are exactly unbiased, and the coordinates affect only their precision.

The construction needs only an estimator linear in the errors, residuals that do not depend on the regression mean, and a declared dependence support. It therefore extends from one-way clustering to two-way clustering and to m-dependent serial and spatial errors, and from the fully robust model to covariance classes between a working model and full robustness, each estimator exact on its class.

Exactness has a price in sampling variance, and the paper offers two responses. A Satterthwaite degrees-of-freedom correction from the estimator's own representer restores nominal coverage when clusters are numerous. With few clusters, a declared covariance class shrinks the representer: in the simulations it has a smaller root mean squared error than CR2 while remaining unbiased on the class, and it keeps close to nominal coverage with as few as ten clusters.

The two applications show both halves of the problem. In the colonial-governor regressions of Xu (2018), many fixed effects make κ_n large, and the variance estimator alone changes the conclusions. In the county regressions of Mian and Sufi (2014), p/n is negligible, but household weights leave about two effective clusters; the variance estimator and the reference distribution both change the conclusions, which, with one exception, survive under a restricted class that the residuals do not reject. Reporting κ_n, G^{eff} and the degrees of freedom ν alongside the standard error shows which of these situations a regression is in.

Several questions remain open. Estimated weights, as in feasible GLS, and instrumental variables, whose influence weights depend on the endogenous regressor, fall outside Assumption A6: the first needs a limit theory uniform in the weights or based on cross-fitting, and the second an extension to influence weights that depend on the errors. Weights that couple clusters call for the working-covariance estimator of Proposition 9.5, for which the paper gives simulation evidence but no limit theory, and outside the two cases of Proposition 9.6 the comparison of the full and restricted systems rests on simulation. Weakly dependent errors with a growing bandwidth need mixing versions of the variance bound and the central limit theorem, and degenerate two-way designs, whose limit is not normal, need non-Gaussian critical values.

References

Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1):1–35, 2023.

John M Abowd, Francis Kramarz, and David N Margolis. High wage workers and high wage firms. *Econometrica*, 67(2):251–333, 1999.

Stanislav Anatolyev and Cheuk Fai Ng. Many covariate and cluster robust estimation and inference. *Econometric Reviews*, 45(5):703–732, 2026.

Stanislav Anatolyev and Maksim Smirnov. Off-diagonal elements of projection matrices and dimension asymptotics. *Economics Letters*, 239:111761, 2024.

Stanislav Anatolyev and Pavel Yaskov. Asymptotics of diagonal elements of projection matrices under many instruments/regressors. *Econometric Theory*, 33(3):717–738, 2017.

James E Anderson and Eric Van Wincoop. Gravity with gravitas: A solution to the border puzzle. *American economic review*, 93(1):170–192, 2003.

Donald WK Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica: Journal of the Econometric Society*, pages 817–858, 1991.

Manuel Arellano. Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics & Statistics*, 49(4):431, 1987.

Jushan Bai, Sung Hoon Choi, and Yuan Liao. Feasible generalized least squares for panel data with cross-sectional and serial correlations. *Empirical Economics*, 60(1):309–326, 2021.

Robert M Bell and Daniel F McCaffrey. Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, 28(2):169–182, 2002.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119(1):249–275, 2004.

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

Mike Brewer, Thomas F Crossley, and Robert Joyce. Inference with difference-in-differences revisited. *Journal of Econometric Methods*, 7(1):20170005, 2018.

Sebastian Calonico, Matias D Cattaneo, and Rocio Titiunik. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326, 2014.

A Colin Cameron, Jonah B Gelbach, and Douglas L Miller. Bootstrap-based improvements for inference with clustered errors. *The review of economics and statistics*, 90(3):414–427, 2008.

A Colin Cameron, Jonah B Gelbach, and Douglas L Miller. Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29(2):238–249, 2011.

David Card. The impact of the mariel boatlift on the miami labor market. *Ilr Review*, 43(2):245–257, 1990.

David Card and Alan B Krueger. Minimum wages and employment: a case study of the fast-food industry in new jersey and pennsylvania: reply. *American Economic Review*, 90(5):1397–1420, 2000.

David Card, Jörg Heining, and Patrick Kline. Workplace heterogeneity and the rise of west german wage inequality. *The Quarterly journal of economics*, 128(3):967–1015, 2013.

Raymond J Carroll. Adapting for heteroscedasticity in linear models. *The Annals of Statistics*, 10(4): 1224–1233, 1982.

Andrew V Carter, Kevin T Schnepel, and Douglas G Steigerwald. Asymptotic behavior of at-test robust to cluster heterogeneity. *Review of Economics and Statistics*, 99(4):698–709, 2017.

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

Samprit Chatterjee and Ali S Hadi. *Sensitivity analysis in linear regression*. John Wiley & Sons, 1988.

Louis HY Chen and Qi-Man Shao. Normal approximation under local dependence. *The Annuals of Probability*, (3):1985–2028, 2004.

Yanping Chen. Covariance identification from regression residuals. *arXiv preprint arXiv:2610.41722*, 2026.

Andrew Chesher and Ian Jewitt. The bias of a heteroskedasticity consistent covariance matrix estimator. *Econometrica: Journal of the Econometric Society*, pages 1217–1222, 1987.

Raj Chetty, John N Friedman, and Jonah E Rockoff. Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American economic review*, 104(9):2633–2679, 2014.

Harold D Chiang, Bruce E Hansen, and Yuya Sasaki. Standard errors for two-way clustering with serially correlated time effects. *Review of Economics and Statistics*, pages 1–40, 2024.

A Colin Cameron and Douglas L Miller. A practitioner's guide to cluster-robust inference. *Journal of human resources*, 50(2):317–372, 2015.

Timothy G Conley. Gmm estimation with cross sectional dependence. *Journal of econometrics*, 92(1): 1–45, 1999.

Timothy G Conley and Christopher R Taber. Inference with "difference in differences" with a small number of policy changes. *The Review of Economics and Statistics*, 93(1):113–125, 2011.

R Dennis Cook and Sanford Weisberg. Residuals and influence in regression. 1982.

Laurent Davezies, Xavier D'haultfœuille, and Yannick Guyonvarch. Empirical process results for exchangeable arrays. *The Annals of Statistics*, 49(2):845–862, 2021.

Antoine A Djogbenou, James G MacKinnon, and Morten Ørregaard Nielsen. Asymptotic theory and wild bootstrap inference with clustered errors. *Journal of Econometrics*, 212(2):393–412, 2019.

Will Dobbie, Jacob Goldin, and Crystal S Yang. The effects of pre-trial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review*, 108 (2):201–240, 2018.

John C Driscoll and Aart C Kraay. Consistent covariance matrix estimation with spatially dependent panel data. *Review of economics and statistics*, 80(4):549–560, 1998.

William H DuMouchel and Greg J Duncan. Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, 78(383):535–543, 1983.

Friedhelm Eicker. Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *The annals of mathematical statistics*, 34(2):447–456, 1963.

Bruno Ferman and Cristine Pinto. Inference in differences-in-differences with few treated groups and heteroskedasticity. *Review of Economics and Statistics*, 101(3):452–467, 2019.

David Chin-Lung Fong and Michael Saunders. LSMR: An iterative algorithm for sparse least-squares problems. *SIAM Journal on Scientific Computing*, 33(5):2950–2971, 2011.

Gene Golub and William Kahan. Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis*, 2(2): 205–224, 1965.

Bruce E Hansen and Seojeong Lee. Asymptotic theory for clustered samples. *Journal of econometrics*, 210(2):268–290, 2019.

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

Christian B Hansen. Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects. *Journal of econometrics*, 140(2):670–694, 2007b.

Per Christian Hansen. *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*. SIAM, 1998.

Magnus R Hestenes, Eduard Stiefel, et al. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, 49(6):409–436, 1952.

David V Hinkley. Jackknifing in unbalanced situations. *Technometrics*, 19(3):285–292, 1977.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

David C Hoaglin and Roy E Welsch. The hat matrix in regression and anova. *The American Statistician*, 32(1):17–22, 1978.

Susan D Horn, Roger A Horn, and David B Duncan. Estimating heteroscedastic variances in linear models. *Journal of the American Statistical Association*, 70(350):380–385, 1975.

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

Ilse CF Ipsen and Carl D Meyer. The idea behind krylov methods. *The American mathematical monthly*, 105(10):889–899, 1998.

Koen Jochmans. Heteroscedasticity-robust inference in linear regression models with many covariates. *Journal of the American Statistical Association*, 117(538):887–896, 2022.

Harry H Kelejian and Ingmar R Prucha. Hac estimation in a spatial framework. *Journal of Econometrics*, 140(1):131–154, 2007.

Nicholas M Kiefer and Timothy J Vogelsang. A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory*, 21(6):1130–1164, 2005.

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

Jeffrey R Kling. Incarceration length, employment, and earnings. *American Economic Review*, 96(3): 863–876, 2006.

Denis Kojevnikov, Vadim Marmer, and Kyungchul Song. Limit theorems for network dependent random variables. *Journal of Econometrics*, 222(2):882–908, 2021.

Michal Kolesár, Pengjin Min, Wenjie Wang, and Yichong Zhang. Cluster-robust inference for quadratic forms. *arXiv preprint arXiv:2602.13537*, 2026.

Cornelius Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of research of the National Bureau of Standards*, 45(4):255–282, 1950.

Eben Lazarus, Daniel J Lewis, James H Stock, and Mark W Watson. Har inference: Recommendations for practice. *Journal of Business & Economic Statistics*, 36(4):541–559, 2018.

Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *biometrika*, pages 13–22, 1986.

James G MacKinnon and Matthew D Webb. Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics*, 32(2):233–254, 2017.

James G MacKinnon and Matthew D Webb. The wild bootstrap for few (treated) clusters. *The Econometrics Journal*, 21(2):114–135, 2018.

James G MacKinnon and Matthew D Webb. Randomization inference for difference-in-differences with few treated clusters. *Journal of Econometrics*, 218(2):435–450, 2020.

James G MacKinnon and Halbert White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics*, 29(3):305–325, 1985.

James G MacKinnon, Morten Ørregaard Nielsen, and Matthew D Webb. Wild bootstrap and asymptotic inference with multiway clustering. *Journal of Business & Economic Statistics*, 39(2):505–519, 2021.

James G MacKinnon, Morten Ørregaard Nielsen, and Matthew D Webb. Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics*, 232(2):272–299, 2023a.

James G MacKinnon, Morten Ørregaard Nielsen, and Matthew D Webb. Fast and reliable jackknife and bootstrap methods for cluster-robust inference. *Journal of Applied Econometrics*, 38(5):671–694, 2023b.

James G MacKinnon, Morten Ørregaard Nielsen, and Matthew D Webb. Leverage, influence, and the jackknife in clustered regression models: Reliable inference using summclust. *The Stata Journal*, 23(4):942–982, 2023c.

Konrad Menzel. Bootstrap with cluster-dependence in two or more dimensions. *Econometrica*, 89 (5):2143–2188, 2021.

Atif Mian and Amir Sufi. What explains the 2007–2009 drop in employment? *Econometrica*, 82(6): 2197–2223, 2014.

Mohammad Mohammadi. On the bounds for diagonal and off-diagonal elements of the hat matrix in the linear regression model. *REVSTAT-Statistical Journal*, 14(1):75–87, 2016.

Brent R Moulton. An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The review of Economics and Statistics*, pages 334–338, 1990.

Whitney K Newey and Kenneth D West. A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix. 1986.

Christopher C Paige and Michael A Saunders. Lsqr: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software (TOMS)*, 8(1):43–71, 1982.

James E Pustejovsky and Elizabeth Tipton. Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, 36 (4):672–683, 2018.

C Radhakrishna Rao. Estimation of heteroscedastic variances in linear models. *Journal of the American Statistical Association*, 65(329):161–172, 1970.

C Radhakrishna Rao. Estimation of variance and covariance components—minque theory. *Journal of multivariate analysis*, 1(3):257–275, 1971.

Peter M Robinson. Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica: Journal of the Econometric Society*, pages 875–891, 1987.

Joseph P. Romano and Michael Wolf. Resurrecting weighted least squares. *Journal of Econometrics*, 197(1):1–19, 2017.

Nathan Ross. Fundamentals of Stein's method. *Probability Surveys*, 8:210–293, 2011.

Yousef Saad. *Iterative methods for sparse linear systems*. SIAM, 2003.

Albert Saiz. The geographic determinants of housing supply. *Quarterly Journal of Economics*, 125 (3):1253–1296, 2010.

Franklin E Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics bulletin*, 2(6):110–114, 1946.

Gary Solon, Steven J. Haider, and Jeffrey M. Wooldridge. What are we weighting for? *Journal of Human Resources*, 50(2):301–316, 2015.

Trond Steihaug. The conjugate gradient method and trust regions in large scale optimization. *SIAM Journal on Numerical Analysis*, 20(3):626–637, 1983.

James H Stock and Mark W Watson. Heteroskedasticity-robust standard errors for fixed effects panel data regression. *Econometrica*, 76(1):155–174, 2008.

Samuel B Thompson. Simple formulas for standard errors that cluster by both firm and time. *Journal of financial Economics*, 99(1):1–10, 2011.

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

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, pages 817–838, 1980.

Jeffrey M Wooldridge. Cluster-sample methods in applied econometrics. *American Economic Review*, 93(2):133–138, 2003.

Guo Xu. The costs of patronage: Evidence from the british empire. *American Economic Review*, 108 (11):3170–3198, 2018.

Alwyn Young. Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The quarterly journal of economics*, 134(2):557–598, 2019.

Appendix

A Preliminary Lemmas

Lemma A.1 (Operator norms of $\mathcal{A}_n$ and $\mathcal{A}_n^*$)**.** *For every design, the operator $\mathcal{A}_n$ of (17) and its adjoint (18) satisfy*

$$\|\mathcal{A}_n\|_{\mathrm{op}} \;\le\; 1, \qquad \|\mathcal{A}_n^*\|_{\mathrm{op}} \;\le\; 1.$$

Proof. By Section 3, $M = I_n - \Pi_{[X\,W]}$ is an orthogonal projector, so $\|M\|_{\mathrm{op}} \le 1$. Read $\Sigma \in \mathcal{H}_n$ as the block-diagonal matrix $\mathrm{diag}(\Sigma_1, \dots, \Sigma_G)$. Since its off-diagonal blocks are zero, the definition of $\langle \cdot, \cdot \rangle_{\mathcal{H}_n}$ gives $\|\Sigma\|_{\mathcal{H}_n} = n^{-1}\|\Sigma\|_F$. Since $[M\Sigma M]_{hh'} = \sum_g M_{hg}\Sigma_g M_{gh'} = [\mathcal{A}_n(\Sigma)]_{hh'}$, the matrix $M\Sigma M$ is $\mathcal{A}_n(\Sigma)$, and so $\|\mathcal{A}_n(\Sigma)\|_{\mathcal{K}_n} = n^{-1}\|M\Sigma M\|_F$. Since $\|BCD\|_F \le \|B\|_{\mathrm{op}}\|C\|_F\|D\|_{\mathrm{op}}$ for conformable matrices,

$$\|\mathcal{A}_n(\Sigma)\|_{\mathcal{K}_n} = \frac{1}{n}\|M\Sigma M\|_F \le \frac{1}{n}\|M\|_{\mathrm{op}}^2\|\Sigma\|_F \le \|\Sigma\|_{\mathcal{H}_n},$$

so $\|\mathcal{A}_n\|_{\mathrm{op}} \le 1$. For the adjoint, (18) gives $[\mathcal{A}_n^*(D)]_g = \sum_{h,h'} M_{gh}D_{hh'}M_{h'g} = [MDM]_{gg}$, so $\mathcal{A}_n^*(D)$ is the block-diagonal part of MDM. Dropping the off-diagonal blocks does not increase the Frobenius norm, and so, by the same bound,

$$\|\mathcal{A}_n^*(D)\|_{\mathcal{H}_n} \le \frac{1}{n}\|MDM\|_F \le \frac{1}{n}\|D\|_F = \|D\|_{\mathcal{K}_n},$$

so $\|\mathcal{A}_n^*\|_{\mathrm{op}} \le 1$. □

Lemma A.2. *Suppose Assumption A2 holds. Then, for any $1 \le \theta \le 2 + \lambda$,*

(i) $\sup_g N_g^{-\theta}\,\mathbb{E}\big\|\widetilde{X}_g'\widetilde{X}_g\big\|_F^\theta = O(1)$, *and consequently* $\sum_{g=1}^G \big\|\widetilde{X}_g'\widetilde{X}_g\big\|_F^\theta = O_p\Big(\sum_{g=1}^G N_g^\theta\Big)$;

(ii) $\sup_g N_g^{-\theta}\,\mathbb{E}\big[\|U_gU_g'\|_F^\theta \mid X, W\big] = O(1)$;

(iii) $\sum_{g=1}^G \mathbb{E}\big[\|\widetilde{X}_g'U_g\|^\theta \mid X, W\big] = O_p\Big(\sum_{g=1}^G N_g^\theta\Big)$.

Proof. Two elementary inequalities are used throughout. Since $\theta \ge 1$, the map $t \mapsto t^\theta$ is convex on $[0, \infty)$, so Jensen's inequality applied to the average of nonnegative numbers $a_1, \dots, a_N$ gives

$$\Big(\sum_{i=1}^N a_i\Big)^\theta \le N^{\theta-1}\sum_{i=1}^N a_i^\theta. \tag{A.1}$$

Since $\theta \le 2 + \lambda$, so that $2\theta \le 4 + 2\lambda$, Lyapunov's inequality gives $\mathbb{E}Y^{2\theta} \le (\mathbb{E}Y^{4+2\lambda})^{\theta/(2+\lambda)}$ for every nonnegative random variable Y, and the same holds for conditional expectations. Let $C_{\tilde{x}} := \sup_g N_g^{-1}\sum_i \mathbb{E}\|\tilde{x}_{gi}\|^{4+2\lambda}$, and let C_u be the finite constant that bounds $\sup_{g,i}\mathbb{E}[|u_{gi}|^{4+2\lambda} \mid X, W]$ almost surely; both exist by Assumption A2.

(i) Write $\widetilde{X}_g'\widetilde{X}_g = \sum_i \tilde{x}_{gi}\tilde{x}_{gi}'$. By the triangle inequality and $\|vv'\|_F = \|v\|^2$, $\|\widetilde{X}_g'\widetilde{X}_g\|_F \le \sum_i \|\tilde{x}_{gi}\|^2$. Applying (A.1) with $a_i = \|\tilde{x}_{gi}\|^2$ and $N = N_g$,

$$N_g^{-\theta}\big\|\widetilde{X}_g'\widetilde{X}_g\big\|_F^{\theta} \;\le\; N_g^{-\theta}N_g^{\theta-1}\sum_{i=1}^{N_g}\|\tilde{x}_{gi}\|^{2\theta} = \frac{1}{N_g}\sum_{i=1}^{N_g}\|\tilde{x}_{gi}\|^{2\theta}.$$

Taking the expectation for every g, the Lyapunov's inequality gives,

$$N_g^{-\theta}\,\mathbb{E}\big\|\widetilde{X}_g'\widetilde{X}_g\big\|_F^{\theta} \;\le\; \Big(\frac{1}{N_g}\sum_{i=1}^{N_g}\mathbb{E}\|\tilde{x}_{gi}\|^{4+2\lambda}\Big)^{\frac{\theta}{2+\lambda}} \;\le\; C_{\tilde{x}}^{\frac{\theta}{2+\lambda}},$$

which is the first claim. For the second, let $S := \sum_g \|\widetilde{X}_g'\widetilde{X}_g\|_F^{\theta}$. Summing the bound over g gives $\mathbb{E}S \le C_{\tilde{x}}^{\theta/(2+\lambda)}\sum_g N_g^{\theta}$, and Markov's inequality gives, for every $K > 0$,

$$\Pr\Big(S > K\sum_{g=1}^{G}N_g^{\theta}\Big) \le \frac{C_{\tilde{x}}^{\theta/(2+\lambda)}}{K},$$

which is arbitrarily small for K large. Hence $S = O_p(\sum_g N_g^{\theta})$.

(ii) Here the first step is an equality, $\|U_gU_g'\|_F = \|U_g\|^2 = \sum_i u_{gi}^2$. Applying (A.1) with $a_i = u_{gi}^2$ gives $N_g^{-\theta}\|U_gU_g'\|_F^{\theta} \le N_g^{-1}\sum_i |u_{gi}|^{2\theta}$. Taking conditional expectations and applying the conditional Lyapunov inequality to each i, $\mathbb{E}[|u_{gi}|^{2\theta} \mid X, W] \le (\mathbb{E}[|u_{gi}|^{4+2\lambda} \mid X, W])^{\theta/(2+\lambda)} \le C_u^{\theta/(2+\lambda)}$. Averaging over i gives, for every g and almost surely,

$$N_g^{-\theta}\,\mathbb{E}\big[\|U_gU_g'\|_F^{\theta} \mid X, W\big] \;\le\; \frac{1}{N_g}\sum_{i=1}^{N_g}\mathbb{E}\big[|u_{gi}|^{2\theta} \mid X, W\big] \;\le\; C_u^{\frac{\theta}{2+\lambda}}.$$

(iii) For a matrix A and a vector v, $\|A'v\|^2 = v'AA'v = \mathrm{tr}(AA'vv') \le \|AA'\|_F\|vv'\|_F$, by the Cauchy–Schwarz inequality for the Frobenius inner product. Take $A = \widetilde{X}_g$ and $v = U_g$. Since $\|AA'\|_F^2 = \mathrm{tr}(AA'AA') = \mathrm{tr}(A'AA'A) = \|A'A\|_F^2$ by cyclicity of the trace, raising to the power $\theta/2$,

$$\|\widetilde{X}_g'U_g\|^{\theta} \;\le\; \|\widetilde{X}_g'\widetilde{X}_g\|_F^{\theta/2}\,\|U_gU_g'\|_F^{\theta/2}.$$

Since $\widetilde{X} = M_W X$ is a function of (X, W), the first factor comes out of the conditional expectation. For the second, the conditional Jensen inequality gives $\mathbb{E}[Z^{\theta/2} \mid X, W] \le (\mathbb{E}[Z^{\theta} \mid X, W])^{1/2}$ for $Z \ge 0$, and part (ii) bounds $\mathbb{E}[\|U_gU_g'\|_F^{\theta} \mid X, W]$ by $C_u^{\theta/(2+\lambda)}N_g^{\theta}$. Hence

$$\mathbb{E}\big[\|\widetilde{X}_g'U_g\|^{\theta} \mid X, W\big] \;\le\; C_u^{\frac{\theta}{4+2\lambda}}\,N_g^{\theta/2}\,\|\widetilde{X}_g'\widetilde{X}_g\|_F^{\theta/2}.$$

Summing over g and applying the Cauchy–Schwarz inequality to the sum,

$$\sum_{g=1}^{G}\mathbb{E}\big[\|\widetilde{X}_g'U_g\|^{\theta} \mid X, W\big] \;\le\; C_u^{\frac{\theta}{4+2\lambda}}\Big(\sum_{g=1}^{G}N_g^{\theta}\Big)^{1/2}\Big(\sum_{g=1}^{G}\big\|\widetilde{X}_g'\widetilde{X}_g\big\|_F^{\theta}\Big)^{1/2},$$

and the last factor is $O_p((\sum_g N_g^{\theta})^{1/2})$ by part (i), which gives the claim. □

Lemma A.3 (Boundedness of Q_n and $R_n^\star$). *Let Q_n and $R_n^\star$ be as in Assumption A5, for a fixed contrast a. Under Assumptions A2 and A4,*

$$\|Q_n\|_F = O_p(1) \qquad \text{and} \qquad \|R_n^\star\|_{\mathcal{H}_n} = O_p\Big(\sqrt{\frac{\sup_g N_g}{n}}\Big).$$

Proof. Q_n. By Assumption A4, $\Gamma_n \to_p \Gamma_0 \succ 0$. The smallest eigenvalue is a continuous function of a symmetric matrix, so $\lambda_{\min}(\Gamma_n) \to_p \lambda_{\min}(\Gamma_0) > 0$. Hence Γ_n is invertible with probability approaching one, which is the event on which Q_n and $R_n^\star$ are defined, and $\|\Gamma_n^{-1}\|_{\mathrm{op}} = 1/\lambda_{\min}(\Gamma_n) = O_p(1)$. Write $v_n := \Gamma_n^{-1} a$, so that $Q_n = v_n v_n'$ has rank one. Then

$$\|Q_n\|_F = \|v_n v_n'\|_F = \|v_n\|^2 \le \|\Gamma_n^{-1}\|_{\mathrm{op}}^2 \|a\|^2 = O_p(1),$$

since the contrast a is fixed.

$R_n^\star$. Each block is again of rank one: $R_{n,g}^\star = \widetilde{X}_g v_n v_n' \widetilde{X}_g' = z_g z_g'$ with $z_g := \widetilde{X}_g v_n$. Hence, using $\mathrm{tr}(v'Bv) = \mathrm{tr}(Bvv')$ and the Cauchy–Schwarz inequality for the Frobenius inner product,

$$\|R_{n,g}^\star\|_F = \|z_g\|^2 = \mathrm{tr}\left(v_n' \widetilde{X}_g' \widetilde{X}_g v_n\right) = \mathrm{tr}\left(\widetilde{X}_g' \widetilde{X}_g\, Q_n\right) \le \|Q_n\|_F \left\|\widetilde{X}_g' \widetilde{X}_g\right\|_F.$$

Squaring and summing over g, by the definition of $\|\cdot\|_{\mathcal{H}_n}$,

$$\|R_n^\star\|_{\mathcal{H}_n}^2 = \frac{1}{n^2} \sum_{g=1}^G \|R_{n,g}^\star\|_F^2 \le \frac{\|Q_n\|_F^2}{n^2} \sum_{g=1}^G \left\|\widetilde{X}_g' \widetilde{X}_g\right\|_F^2.$$

Lemma A.2(i) with $\theta = 2$, admissible since $\lambda > 0$, gives $\sum_g \|\widetilde{X}_g' \widetilde{X}_g\|_F^2 = O_p(\sum_g N_g^2)$, and $\sum_g N_g^2 \le \sup_g N_g \sum_g N_g = n \sup_g N_g$. Since $\|Q_n\|_F^2 = O_p(1)$ and a product of O_p terms is O_p of the product,

$$\|R_n^\star\|_{\mathcal{H}_n}^2 = O_p\Big(\frac{1}{n^2} \sum_{g=1}^G N_g^2\Big) = O_p\Big(\frac{\sup_g N_g}{n}\Big),$$

and taking square roots gives the claim. □

Lemma A.4 (Boundedness of Σ_n and C_n). *Under Assumptions A1 and A2,*

$$\|C_n\|_{\mathcal{K}_n} \;\le\; \|\Sigma_n\|_{\mathcal{H}_n} \;=\; O_p\Big(\sqrt{\tfrac{\sup_g N_g}{n}}\Big).$$

Proof. Σ_n. Since $\Sigma_g = \mathbb{E}[U_g U_g' \mid X, W]$ and the Frobenius norm is convex, Jensen's inequality and Lemma A.2(ii) with $\theta = 1$ give, almost surely,

$$\|\Sigma_g\|_F \le \mathbb{E}\big[\|U_g U_g'\|_F \mid X, W\big] \le N_g\, C_u^{1/(2+\lambda)},$$

with C_u as in the proof of Lemma A.2. Squaring, summing over g and using $\sum_g N_g^2 \le n \sup_g N_g$,

$$\|\Sigma_n\|_{\mathcal{H}_n}^2 = \frac{1}{n^2} \sum_{g=1}^G \|\Sigma_g\|_F^2 \le C_u^{\frac{2}{2+\lambda}} \frac{1}{n^2} \sum_{g=1}^G N_g^2 \le C_u^{\frac{2}{2+\lambda}} \frac{\sup_g N_g}{n}$$

almost surely. Taking square roots,

$$\|\Sigma_n\|_{\mathcal{H}_n} \le C_u^{1/(2+\lambda)} \Big(\frac{\sup_g N_g}{n}\Big)^{1/2} \quad \text{almost surely,}$$

which implies the stated O_p bound.

C_n. By Assumption A1, the errors have conditional mean zero and are independent across clusters, so $\mathbb{E}[U_g U_{g'}' \mid X, W] = 0$ for $g \neq g'$ and $\mathbb{E}[UU' \mid X, W] = \mathrm{diag}(\Sigma_1, \dots, \Sigma_G)$. Since $\widehat{U} = MU$ with M a function of (X, W), $C_n = \mathbb{E}[\widehat{U}\widehat{U}' \mid X, W] = M\ \mathrm{diag}(\Sigma_1, \dots, \Sigma_G)\, M = \mathcal{A}_n(\Sigma_n)$, as in the proof of Lemma A.1. Hence $\|C_n\|_{\mathcal{K}_n} \le \|\mathcal{A}_n\|_{\mathrm{op}} \|\Sigma_n\|_{\mathcal{H}_n} \le \|\Sigma_n\|_{\mathcal{H}_n}$ by Lemma A.1. □

Lemma A.5 (Boundedness of the minimum-norm representer)**.** *Let $D_n^\star = (\mathcal{A}_n^*)^+ R_n^\star$ be the representer of Definition 4.3 and $s_n = 1/\sigma_{\min}^+(\mathcal{A}_n)$. Then $\|D_n^\star\|_{\mathcal{K}_n} \le s_n \|R_n^\star\|_{\mathcal{H}_n}$ for every design, and under Assumptions A2 and A4*

$$\|D_n^\star\|_{\mathcal{K}_n} \;=\; O_p\Big(s_n \sqrt{\frac{\sup_g N_g}{n}}\Big).$$

Proof. If $\mathcal{A}_n = 0$, then $\mathcal{A}_n^* = 0$, $D_n^\star = 0$, and the bound holds with the convention $s_n = \infty$. Since $\mathcal{A}_n(I_n) = M^2 = M$, $\mathcal{A}_n = 0$ only if $M = 0$, which the rank condition of Assumption A4 excludes with probability approaching one. Otherwise, $\mathcal{A}_n^* \colon \mathcal{K}_n \to \mathcal{H}_n$ is a nonzero linear map between finite-dimensional inner-product spaces, so it has a singular value decomposition

$$\mathcal{A}_n^*(D) = \sum_{j=1}^{r} \sigma_j \, \langle D, e_j \rangle_{\mathcal{K}_n} \, f_j, \qquad \sigma_1 \ge \cdots \ge \sigma_r > 0,$$

with $\{e_j\}$ orthonormal in $\mathcal{K}_n$ and $\{f_j\}$ orthonormal in $\mathcal{H}_n$. The defining identity of the adjoint then gives $\mathcal{A}_n(\Sigma) = \sum_j \sigma_j \langle \Sigma, f_j \rangle_{\mathcal{H}_n} e_j$, so $\mathcal{A}_n$ and $\mathcal{A}_n^*$ have the same nonzero singular values, and $\sigma_{\min}^+(\mathcal{A}_n) = \sigma_r = 1/s_n$. The Moore–Penrose inverse inverts each singular value and annihilates the orthogonal complement of $\mathrm{span}\{f_j\}$,

$$(\mathcal{A}_n^*)^+(R) = \sum_{j=1}^{r} \sigma_j^{-1} \, \langle R, f_j \rangle_{\mathcal{H}_n} \, e_j.$$

By orthonormality of the e_j and Bessel's inequality,

$$\|D_n^\star\|_{\mathcal{K}_n}^2 = \sum_{j=1}^{r} \sigma_j^{-2} \langle R_n^\star, f_j \rangle_{\mathcal{H}_n}^2 \le \sigma_r^{-2} \sum_{j=1}^{r} \langle R_n^\star, f_j \rangle_{\mathcal{H}_n}^2 \le s_n^2 \, \|R_n^\star\|_{\mathcal{H}_n}^2,$$

whether or not $R_n^\star \in \mathrm{Range}(\mathcal{A}_n^*)$.

Since s_n depends on the design, the claim means that the left side below is $O_p(1)$; by the deterministic bound above,

$$\frac{\|D_n^\star\|_{\mathcal{K}_n}}{s_n \sqrt{\sup_g N_g / n}} \;\le\; \frac{\|R_n^\star\|_{\mathcal{H}_n}}{\sqrt{\sup_g N_g / n}},$$

and the right side is $O_p(1)$ by Lemma A.3 under Assumptions A2 and A4. □

B Proofs for Section 4

Proof of Proposition 4.1

Proof. The representer lies in $\mathcal{H}$. Since Γ is nonsingular, $Q = \Gamma^{-1} a a' \Gamma^{-1}$ is defined, and it is symmetric. Hence each $R_g^\star := \widetilde{X}_g Q \widetilde{X}_g'$ is symmetric, $(R_g^\star)' = \widetilde{X}_g Q' \widetilde{X}_g' = R_g^\star$, and $R^\star := \{R_g^\star\}_{g=1}^G \in \mathcal{H}$.

Representation, linearity and boundedness. Every $\Sigma \in \mathcal{H}$ has symmetric blocks, so $\operatorname{tr}(\Sigma_g R_g^\star) = \operatorname{tr}(\Sigma_g' R_g^\star)$, and by the definition of $\langle \cdot, \cdot \rangle_{\mathcal{H}}$,

$$\Phi(\Sigma) = \frac{1}{n^2} \sum_{g=1}^{G} \operatorname{tr}\left(\Sigma_g' R_g^\star\right) = \langle \Sigma, R^\star \rangle_{\mathcal{H}} \qquad \text{for all } \Sigma \in \mathcal{H}.$$

The inner product is linear in its first argument, so Φ is linear, and by the Cauchy–Schwarz inequality $|\Phi(\Sigma)| \le \|R^\star\|_{\mathcal{H}} \|\Sigma\|_{\mathcal{H}}$, so Φ is bounded.

Uniqueness. If $R \in \mathcal{H}$ also satisfies $\Phi(\Sigma) = \langle \Sigma, R \rangle_{\mathcal{H}}$ for all $\Sigma \in \mathcal{H}$, then $\langle \Sigma, R - R^\star \rangle_{\mathcal{H}} = 0$ for all $\Sigma \in \mathcal{H}$. Since $R - R^\star \in \mathcal{H}$, taking $\Sigma = R - R^\star$ gives $\|R - R^\star\|_{\mathcal{H}}^2 = 0$, so $R = R^\star$. □

Proof of Lemma 4.2

Proof. The map (18) sends $\mathcal{K}$ into $\mathcal{H}$. For $D \in \mathcal{K}$, $[\mathcal{A}^*(D)]_g = \sum_{h,h'} M_{gh} D_{hh'} M_{h'g} = [MDM]_{gg}$, so $\mathcal{A}^*(D)$ collects the diagonal blocks of MDM. Since M and D are symmetric, so is MDM, and hence each diagonal block; thus $\mathcal{A}^*(D) \in \mathcal{H}$, and $\mathcal{A}^*$ is linear in D.

The adjoint identity. Read $\Sigma \in \mathcal{H}$ as the block-diagonal matrix $\operatorname{diag}(\Sigma_1, \ldots, \Sigma_G)$, so that $\mathcal{A}(\Sigma) = M\Sigma M$ by (17). For $D \in \mathcal{K}$, the symmetry of $M\Sigma M$, the cyclicity of the trace and the block-diagonal form of Σ give

$$\langle \mathcal{A}(\Sigma), D \rangle_{\mathcal{K}} = \frac{1}{n^2} \operatorname{tr}(M\Sigma M D) = \frac{1}{n^2} \operatorname{tr}(\Sigma\, MDM) = \frac{1}{n^2} \sum_{g=1}^{G} \operatorname{tr}\left(\Sigma_g\, [MDM]_{gg}\right).$$

Each Σ_g is symmetric, so $\operatorname{tr}(\Sigma_g [MDM]_{gg}) = \operatorname{tr}(\Sigma_g' [MDM]_{gg})$, and the right side is $\langle \Sigma, \mathcal{A}^*(D) \rangle_{\mathcal{H}}$.

Uniqueness. If $\mathcal{B} \colon \mathcal{K} \to \mathcal{H}$ also satisfies the identity, then $\langle \Sigma, \mathcal{A}^*(D) - \mathcal{B}(D) \rangle_{\mathcal{H}} = 0$ for all $\Sigma \in \mathcal{H}$ and $D \in \mathcal{K}$. Taking $\Sigma = \mathcal{A}^*(D) - \mathcal{B}(D) \in \mathcal{H}$ gives $\mathcal{B}(D) = \mathcal{A}^*(D)$ for every D. □

Remark. Both proofs take the direct route: they write down the representer and the adjoint explicitly and verify them by a trace computation. In finite dimension this proves the Riesz representation theorem for the maps at hand, since exhibiting an element that satisfies the identity gives existence, and testing the identity against the difference of two such elements gives uniqueness.

Proof of Proposition 4.4

Proof. Since $\mathcal{K} = \mathbb{S}^n$, D, $D^\star$ and Z are symmetric $n \times n$ matrices; write $\langle A, B \rangle_F = \operatorname{tr}(A'B)$, so that $\langle \cdot, \cdot \rangle_{\mathcal{K}} = n^{-2} \langle \cdot, \cdot \rangle_F$ and orthogonality is the same under both.

The norm identity. The minimum-norm solution has no component in $\operatorname{Ker}(\mathcal{A}^*)$, since removing one would leave $\mathcal{A}^* D^\star$ unchanged and reduce the norm. Hence $D^\star \in \operatorname{Ker}(\mathcal{A}^*)^\perp = \operatorname{Range}(\mathcal{A})$, so

$D^\star = \mathcal{A}(H) = MHM$ for some $H \in \mathcal{H}$ by (17), and $MD^\star M = D^\star$ because M is idempotent. Therefore $MDM = D^\star + MZM$. The cross term vanishes,

$$\langle D^\star, MZM\rangle_F = \operatorname{tr}(MD^\star M\, Z) = \langle D^\star, Z\rangle_F = 0,$$

because $Z \in \operatorname{Ker}(\mathcal{A}^*) = \operatorname{Range}(\mathcal{A})^\perp$ and $D^\star \in \operatorname{Range}(\mathcal{A})$. This gives $\|MDM\|_F^2 = \|D^\star\|_F^2 + \|MZM\|_F^2$.

The variance under independent errors. Since $\widehat{C} = \widehat{U}\widehat{U}'$ and $\widehat{U} = MU$, symmetry of D gives $\widehat{V} = \langle \widehat{C}, D\rangle_{\mathcal{K}} = n^{-2}\widehat{U}'D\widehat{U} = U'SU$ with $S := n^{-2}MDM$, which is symmetric and (X, W)-measurable. Let the errors be independent given (X, W), with mean zero, common variance σ^2 and fourth moments $\mu_{4,i}$, and split the quadratic form into its diagonal and off-diagonal parts,

$$U'SU = T_1 + T_2, \qquad T_1 := \sum_{i=1}^n S_{ii}u_i^2, \qquad T_2 := \sum_{i\neq j} S_{ij}u_iu_j.$$

By independence, $\operatorname{Var}(T_1 \mid X, W) = \sum_i S_{ii}^2 \operatorname{Var}(u_i^2 \mid X, W) = \sum_i S_{ii}^2(\mu_{4,i} - \sigma^4)$. The term T_2 has mean zero, and for $i \neq j$ and $k \neq l$, $\mathbb{E}[u_iu_ju_ku_l \mid X, W]$ is nonzero only if $\{k, l\} = \{i, j\}$, where it equals σ^4; since S is symmetric, each unordered pair is counted twice, and $\operatorname{Var}(T_2 \mid X, W) = 2\sigma^4 \sum_{i\neq j} S_{ij}^2$. The two parts are uncorrelated, because $\mathbb{E}[u_k^2u_iu_j \mid X, W] = 0$ whenever $i \neq j$: one of u_i, u_j appears alone and has mean zero. Adding, and using $\sum_{i\neq j} S_{ij}^2 = \|S\|_F^2 - \sum_i S_{ii}^2$,

$$\begin{aligned}\operatorname{Var}(\widehat{V} \mid X, W) &= \sum_{i=1}^n S_{ii}^2\big(\mu_{4,i} - \sigma^4\big) + 2\sigma^4 \sum_{i\neq j} S_{ij}^2 \\ &= 2\sigma^4\|S\|_F^2 + \sum_{i=1}^n \big(\mu_{4,i} - 3\sigma^4\big)S_{ii}^2.\end{aligned}$$

When $\mu_{4,i} = 3\sigma^4$ for every i, as for normal errors, the second term vanishes, so $\operatorname{Var}(\widehat{V} \mid X, W) = 2\sigma^4 n^{-4}\|MDM\|_F^2$, which the norm identity minimizes at $Z = 0$.

The bound for general errors. Under Assumption A1 the clusters are independent, so the errors form a dependency graph whose neighbourhoods are the clusters, with $d_n = \sup_g N_g$; under Assumption A2, Lyapunov's inequality bounds $m_4 := \sup_i \mathbb{E}[u_i^4 \mid X, W]$. Lemma 8.1, with $\|S\|_F^2 = n^{-4}\|MDM\|_F^2 = n^{-2}\|MDM\|_{\mathcal{K}}^2$, gives

$$\operatorname{Var}(\widehat{V} \mid X, W) \le C_0 m_4 \Big(\sup_g N_g\Big)^2 \|S\|_F^2 = C_0 m_4 \Big(\sup_g N_g\Big)^2 n^{-2}\|MDM\|_{\mathcal{K}}^2.$$

The constant $C_0 m_4 (\sup_g N_g)^2$ does not depend on D, so the norm identity again gives the minimum at $D = D^\star$. □

Proof of Proposition 4.5

Proof. No step below uses $R^\star \in \operatorname{Range}(\mathcal{A}^*)$, which is why neither identity requires it.

The bias identity. Since $C = \mathcal{A}(\Sigma)$, Lemma 4.2 gives $\langle C, D\rangle_{\mathcal{K}} = \langle \Sigma, \mathcal{A}^*(D)\rangle_{\mathcal{H}}$, and Proposition 4.1 gives $\Phi(\Sigma) = \langle \Sigma, R^\star\rangle_{\mathcal{H}}$. Subtracting gives (23), and the Cauchy–Schwarz inequality gives the bound that follows it.

The minimum-norm representer. By the defining property of the Moore–Penrose inverse, $\mathcal{A}^*(\mathcal{A}^*)^+$ is the orthogonal projection onto $\mathrm{Range}(\mathcal{A}^*)$. As in the proof of Proposition 4.4, with the roles of $\mathcal{A}$ and $\mathcal{A}^*$ exchanged, $\mathrm{Range}(\mathcal{A}^*) = \mathrm{Ker}(\mathcal{A})^\perp = \mathcal{N}^\perp$. Hence

$$\mathcal{A}^*(D^\star) = \mathcal{A}^*(\mathcal{A}^*)^+ R^\star = \Pi_{\mathcal{N}^\perp} R^\star = (I - \Pi_{\mathcal{N}}) R^\star.$$

Substituting into (23), using that the orthogonal projection $\Pi_{\mathcal{N}}$ is self-adjoint, and applying Proposition 4.1 to $\Pi_{\mathcal{N}}\Sigma \in \mathcal{N} \subseteq \mathcal{H}$,

$$\langle C, D^\star\rangle_{\mathcal{K}} - \Phi(\Sigma) = -\big\langle \Sigma, \Pi_{\mathcal{N}} R^\star\big\rangle_{\mathcal{H}} = -\big\langle \Pi_{\mathcal{N}}\Sigma, R^\star\big\rangle_{\mathcal{H}} = -\Phi\big(\Pi_{\mathcal{N}}\Sigma\big),$$

which, with the linearity of Φ, is (24).

The floor of footnote 8. For any $D \in \mathcal{K}$, both $\mathcal{A}^*(D)$ and $(I - \Pi_{\mathcal{N}})R^\star$ lie in $\mathcal{N}^\perp$, while $\Pi_{\mathcal{N}} R^\star \in \mathcal{N}$. The decomposition $\mathcal{A}^*(D) - R^\star = \{\mathcal{A}^*(D) - (I - \Pi_{\mathcal{N}})R^\star\} - \Pi_{\mathcal{N}} R^\star$ is therefore orthogonal, and by Pythagoras

$$\|\mathcal{A}^*(D) - R^\star\|_{\mathcal{H}}^2 = \big\|\mathcal{A}^*(D) - (I - \Pi_{\mathcal{N}})R^\star\big\|_{\mathcal{H}}^2 + \|\Pi_{\mathcal{N}} R^\star\|_{\mathcal{H}}^2 \geq \delta^2,$$

with equality at $D = D^\star$. □

Proof of Theorem 4.6

Proof. Let $D \in \mathcal{K}$ depend on (X, W) only. By Assumption A1 the errors have conditional mean zero and are independent across clusters, so $\mathbb{E}[UU' \mid X, W] = \mathrm{diag}(\Sigma_1, \ldots, \Sigma_G)$, and since $\widehat{C} = \widehat{U}\widehat{U}' = MUU'M$ with M a function of (X, W), $\mathbb{E}[\widehat{C} \mid X, W] = M\Sigma M = C$. The inner product is linear, so

$$\mathbb{E}\big[\langle \widehat{C}, D\rangle_{\mathcal{K}} \mid X, W\big] = \langle C, D\rangle_{\mathcal{K}}.$$

With $D = D^\star$, (24) gives $\mathbb{E}[\widehat{V}_{\mathrm{Riesz}} \mid X, W] = \Phi\big((I - \Pi_{\mathcal{N}})\Sigma\big) = \Phi(\Sigma) - \Phi(\Pi_{\mathcal{N}}\Sigma)$ for every $\Sigma \in \mathcal{H}$, whether or not the range condition holds; this is the second claim. The estimator is therefore unbiased for every $\Sigma \in \mathcal{H}$ if and only if $\Phi(\Pi_{\mathcal{N}}\Sigma) = \langle \Sigma, \Pi_{\mathcal{N}} R^\star\rangle_{\mathcal{H}}$ vanishes for every $\Sigma \in \mathcal{H}$. Taking $\Sigma = \Pi_{\mathcal{N}} R^\star \in \mathcal{H}$ shows that this holds if and only if $\Pi_{\mathcal{N}} R^\star = 0$, that is, $R^\star \in \mathcal{N}^\perp = \mathrm{Range}(\mathcal{A}^*)$.

Any estimator of the form $\langle \widehat{C}, D\rangle_{\mathcal{K}}$. By the first step and (23), $\mathbb{E}[\langle \widehat{C}, D\rangle_{\mathcal{K}} \mid X, W] - V = \langle \Sigma, \mathcal{A}^*(D) - R^\star\rangle_{\mathcal{H}}$. If this vanishes for every $\Sigma \in \mathcal{H}$, taking $\Sigma = \mathcal{A}^*(D) - R^\star$ gives $\mathcal{A}^*(D) = R^\star$, so $R^\star \in \mathrm{Range}(\mathcal{A}^*)$. When the range condition fails, no such D exists. □

C The LSQR Algorithm and Proofs for Section 5

C.1 The LSQR algorithm

This subsection explains Algorithm 1, following Paige and Saunders (1982). The algorithm builds orthonormal vectors $\{v_j\}$ in $\mathcal{K}$, spanning the subspace in which the iterate is sought, and orthonormal vectors $\{u_j\}$ in $\mathcal{H}$, in which its residual lies. In these bases $\mathcal{A}^*$ is bidiagonal, so each step reduces to a small least-squares problem solved by one rotation. We write $V_t y$ for $\sum_j y_j v_j$, and likewise $U_t y$. Table 13 collects the quantities of Algorithm 1 and their meaning; the paragraphs below derive them.

Table 13: Quantities in Algorithm 1

Symbol	Defined by	Meaning
$u_t \in \mathcal{H}$	bidiagonalization	Orthonormal basis of the space containing the residual $R^\star - \mathcal{A}^* D_t$
$v_t \in \mathcal{K}$	bidiagonalization	Orthonormal basis of the search space, in which D_t lies
$a_t,\ b_{t+1} \geq 0$	normalizations	Aligned and new parts of each image: in $\mathcal{A}^* v_t = a_t u_t + b_{t+1} u_{t+1}$, a_t is the component along the latest direction u_t and b_{t+1} the length of the part not aligned with any existing direction
B_t, y_t	$\mathcal{A}^* V_t = U_{t+1} B_t$, $D_t = V_t y_t$	B_t maps the coordinates of D in the v_j to those of $\mathcal{A}^* D$ in the u_j; y_t, the coordinates of D_t, minimizes $\|B_t y - b_1 e_1\|$
Q_t; $c_t,\ s_t$	rotations	Q_t rotates the u_j to separate the fitted directions from the residual; c_t, s_t are the cosine and sine of the angle between the new fitted direction and the residual
w_t	$w_{t+1} = v_{t+1} - (\theta_{t+1}/\rho_t) w_t$	Direction in which D_t is updated: v_t corrected by w_{t-1} so that the images $\mathcal{A}^* w_j$ are mutually orthogonal; each step then fits $R^\star$ along one new direction without changing earlier coefficients
ρ_t, θ_{t+1}	entries of R_t	$\rho_t = \|\mathcal{A}^* w_t\|_{\mathcal{H}}$, the length of the image of the search direction; θ_{t+1}, the component of the next image $\mathcal{A}^* v_{t+1}$ along $\mathcal{A}^* w_t$
$\bar{\rho}_t$	provisional diagonal	The component of $\mathcal{A}^* v_t$ along the previous residual $R^\star - \mathcal{A}^* D_{t-1}$, normalized; the (t,t) entry before the t-th rotation
ζ_t	entries of f_t	The coordinate of $R^\star$ along the fitted direction $\mathcal{A}^* w_t/\rho_t$, the part of the residual removed at step t; hence $D_t = D_{t-1} + (\zeta_t/\rho_t) w_t$
$\bar{\zeta}_{t+1}$	last entry of $Q_t(b_1 e_1)$	The residual left after step t, $\phi_t = \|\bar{\zeta}_{t+1}\|$; step t splits the previous residual $\bar{\zeta}_t$ into $\zeta_t = c_t \bar{\zeta}_t$ removed and $\bar{\zeta}_{t+1} = s_t \bar{\zeta}_t$ left
ϕ_t, ψ_t	(27)	Primal and dual residual norms, from scalars only (Lemma C.1)

Bidiagonalization. Step (1) of the algorithm is the Golub–Kahan recursion

$$b_1 u_1 = R^\star, \qquad a_1 v_1 = \mathcal{A} u_1, \qquad b_{t+1} u_{t+1} = \mathcal{A}^* v_t - a_t u_t, \qquad a_{t+1} v_{t+1} = \mathcal{A} u_{t+1} - b_{t+1} v_t.$$

Each new vector is the image of the latest vector of the other basis, with its component along the latest vector of its own basis removed, and $a_t, b_t \geq 0$ normalize it, $\|u_t\|_{\mathcal{H}} = \|v_t\|_{\mathcal{K}} = 1$. In exact arithmetic $\{u_j\}$ is orthonormal in $\mathcal{H}$ and spans $\mathscr{K}_t(\mathcal{A}^*\mathcal{A}, R^\star)$, and $\{v_j\}$ is orthonormal in $\mathcal{K}$ and spans $\mathscr{K}_t(\mathcal{A}\mathcal{A}^*, \mathcal{A}R^\star)$. The iterate is sought in the span of the v_j, $D = V_t y$, with y chosen to minimize $\|\mathcal{A}^* D - R^\star\|_{\mathcal{H}}$; this is (26). The recursion says that $\mathcal{A}^*$ maps the v_j into the span of the u_j with bidiagonal coefficients. Collecting the first t steps,

$$\mathcal{A}^* V_t = U_{t+1} B_t, \qquad \mathcal{A} U_{t+1} = V_t B_t' + a_{t+1} v_{t+1} e_{t+1}', \qquad B_t = \begin{pmatrix} a_1 & & \\ b_2 & \ddots & \\ & \ddots & a_t \\ & & b_{t+1} \end{pmatrix} \in \mathbb{R}^{(t+1)\times t}, \quad \text{(C.1)}$$

where e_j denotes the j-th standard basis vector of $\mathbb{R}^{t+1}$ and the second identity is used in Lemma C.1. The target lies in the same span, $R^\star = b_1 u_1 = U_{t+1}(b_1 e_1)$, so $\mathcal{A}^* D - R^\star = U_{t+1}(B_t y - b_1 e_1)$, and orthonormality of the u_j gives

$$\|\mathcal{A}^* D - R^\star\|_{\mathcal{H}} = \|B_t y - b_1 e_1\|.$$

The u_j are thus the coordinates in which the residual becomes a vector of length $t+1$, and the minimization over the t-dimensional Krylov space is the $(t+1)\times t$ least-squares problem $y_t = \arg\min_y \|B_t y - b_1 e_1\|$, with $D_t = V_t y_t$.

QR update. Let Q_t be the product of t plane rotations that reduces B_t to upper triangular form,

$$Q_t\begin{bmatrix} B_t & b_1 e_1 \end{bmatrix} = \begin{pmatrix} R_t & f_t \\ 0 & \bar\zeta_{t+1} \end{pmatrix}, \qquad R_t = \begin{pmatrix} \rho_1 & \theta_2 & & \\ & \rho_2 & \ddots & \\ & & \ddots & \theta_t \\ & & & \rho_t \end{pmatrix}, \qquad f_t = \begin{pmatrix} \zeta_1 \\ \vdots \\ \zeta_t \end{pmatrix}.$$

The rotations act on consecutive pairs of rows. After the first $t-1$ of them, rows t and $t+1$ of $[\,B_{t+1}\ \ b_1 e_1\,]$, in columns t, $t+1$ and the right side, are $(\bar\rho_t, 0, \bar\zeta_t)$ and $(b_{t+1}, a_{t+1}, 0)$, with $\bar\rho_1 = a_1$ and $\bar\zeta_1 = b_1$. The t-th rotation acts on these two rows,

$$\begin{pmatrix} c_t & s_t \\ s_t & -c_t \end{pmatrix}\begin{pmatrix} \bar\rho_t & 0 & \bar\zeta_t \\ b_{t+1} & a_{t+1} & 0 \end{pmatrix} = \begin{pmatrix} \rho_t & \theta_{t+1} & \zeta_t \\ 0 & \bar\rho_{t+1} & \bar\zeta_{t+1} \end{pmatrix}, \qquad c_t = \frac{\bar\rho_t}{\rho_t}, \quad s_t = \frac{b_{t+1}}{\rho_t},$$

with $\rho_t = (\bar\rho_t^2 + b_{t+1}^2)^{1/2}$; it removes the subdiagonal entry b_{t+1} and gives the scalar updates of step (2).[21] The entries already in R_t and f_t are not changed by later rotations, so f_t extends f_{t-1} by the single entry ζ_t, and $y_t = R_t^{-1} f_t$. Define w_j by $V_t R_t^{-1} = [\,w_1/\rho_1, \dots, w_t/\rho_t\,]$. Equating column $j+1$ of $V_t = (V_t R_t^{-1})R_t$ gives $w_1 = v_1$ and $w_{j+1} = v_{j+1} - (\theta_{j+1}/\rho_j) w_j$, and then

$$D_t = V_t R_t^{-1} f_t = \sum_{j\le t} \frac{\zeta_j}{\rho_j} w_j = D_{t-1} + \frac{\zeta_t}{\rho_t} w_t,$$

which is the two-term update of step (2). Each step thus costs one rotation and two vector updates, whatever t, and neither basis needs to be stored. Every v_j lies in $\mathrm{Range}(\mathcal{A})$, since $v_1 = \mathcal{A}u_1/a_1$ and v_{j+1} combines $\mathcal{A}u_{j+1}$ and v_j; hence so do w_t and D_t. As $\mathrm{Range}(\mathcal{A}) = \mathrm{Ker}(\mathcal{A}^*)^\perp$, the iterates never acquire a component in $\mathrm{Ker}(\mathcal{A}^*)$, which would raise $\|D\|_{\mathcal{K}}$ without changing $\mathcal{A}^*D$, and they converge to the minimum-norm least-squares solution $D^\star$ (Paige and Saunders, 1982).

Lemma C.1 (Residual scalars)**.** *In exact arithmetic, the residuals* (27) *of the LSQR iterate satisfy*

$$\phi_t = \|R^\star - \mathcal{A}^* D_t\|_{\mathcal{H}} = |\bar\zeta_{t+1}|, \qquad \psi_t = \|\mathcal{A}(R^\star - \mathcal{A}^* D_t)\|_{\mathcal{K}} = |\bar\zeta_{t+1}|\, a_{t+1}\, |c_t|.$$

In particular $\phi_t = b_1 \prod_{j\le t} s_j$, so s_t is the fraction of the remaining residual that step t fails to remove.

Proof. Exact arithmetic enters only through the orthonormality of $\{u_j\}$ and $\{v_j\}$.

The primal residual. Write $q_{t+1} := b_1 e_1 - B_t y_t \in \mathbb{R}^{t+1}$. Since $D_t = V_t y_t$, $R^\star = U_{t+1}(b_1 e_1)$ and

[21] Geometrically, each step is Gram–Schmidt. The columns $\tilde u_1, \dots, \tilde u_t, \bar u_{t+1}$ of $U_{t+1}Q_t'$ are orthonormal, with $\bar u_1 := u_1$; $\tilde u_j = \mathcal{A}^* w_j/\rho_j$ is the normalized image of the j-th search direction, and $\bar u_{t+1}$ is the direction of the residual $R^\star - \mathcal{A}^* D_t = \bar\zeta_{t+1}\bar u_{t+1}$. Because B_t is bidiagonal, the new image $\mathcal{A}^* v_t = \theta_t \tilde u_{t-1} + \bar\rho_t \bar u_t + b_{t+1} u_{t+1}$ has no component along $\tilde u_1, \dots, \tilde u_{t-2}$. Removing the one along $\tilde u_{t-1}$ and normalizing gives $\tilde u_t = (\bar\rho_t \bar u_t + b_{t+1} u_{t+1})/\rho_t$, which is the rotation of step t; the orthogonal direction in that plane, $\bar u_{t+1} = s_t \bar u_t - c_t u_{t+1}$, carries the new residual. Eliminating b_{t+1} and aligning $\tilde u_t$ with the new part of the image are the same operation.

$\mathcal{A}^*V_t = U_{t+1}B_t$ by (C.1),

$$R^\star - \mathcal{A}^*D_t = U_{t+1}\big(b_1e_1 - B_ty_t\big) = U_{t+1}q_{t+1},$$

and orthonormality of the u_j gives $\phi_t = \|q_{t+1}\|$. Since Q_t is orthogonal, $q_{t+1} = Q_t'Q_tq_{t+1}$, and by the definition of the QR update, $Q_t(b_1e_1) = (f_t', \bar\zeta_{t+1})'$ and $Q_tB_t = (R_t', 0)'$. With $y_t = R_t^{-1}f_t$,

$$q_{t+1} = Q_t'\begin{pmatrix} f_t - R_ty_t \\ \bar\zeta_{t+1}\end{pmatrix} = \bar\zeta_{t+1}\,Q_t'e_{t+1},$$

so $\phi_t = |\bar\zeta_{t+1}|$. The rotation of step t gives $\bar\zeta_{t+1} = s_t\bar\zeta_t$, and $\bar\zeta_1 = b_1$, so $\phi_t = b_1\prod_{j\le t}s_j$ and $\phi_t/\phi_{t-1} = s_t$.

The dual residual. By the second identity in (C.1),

$$\mathcal{A}\big(R^\star - \mathcal{A}^*D_t\big) = \mathcal{A}U_{t+1}q_{t+1} = V_tB_t'q_{t+1} + a_{t+1}v_{t+1}\,\big(e_{t+1}'q_{t+1}\big).$$

The first term vanishes: y_t minimizes $\|B_ty - b_1e_1\|$, and the first-order condition is $B_t'(b_1e_1 - B_ty_t) = B_t'q_{t+1} = 0$. For the second, write $Q_t = G_t\cdots G_1$, where G_j is the rotation of step j, acting on rows j and $j+1$ through the block $\left(\begin{smallmatrix} c_j & s_j \\ s_j & -c_j\end{smallmatrix}\right)$. The rotations $G_1,\dots,G_{t-1}$ leave e_{t+1} unchanged, and G_te_{t+1} has entries s_t and $-c_t$ in rows t and $t+1$, so $e_{t+1}'Q_te_{t+1} = -c_t$, and the same holds for Q_t', the scalar being its own transpose. Hence $e_{t+1}'q_{t+1} = -c_t\bar\zeta_{t+1}$, and since $\|v_{t+1}\|_{\mathcal{K}} = 1$, $\psi_t = a_{t+1}\,|c_t|\,|\bar\zeta_{t+1}|$. □

Both residuals are therefore available from scalars the recursion already computes, and the stopping rule (28) costs no operator application. The rotation has a direct reading. When b_{t+1} is small relative to $\bar\rho_t$, $\mathcal{A}^*v_t$ adds little outside the directions already found, s_t is small, and the step removes most of the remaining residual. When $|c_t|$ is small, the remaining residual is nearly orthogonal to $\mathrm{Range}(\mathcal{A}^*)$, which is the case the dual criterion detects.

Implementation in $\mathcal{H}$. Since v_t, w_t and D_t lie in $\mathrm{Range}(\mathcal{A})$, the implementation carries preimages $z_t, p_t, H_t \in \mathcal{H}$ with $v_t = \mathcal{A}(z_t)$, $w_t = \mathcal{A}(p_t)$ and $D_t = \mathcal{A}(H_t)$. Linearity of $\mathcal{A}$ turns the three $\mathcal{K}$-recursions into

$$\tilde z_{t+1} = u_{t+1} - b_{t+1}z_t, \qquad a_{t+1} = \|\mathcal{A}\tilde z_{t+1}\|_{\mathcal{K}}, \qquad z_{t+1} = \tilde z_{t+1}/a_{t+1},$$
$$p_{t+1} = z_{t+1} - \frac{\theta_{t+1}}{\rho_t}\,p_t, \qquad H_t = H_{t-1} + \frac{\zeta_t}{\rho_t}\,p_t,$$

started from $z_1 = p_1 = u_1/a_1$ and $H_0 = 0$. The estimate needs only H_t (Section 5). The solver thus keeps four elements of $\mathcal{H}$ across iterations, u_t, z_t, p_t, H_t, each with $\sum_g N_g(N_g+1)/2$ free entries, and a single element of $\mathcal{K}$ as workspace for $v = \mathcal{A}(\tilde z_{t+1})$, from which a_{t+1} and then $\mathcal{A}^*v_{t+1}$ are computed; no basis of $\mathcal{K}$ is stored and the representer is never assembled. Each iteration applies $\mathcal{A}$ and $\mathcal{A}^*$ once, and each application is a pair of multiplications by M (25).

Floating point. The identities above hold in exact arithmetic; in floating point the recurrences gradually lose the orthogonality of $\{u_j\}$ and $\{v_j\}$, since nothing reorthogonalizes them. We therefore take the recursion of Algorithm 1 as the definition of what is computed, and the stopping rule

monitors the scalars of Lemma C.1 as it produces them. The conditioning results of Section H.7.1 concern floating point.

C.2 Proof of Lemma 5.1

Proof. Throughout, D_t denotes the LSQR iterate from $D_0 = 0$ and $D^\star = (\mathcal{A}^*)^+ R^\star$. Two facts are used repeatedly: $\mathrm{Range}(\mathcal{A}^*) = \mathrm{Ker}(\mathcal{A})^\perp = \mathcal{N}^\perp$, and $\mathcal{A}$ and $\mathcal{A}^*$ have the same nonzero singular values, so that $\|\mathcal{A}x\|_\mathcal{K} \geq s_n^{-1}\|x\|_\mathcal{H}$ for $x \in \mathrm{Ker}(\mathcal{A})^\perp$ and $\|\mathcal{A}^* y\|_\mathcal{H} \geq s_n^{-1}\|y\|_\mathcal{K}$ for $y \in \mathrm{Ker}(\mathcal{A}^*)^\perp$.

(i). The decomposition $r_t = \Pi_\mathcal{N} R^\star + r_t^{\mathrm{red}}$ holds by definition. Both $(I - \Pi_\mathcal{N})R^\star$ and $\mathcal{A}^* D_t$ lie in $\mathcal{N}^\perp = \mathrm{Range}(\mathcal{A}^*)$, hence so does r_t^{red}, which is therefore orthogonal to $\Pi_\mathcal{N} R^\star \in \mathcal{N}$; Pythagoras gives the first identity in (29). Since $\mathcal{A}$ annihilates $\mathcal{N}$, $\mathcal{A}r_t = \mathcal{A}r_t^{\mathrm{red}}$, which is the second. Under Assumption A5(i), $R^\star \in \mathrm{Range}(\mathcal{A}^*) = \mathcal{N}^\perp$, so $\Pi_\mathcal{N} R^\star = 0$ and $\delta_n = 0$.

(ii). Since $r_t^{\mathrm{red}} \in \mathrm{Ker}(\mathcal{A})^\perp$, the singular-value bound gives $\|r_t^{\mathrm{red}}\|_\mathcal{H} \leq s_n \|\mathcal{A}r_t^{\mathrm{red}}\|_\mathcal{K} = s_n \psi_t$. For the other two bounds, let $e_t := D_t - D^\star$. Since $\mathcal{A}^* D^\star = (I - \Pi_\mathcal{N})R^\star$ (Proposition 4.5), $\mathcal{A}^* e_t = -r_t^{\mathrm{red}}$, so $\|e_t\|_\mathcal{T} = \|\mathcal{A}^* e_t\|_\mathcal{H} = \|r_t^{\mathrm{red}}\|_\mathcal{H} \leq s_n \psi_t$. Both D_t and $D^\star$ lie in $\mathrm{Range}(\mathcal{A}) = \mathrm{Ker}(\mathcal{A}^*)^\perp$ (Section C.1 and the proof of Proposition 4.4), hence so does e_t, and the singular-value bound for $\mathcal{A}^*$ gives $\|e_t\|_\mathcal{K} \leq s_n \|\mathcal{A}^* e_t\|_\mathcal{H} = s_n \|e_t\|_\mathcal{T} \leq s_n^2 \psi_t$.

(iii). By (26), ϕ_t is the minimum of $\|R^\star - \mathcal{A}^* D\|_\mathcal{H}$ over $\mathscr{K}_t$, and these subspaces are nested, so ϕ_t is nonincreasing; $\phi_0 = \|R^\star\|_\mathcal{H}$ because $\mathscr{K}_0 = \{0\}$. For the iterates, in exact arithmetic LSQR from $D_0 = 0$ produces the iterates of the conjugate gradient method applied to the normal equations $\mathcal{T}D = \mathcal{A}R^\star$ (Paige and Saunders, 1982). These equations are consistent, since $\mathcal{A}R^\star \in \mathrm{Range}(\mathcal{A}) = \mathrm{Range}(\mathcal{T})$, and the iterates stay in $\mathrm{Range}(\mathcal{T})$, on which $\mathcal{T}$ is positive definite; the norms of such conjugate gradient iterates are nondecreasing (Hestenes et al., 1952); see also Steihaug (1983). In finite dimension the iteration reaches $D^\star$ after finitely many steps, so $\|D_t\|_\mathcal{K} \leq \|D^\star\|_\mathcal{K}$ for every t.

(iv). By (29) and the first bound of (30), $\delta_n^2 = \phi_t^2 - \|r_t^{\mathrm{red}}\|_\mathcal{H}^2 \geq \phi_t^2 - s_n^2 \psi_t^2$, and $\delta_n \geq 0$; the upper bound $\delta_n \leq \phi_t$ is immediate from the same identity.

Finally, $\psi_t = \|\mathcal{A}r_t\|_\mathcal{K} \leq \|\mathcal{A}\|_{\mathrm{op}} \phi_t \leq \phi_t$ by Lemma A.1, which is the comparison stated after the lemma. □

Under Assumption A5(i) the first bound of (30) may be replaced by the identity $\|r_t^{\mathrm{red}}\|_\mathcal{H} = \phi_t$ of part (i), which is what makes the bound in the primal case free of s_n.

D Proofs for Section 6

D.1 Asymptotic normality and feasible implementation

Proposition D.1 (Sufficient conditions for a weighting array)**.** *Suppose Assumptions A1 to A4 hold. Let $\{D_n\}$ be any sequence in $\mathcal{K}_n$ measurable with respect to (X_n, W_n), and set $\widehat{V}(D_n) := \langle \widehat{C}_n, D_n \rangle_{\mathcal{K}_n}$. If*

$$\text{(a) (bias)} \qquad \mu_n \, \|\Sigma_n\|_{\mathcal{H}_n} \, \big\| \mathcal{A}_n^*(D_n) - R_n^\star \big\|_{\mathcal{H}_n} \xrightarrow{p} 0, \tag{D.1}$$

$$\text{(b) (norm)} \qquad \mu_n \, \frac{\sup_g N_g}{n} \, \|M D_n M\|_{\mathcal{K}_n} \xrightarrow{p} 0, \tag{D.2}$$

then $\mu_n\big(\widehat{V}(D_n) - V\big) = o_p(1)$*, and consequently* $\widehat{V}(D_n)/V \to_p 1$*. By Lemma A.4, condition (D.1) is implied by* $\mu_n \|\mathcal{A}_n^*(D_n) - R_n^\star\|_{\mathcal{H}_n} \sqrt{\sup_g N_g/n} \to_p 0$*, and since* $\|M\|_{\mathrm{op}} \le 1$*, condition (D.2) is implied by the same condition with* $\|D_n\|_{\mathcal{K}_n}$ *in place of* $\|MD_nM\|_{\mathcal{K}_n}$*.*

Remark. The proposition is the common core of four results. Taking $D_n = D_n^\star$ under Assumption A5 gives Step 2 of Theorem 6.1: (D.1) holds with left side identically zero, and (D.2) is Assumption A5(ii) via Lemma A.5. Taking $D_n = \widehat{D}_n$ gives Theorem 6.2; the same argument with V replaced by $\Phi_n((I - \Pi_{\mathcal{N}_n})\Sigma_n)$ gives Proposition 6.4; taking $D_n = D_n^{\mathrm{diag}}$ gives Proposition E.5, whose Assumption A10 is (D.1) and (D.2) written out for a block-diagonal array. The proposition uses only Assumptions A1 to A4; the normal limit of the studentized statistic follows in each application from (32), proved in Step 1 of the proof of Theorem 6.1, and Slutsky's theorem.

Proof. Decompose

$$\mu_n\big(\widehat{V}(D_n) - V\big) = \underbrace{\mu_n \big\langle \widehat{C}_n - C_n,\, D_n \big\rangle_{\mathcal{K}_n}}_{\text{(I)}} + \underbrace{\mu_n \Big(\big\langle C_n, D_n \big\rangle_{\mathcal{K}_n} - V \Big)}_{\text{(II)}}. \tag{D.3}$$

Term (II). By Assumption A1, $C_n = \mathcal{A}_n(\Sigma_n)$ (Lemma A.4), so Proposition 4.5 with $\Sigma = \Sigma_n$ and $D = D_n$, and Cauchy–Schwarz, give

$$|\text{(II)}| = \mu_n \big|\big\langle \Sigma_n, \mathcal{A}_n^*(D_n) - R_n^\star \big\rangle_{\mathcal{H}_n}\big| \le \mu_n \, \|\Sigma_n\|_{\mathcal{H}_n} \, \big\|\mathcal{A}_n^*(D_n) - R_n^\star\big\|_{\mathcal{H}_n} = o_p(1)$$

by (D.1).

Term (I). Let $\Psi_n := n^{-2} M D_n M$. Since $\widehat{U} = MU$, $\widehat{C}_n = \widehat{U}\widehat{U}' = MU_nU_n'M$ and $C_n = \mathbb{E}[\widehat{C}_n \mid X_n, W_n] = M\,\mathbb{E}[U_nU_n' \mid X_n, W_n]\,M$. Writing the inner product as a trace and using its cyclicity,

$$\big\langle \widehat{C}_n - C_n, D_n \big\rangle_{\mathcal{K}_n} = \frac{1}{n^2} \operatorname{tr}\Big(M\big(U_nU_n' - \mathbb{E}[U_nU_n' \mid X_n, W_n]\big) M D_n \Big) = U_n'\Psi_n U_n - \mathbb{E}\big[U_n'\Psi_n U_n \mid X_n, W_n\big],$$

a centred quadratic form in the errors, since Ψ_n is (X_n, W_n)-measurable. Its conditional second moment is therefore its conditional variance, $\mathbb{E}[(\text{I})^2 \mid X_n, W_n] = \mu_n^2 \operatorname{Var}(U_n'\Psi_n U_n \mid X_n, W_n)$.

To bound this variance, write $U_n'\Psi_n U_n = \sum_{i,j} (\Psi_n)_{ij} u_i u_j$, so that its variance is a sum of the terms $(\Psi_n)_{ij}(\Psi_n)_{kl} \operatorname{Cov}(u_iu_j, u_ku_l)$. Under Assumption A1, errors in different clusters are independent, so a covariance is nonzero only when the pairs $\{i,j\}$ and $\{k,l\}$ are linked through shared clusters: the one-way support $\mathcal{S}_{1W} = \bigcup_g C_g \times C_g$ is a dependency graph for U_n in the sense of Assumption A7, in which each observation has at most $d_n = \sup_g N_g$ neighbours. Each covariance is bounded by a multiple of the fourth moment $m_4 := \sup_i \mathbb{E}[u_i^4 \mid X_n, W_n]$, which Assumption A2 and Lyapunov's inequality bound almost surely. Counting the nonzero terms in this way is Lemma 8.1, whose proof uses no result of this appendix, and it gives $\operatorname{Var}(U_n'\Psi_n U_n \mid X_n, W_n) \le C_0 m_4 (\sup_g N_g)^2 \|\Psi_n\|_F^2$. Since $\|\Psi_n\|_F = n^{-2}\|MD_nM\|_F = n^{-1}\|MD_nM\|_{\mathcal{K}_n}$,

$$\mathbb{E}\big[(\text{I})^2 \mid X_n, W_n\big] \;\le\; C_0\, m_4 \Big(\mu_n \frac{\sup_g N_g}{n} \, \|MD_nM\|_{\mathcal{K}_n} \Big)^2, \tag{D.4}$$

and the quantity in brackets is exactly the left side of (D.2), so the bound is $o_p(1)$.

It remains to pass from the second moment to (I) itself. By the conditional Chebyshev inequality,

for every $\varepsilon > 0$,

$$\Pr\left(|(\mathrm{I})| > \varepsilon \mid X_n, W_n\right) \;\le\; \varepsilon^{-2}\,\mathbb{E}\big[(\mathrm{I})^2 \mid X_n, W_n\big] \;\xrightarrow{p}\; 0.$$

The conditional probability lies in $[0,1]$, so its expectation, $\Pr(|(\mathrm{I})| > \varepsilon)$, also tends to zero by dominated convergence. Hence $(\mathrm{I}) = o_p(1)$.

Summing, $\mu_n(\widehat{V}(D_n) - V) = (\mathrm{I}) + (\mathrm{II}) = o_p(1)$. Finally, $\widehat{V}(D_n)/V - 1 = \mu_n(\widehat{V}(D_n) - V)/(\mu_n V)$, and $\mu_n V \to_p \nu_a > 0$ by Assumption A3, so $\widehat{V}(D_n)/V \to_p 1$.

The simpler conditions. By Lemma A.4, $\|\Sigma_n\|_{\mathcal{H}_n} \le C_u^{1/(2+\lambda)}(\sup_g N_g/n)^{1/2}$ almost surely, so the left side of (D.1) is at most a constant times $\mu_n\|\mathcal{A}_n^*(D_n) - R_n^\star\|_{\mathcal{H}_n}\sqrt{\sup_g N_g/n}$. Since $\|M\|_{\mathrm{op}} \le 1$, $\|MD_nM\|_{\mathcal{K}_n} \le \|D_n\|_{\mathcal{K}_n}$, so (D.2) holds whenever it holds with $\|D_n\|_{\mathcal{K}_n}$ in place of $\|MD_nM\|_{\mathcal{K}_n}$. □

Proof of Theorem 6.1

Proof. Step 1: Asymptotic normality. Throughout this step we condition on (X, W), so that Γ_n, $\{\Sigma_g\}$ and V are fixed numbers, and work on the event $V > 0$, whose probability tends to one by Assumption A3. By (11), $a'(\widehat{\beta} - \beta) = n^{-1}a'\Gamma_n^{-1}\sum_g \widetilde{X}_g'U_g$, a sum of one term per cluster. Define

$$Z_g := V^{-1/2}\,n^{-1}a'\Gamma_n^{-1}\widetilde{X}_g'U_g, \qquad \text{so that} \qquad V^{-1/2}a'(\widehat{\beta} - \beta) = \sum_{g=1}^{G} Z_g$$

exactly. By Assumption A1, conditional on (X, W) the Z_g are independent with mean zero, and since $\mathbb{E}[\widetilde{X}_g'U_gU_g'\widetilde{X}_g \mid X, W] = \widetilde{X}_g'\Sigma_g\widetilde{X}_g$,

$$\sum_{g=1}^{G}\mathbb{E}[Z_g^2 \mid X, W] = V^{-1}\,a'\Gamma_n^{-1}\left(\frac{1}{n^2}\sum_{g=1}^{G}\widetilde{X}_g'\Sigma_g\widetilde{X}_g\right)\Gamma_n^{-1}a = 1$$

by the definition (15) of V. So $\sum_g Z_g$ is a sum of independent, mean-zero terms with total variance one, and a central limit theorem applies if no single cluster dominates.

That is Lyapunov's condition, $S_n := \sum_g \mathbb{E}[|Z_g|^{2+\xi} \mid X, W] \to_p 0$ for some $\xi > 0$; we take $\xi = \lambda$. By Cauchy–Schwarz, $|Z_g| \le V^{-1/2}n^{-1}\|\Gamma_n^{-1}a\|\,\|\widetilde{X}_g'U_g\|$, so

$$S_n \;\le\; \underbrace{V^{-\frac{2+\xi}{2}}}_{(a)}\;\underbrace{\|\Gamma_n^{-1}a\|^{2+\xi}}_{(b)}\;n^{-(2+\xi)}\;\underbrace{\sum_{g=1}^{G}\mathbb{E}\big[\|\widetilde{X}_g'U_g\|^{2+\xi} \mid X, W\big]}_{(c)}.$$

We bound the three factors in turn. (a) By Assumption A3, $\mu_n V \to_p \nu_a > 0$, so $V^{-1} = O_p(\mu_n)$ and $(a) = O_p\big(\mu_n^{(2+\xi)/2}\big)$. (b) By Assumption A4, $\Gamma_n \to_p \Gamma_0 \succ 0$, so $\|\Gamma_n^{-1}a\| \le \|\Gamma_n^{-1}\|_{\mathrm{op}}\|a\| = O_p(1)$ and $(b) = O_p(1)$. (c) By Lemma A.2(iii) with $\theta = 2 + \xi$, which is allowed because $\xi = \lambda$, $(c) = O_p\big(\sum_g N_g^{2+\xi}\big)$. Multiplying, and using $\sum_g N_g^{2+\xi} \le (\sup_g N_g)^{1+\xi}\sum_g N_g = (\sup_g N_g)^{1+\xi}n$,

$$S_n = O_p\left(\mu_n^{\frac{2+\xi}{2}}\,\frac{\sum_g N_g^{2+\xi}}{n^{2+\xi}}\right) = O_p\left(\mu_n^{\frac{2+\xi}{2}}\left(\frac{\sup_g N_g}{n}\right)^{1+\xi}\right) = O_p\left(\left(\mu_n^{\frac{2+\xi}{2+2\xi}}\,\frac{\sup_g N_g}{n}\right)^{1+\xi}\right),$$

where the last step moves μ_n inside the power $1 + \xi$. With $\xi = \lambda$, the quantity in brackets is the one that Assumption A3 sends to zero, so $S_n = o_p(1)$.

The Lyapunov ratio is random, because it depends on (X, W), and it tends to zero in probability rather than almost surely. Every subsequence therefore has a further subsequence along which it tends to zero almost surely, and along that subsequence the Lyapunov central limit theorem for triangular arrays applies conditionally on (X, W). Since every subsequence thus has a further subsequence along which $\Pr\left(\sum_g Z_g \le x \mid X, W\right) \to \Phi(x)$ almost surely, the same characterization gives $\Pr\left(\sum_g Z_g \le x \mid X, W\right) \to_p \Phi(x)$ along the whole sequence, for every $x \in \mathbb{R}$. Conditional probabilities lie in $[0, 1]$, so taking expectations and using dominated convergence gives $\Pr\left(\sum_g Z_g \le x\right) \to \Phi(x)$, which is (32).

Step 2: Consistency of $\widehat{V}_{\mathrm{Riesz}}$. By Assumption A5(i), $R_n^\star \in \mathrm{Range}(\mathcal{A}_n^*)$, and hence $\mathcal{A}_n^* D_n^\star = R_n^\star$, on an event whose probability tends to one. Events of vanishing probability do not affect convergence in probability, so we work on this event. We verify the two conditions of Proposition D.1 for $D_n = D_n^\star$, which is computed from (X_n, W_n) alone.

Condition (D.1) holds with left side identically zero, since $\mathcal{A}_n^* D_n^\star - R_n^\star = 0$; by Proposition 4.5, this says that the estimator is exactly unbiased given the design. For condition (D.2), $\|M D_n^\star M\|_{\mathcal{K}_n} \le \|M\|_{\mathrm{op}}^2 \|D_n^\star\|_{\mathcal{K}_n} \le \|D_n^\star\|_{\mathcal{K}_n}$, and Lemma A.5 gives $\|D_n^\star\|_{\mathcal{K}_n} = O_p\big(s_n \sqrt{\sup_g N_g / n}\big)$. Hence

$$\mu_n \frac{\sup_g N_g}{n} \|M D_n^\star M\|_{\mathcal{K}_n} \;=\; O_p\left(s_n \mu_n \left(\frac{\sup_g N_g}{n}\right)^{3/2}\right) \;=\; o_p(1)$$

by Assumption A5(ii). Proposition D.1 therefore gives

$$\mu_n(\widehat{V}_{\mathrm{Riesz}} - V) \;\xrightarrow{p}\; 0. \tag{D.5}$$

Since $\widehat{V}_{\mathrm{Riesz}}/V - 1 = \mu_n(\widehat{V}_{\mathrm{Riesz}} - V)/(\mu_n V)$ and $\mu_n V \to_p \nu_a > 0$, this is equivalent to (33).

Step 3: Riesz-t statistic. By Step 2, $\widehat{V}_{\mathrm{Riesz}} > 0$ with probability approaching one, so t_{Riesz} is well defined on that event. Write

$$t_{\mathrm{Riesz}} := \widehat{V}_{\mathrm{Riesz}}^{-1/2} a'(\widehat{\beta} - \beta) = \left(\widehat{V}_{\mathrm{Riesz}}/V\right)^{-1/2} \left[V^{-1/2} a'(\widehat{\beta} - \beta)\right].$$

The first factor tends to one in probability by Step 2 and the continuous mapping theorem, and the second tends to $\mathcal{N}(0, 1)$ in distribution by Step 1, so Slutsky's theorem gives (34). □

Proof of Theorem 6.2

Proof. The argument is that of Step 2 of the proof of Theorem 6.1, with $D_n = \widehat{D}_n = D_{t_n}$ in place of $D_n^\star$; the only new work is the bias condition, which no longer holds exactly. The iterates and the stopping time are computed from (X_n, W_n) alone, so $\widehat{D}_n$ is (X_n, W_n)-measurable, as Proposition D.1 requires. As in Section C.1, LSQR is analysed in exact arithmetic. We work on the event $\{R_n^\star \in \mathrm{Range}(\mathcal{A}_n^*)\} \cap \{\tau_n s_n < 1\}$, whose probability tends to one by Assumption A5(i) and the second part of (35).

Which criterion stops the iteration. On this event $R_n^\star$ and every $\mathcal{A}_n^*(D_t)$ lie in $\mathrm{Range}(\mathcal{A}_n^*)$, so every LSQR residual $r_t = R_n^\star - \mathcal{A}_n^*(D_t)$ does too. On $\mathrm{Range}(\mathcal{A}_n^*) = \mathrm{Ker}(\mathcal{A}_n)^\perp$ the operator $\mathcal{A}_n$ is bounded

below by its smallest nonzero singular value, s_n^{-1}, so

$$\psi_t = \|\mathcal{A}_n(r_t)\|_{\mathcal{K}_n} \;\geq\; s_n^{-1}\phi_t \;>\; \tau_n\phi_t$$

whenever $\phi_t > 0$ (if $\phi_t = 0$, the primal criterion fires at that step). The dual ratio ψ_t/ϕ_t therefore never falls below τ_n, and the dual criterion never fires. The system is solvable on this event, so $\phi_t \to 0$, and the primal criterion stops the iteration at a finite t_n with $\phi_{t_n} \leq \tau_n\|R_n^\star\|_{\mathcal{H}_n}$.

Bias. By the primal stop, and Lemmas A.3 and A.4, which make $\|R_n^\star\|_{\mathcal{H}_n}$ and $\|\Sigma_n\|_{\mathcal{H}_n}$ each $O_p\big(\sqrt{\sup_g N_g/n}\big)$,

$$\begin{aligned}\mu_n\|\Sigma_n\|_{\mathcal{H}_n}\big\|\mathcal{A}_n^*(\widehat{D}_n) - R_n^\star\big\|_{\mathcal{H}_n} &= \mu_n\|\Sigma_n\|_{\mathcal{H}_n}\phi_{t_n} \leq \mu_n\tau_n\|\Sigma_n\|_{\mathcal{H}_n}\|R_n^\star\|_{\mathcal{H}_n}\\ &= O_p\left(\mu_n\,\tau_n\,\frac{\sup_g N_g}{n}\right) = o_p(1)\end{aligned}$$

by the first part of (35). This is (D.1).

Norm. LSQR started at zero never overshoots the minimum-norm solution: $\|\widehat{D}_n\|_{\mathcal{K}_n} \leq \|D_n^\star\|_{\mathcal{K}_n}$ by Lemma 5.1(iii). With $\|M\|_{\mathrm{op}} \leq 1$ and Lemma A.5,

$$\mu_n\,\frac{\sup_g N_g}{n}\,\|M\widehat{D}_nM\|_{\mathcal{K}_n} \;\leq\; \mu_n\,\frac{\sup_g N_g}{n}\,\|D_n^\star\|_{\mathcal{K}_n} \;=\; O_p\left(s_n\,\mu_n\Big(\frac{\sup_g N_g}{n}\Big)^{3/2}\right) \;=\; o_p(1)$$

by Assumption A5(ii). This is (D.2).

Proposition D.1 now gives $\mu_n(\widehat{V}_{\mathrm{Riesz}}^{\mathrm{LSQR}} - V) = o_p(1)$, hence (36) as in Step 2 of the proof of Theorem 6.1, and Step 3 of that proof with $\widehat{V}_{\mathrm{Riesz}}^{\mathrm{LSQR}}$ in place of $\widehat{V}_{\mathrm{Riesz}}$ gives (37). □

D.2 The range condition and approximate solvability

Proof of Proposition 6.3

Proof. The range condition as orthogonality. In finite dimensions $\mathrm{Range}(\mathcal{A}_n^*) = \mathrm{Ker}(\mathcal{A}_n)^\perp$, as in the proof of Proposition 4.4. Hence $R_n^\star \in \mathrm{Range}(\mathcal{A}_n^*)$ if and only if $\langle R_n^\star, \Delta\rangle_{\mathcal{H}_n} = 0$ for every $\Delta \in \mathrm{Ker}(\mathcal{A}_n)$.

The inner product. With $z_g = \widetilde{X}_g\Gamma_n^{-1}a$, $R_{n,g}^\star = \widetilde{X}_gQ_n\widetilde{X}_g' = z_gz_g'$. Since the blocks are symmetric, the definition of $\langle\cdot,\cdot\rangle_{\mathcal{H}_n}$ and the cyclicity of the trace give, for every $\Delta \in \mathcal{H}_n$,

$$\langle R_n^\star, \Delta\rangle_{\mathcal{H}_n} = \frac{1}{n^2}\sum_{g=1}^{G}\mathrm{tr}\left(z_gz_g'\Delta_g\right) = \frac{1}{n^2}\sum_{g=1}^{G}z_g'\Delta_gz_g.$$

The factor n^{-2} does not affect whether the sum vanishes, so the condition of the first step is exactly (38). Chen (2026, Corollary 3.4) gives the general case. □

Proof of Proposition 6.4

Proof. Write $V_{\mathcal{N}} := \Phi_n\big((I - \Pi_{\mathcal{N}_n})\Sigma_n\big)$ for the target of part (i). As in Section C.1, LSQR is analysed in exact arithmetic.

Part (i): the decomposition. For any (X_n, W_n)-measurable D_n, decompose $\mu_n(\langle\widehat{C}_n, D_n\rangle_{\mathcal{K}_n} - V_{\mathcal{N}})$ as in (D.3), with V replaced by $V_{\mathcal{N}}$. Term (I) does not involve the target, so by the proof of

Proposition D.1 it is $o_p(1)$ whenever (D.2) holds. For term (II), Proposition 4.5 and $\Phi_n(\Pi_{\mathcal{N}_n}\Sigma_n) = \langle \Sigma_n, \Pi_{\mathcal{N}_n} R_n^\star \rangle_{\mathcal{H}_n}$ give

$$\langle C_n, D_n \rangle_{\mathcal{K}_n} - V_{\mathcal{N}} = \big\langle \Sigma_n,\ \mathcal{A}_n^*(D_n) - (I - \Pi_{\mathcal{N}_n}) R_n^\star \big\rangle_{\mathcal{H}_n}. \tag{D.6}$$

It therefore suffices to verify (D.2) and $\mu_n \|\Sigma_n\|_{\mathcal{H}_n} \|\mathcal{A}_n^*(D_n) - (I - \Pi_{\mathcal{N}_n}) R_n^\star\|_{\mathcal{H}_n} \to_p 0$ for $D_n^\star$ and $\widehat{D}_n$.

Part (i): the exact representer. By (24), $\mathcal{A}_n^* D_n^\star = (I - \Pi_{\mathcal{N}_n}) R_n^\star$, so (D.6) vanishes. The bound $\|D_n^\star\|_{\mathcal{K}_n} \le s_n \|R_n^\star\|_{\mathcal{H}_n}$ of Lemma A.5 holds whatever the defect, so with $\|M D_n^\star M\|_{\mathcal{K}_n} \le \|D_n^\star\|_{\mathcal{K}_n}$ and Lemma A.3, the left side of (D.2) is $O_p\big(s_n \mu_n (\sup_g N_g/n)^{3/2}\big) = o_p(1)$ by Assumption A5(ii).

Part (i): the LSQR iterate. The iteration terminates whatever the defect. If $\delta_n > 0$, LSQR converges to a least-squares solution, so $\psi_t \to 0$ while $\phi_t \to \delta_n > 0$ by Lemma 5.1(i), and the dual ratio ψ_t/ϕ_t falls below τ_n; if $\delta_n = 0$, $\phi_t \to 0$ and the primal criterion fires. Either way $\widehat{D}_n$ is (X_n, W_n)-measurable. By Lemma 5.1(i), $\mathcal{A}_n^*(\widehat{D}_n) - (I - \Pi_{\mathcal{N}_n}) R_n^\star = -r_{t_n}^{\mathrm{red}}$. If the primal criterion stopped the iteration, (29) gives $\|r_{t_n}^{\mathrm{red}}\|_{\mathcal{H}_n} \le \phi_{t_n} \le \tau_n \|R_n^\star\|_{\mathcal{H}_n}$; if the dual criterion did, the first bound of (30) and Lemma 5.1(iii) give $\|r_{t_n}^{\mathrm{red}}\|_{\mathcal{H}_n} \le s_n \psi_{t_n} \le s_n \tau_n \phi_{t_n} \le s_n \tau_n \|R_n^\star\|_{\mathcal{H}_n}$. Since $\|\mathcal{A}_n\|_{\mathrm{op}} \le 1$ (Lemma A.1), $\sigma_{\min}^+(\mathcal{A}_n) \le 1$ and $s_n \ge 1$, so in both cases $\|r_{t_n}^{\mathrm{red}}\|_{\mathcal{H}_n} \le s_n \tau_n \|R_n^\star\|_{\mathcal{H}_n}$, and Lemmas A.3 and A.4 give

$$\mu_n \|\Sigma_n\|_{\mathcal{H}_n} \|r_{t_n}^{\mathrm{red}}\|_{\mathcal{H}_n} = O_p\Big(\mu_n \tau_n s_n \frac{\sup_g N_g}{n}\Big) = o_p(1)$$

by (39). For (D.2), Lemma 5.1(iii) gives $\|\widehat{D}_n\|_{\mathcal{K}_n} \le \|D_n^\star\|_{\mathcal{K}_n}$ whatever the defect, so the bound for $D_n^\star$ applies.

Part (ii). By Proposition 4.5, $V - V_{\mathcal{N}} = \Phi_n(\Pi_{\mathcal{N}_n}\Sigma_n) = \langle \Sigma_n, \Pi_{\mathcal{N}_n} R_n^\star \rangle_{\mathcal{H}_n}$, and Cauchy–Schwarz with Lemma A.4 gives

$$\big|\Phi_n\big(\Pi_{\mathcal{N}_n}\Sigma_n\big)\big| \le \|\Sigma_n\|_{\mathcal{H}_n} \delta_n = O_p\Big(\delta_n \sqrt{\sup_g N_g/n}\Big) = o_p(\mu_n^{-1})$$

by (40). With part (i), $\mu_n(\widehat{V} - V) = o_p(1)$, and since $\mu_n V \to_p \nu_a > 0$ by Assumption A3, $\widehat{V}/V \to_p 1$, which is (33) and (36). Step 1 of the proof of Theorem 6.1 uses only Assumptions A1 to A4, so (32) holds; as in Step 3 there, $\widehat{V} > 0$ with probability approaching one, and Slutsky's theorem gives (34) and (37). □

E Supplement and Proofs for Section 7

This section follows the order of Section 7. It proves the results stated there and adds the supporting lemmas and propositions they rely on, each followed by its proof.

E.1 The restricted and full systems

Proof of Proposition 7.1. Step 1: the operator identity. Read $H \in \mathcal{H}_n$ as $\mathrm{diag}(H_1, \dots, H_G)$. Since $\widehat{U} = MU$, $\mathcal{A}_n(H)$ is the block decomposition of MHM' and $[\mathcal{A}_n^*(D)]_g = [M'DM]_{gg}$; (17) and (18) write these with $M' = M$. Composing and using $M'M = M$, which holds because M is symmetric

and idempotent,

$$\begin{aligned}\big[\mathcal{A}_n^*\mathcal{A}_n(H)\big]_g &= \big[M'\big(MHM'\big)M\big]_{gg} = \big[(M'M)\,H\,(M'M)\big]_{gg} = \big[MHM\big]_{gg} = \sum_{k=1}^{G} M_{gk}H_kM_{kg}\\ &= \big[P_{\mathrm{diag}}(MHM)\big]_{gg} = \big[\mathcal{A}_n^{\mathrm{diag}}(H)\big]_g,\end{aligned}$$

which is (43). Self-adjointness and positive semidefiniteness are then automatic: $(\mathcal{A}_n^*\mathcal{A}_n)^* = \mathcal{A}_n^*\mathcal{A}_n$ and $\langle\Sigma, \mathcal{A}_n^{\mathrm{diag}}\Sigma\rangle_{\mathcal{H}_n} = \|\mathcal{A}_n\Sigma\|^2_{\mathcal{K}_n} \geq 0$.

Step 2: kernels and ranges. If $\mathcal{A}_n^{\mathrm{diag}}\Sigma = 0$ then $0 = \langle\Sigma, \mathcal{A}_n^{\mathrm{diag}}\Sigma\rangle_{\mathcal{H}_n} = \|\mathcal{A}_n\Sigma\|^2_{\mathcal{K}_n}$ by Step 1, so $\Sigma \in \mathrm{Ker}(\mathcal{A}_n)$; the reverse inclusion is immediate. Hence $\mathrm{Ker}(\mathcal{A}_n^{\mathrm{diag}}) = \mathrm{Ker}(\mathcal{A}_n)$, and since $\mathcal{A}_n^{\mathrm{diag}}$ is self-adjoint,

$$\mathrm{Range}\left((\mathcal{A}_n^{\mathrm{diag}})^*\right) = \mathrm{Range}\left(\mathcal{A}_n^{\mathrm{diag}}\right) = \mathrm{Ker}(\mathcal{A}_n^{\mathrm{diag}})^\perp = \mathrm{Ker}(\mathcal{A}_n)^\perp = \mathrm{Range}(\mathcal{A}_n^*).$$

Solvability of (41) and of (19) are both the statement $R_n^\star \in \mathrm{Ker}(\mathcal{A}_n)^\perp$.

Step 3: the variance identity. If $\mathcal{A}_n^*D_n^\star = R_n^\star$ then $\langle\mathcal{A}_n\Sigma_n, D_n^\star\rangle_{\mathcal{K}_n} = \langle\Sigma_n, R_n^\star\rangle_{\mathcal{H}_n} = \Phi(\Sigma_n) = V$ by Proposition 4.1. Likewise, if $(\mathcal{A}_n^{\mathrm{diag}})^*D_n^{\mathrm{diag}} = R_n^\star$, then

$$\langle\mathcal{A}_n^{\mathrm{diag}}\Sigma_n, D_n^{\mathrm{diag}}\rangle_{\mathcal{K}_n^{\mathrm{diag}}} = \langle\Sigma_n, (\mathcal{A}_n^{\mathrm{diag}})^*D_n^{\mathrm{diag}}\rangle_{\mathcal{H}_n} = \langle\Sigma_n, R_n^\star\rangle_{\mathcal{H}_n} = V.$$

Chen (2026, Theorem 5.2) gives the general case. □

Proof of Corollary 7.2. By Proposition 7.1, $\mathcal{A}_n^{\mathrm{diag}} = \mathcal{A}_n^*\mathcal{A}_n$, so $H_n^\star = (\mathcal{A}_n^*\mathcal{A}_n)^+R_n^\star$ and

$$D_n^\star = (\mathcal{A}_n^*)^+R_n^\star = \mathcal{A}_n(\mathcal{A}_n^*\mathcal{A}_n)^+R_n^\star = MH_n^\star M,$$

by the identity $X^+ = X^*(XX^*)^+$ with $X = \mathcal{A}_n^*$ and because $\mathcal{A}_n(H)$ is the block decomposition of MHM; neither identity requires $R_n^\star \in \mathrm{Range}(\mathcal{A}_n^*)$. Since $\widehat{U} = M\widehat{U}$ and $H_n^\star$ is block diagonal,

$$\begin{aligned}\langle\widehat{C}_n, D_n^\star\rangle_{\mathcal{K}_n} &= \frac{1}{n^2}\widehat{U}'MH_n^\star M\widehat{U} = \frac{1}{n^2}\widehat{U}'H_n^\star\widehat{U}\\ &= \frac{1}{n^2}\sum_g \widehat{U}_g'H_{n,g}^\star\widehat{U}_g = \langle P_{\mathrm{diag}}\widehat{C}_n, H_n^\star\rangle_{\mathcal{K}_n^{\mathrm{diag}}}.\end{aligned}$$

If H_n is any other solution of (41), then $K := H_n - H_n^\star \in \mathrm{Ker}(\mathcal{A}_n^{\mathrm{diag}}) = \mathrm{Ker}(\mathcal{A}_n)$, so $MKM = 0$ and $\widehat{U}'K\widehat{U} = U'MKMU = 0$; hence H_n gives the same value. □

Proposition E.1 (Conditioning of the restricted system)**.** *Let $\sigma_1 \geq \cdots \geq \sigma_m > 0$ denote the nonzero singular values of $\mathcal{A}_n$ and $\lambda_1 \geq \cdots \geq \lambda_m > 0$ those of $\mathcal{A}_n^{\mathrm{diag}}$. Then $\lambda_j = \sigma_j^2$ for every j, and on the identified subspace $\mathrm{Ker}(\mathcal{A}_n)^\perp$,*

$$\kappa\big(\mathcal{A}_n^{\mathrm{diag}}\big) = \kappa(\mathcal{A}_n)^2, \qquad \kappa(\cdot) := \sigma_{\max}/\sigma^+_{\min},$$

with equality to $\kappa(\mathcal{A}_n)$ if and only if $\kappa(\mathcal{A}_n) = 1$. In particular, $\|(\mathcal{A}_n^{\mathrm{diag}})^+\|_{\mathrm{op}} = s_n^2$ while $\|(\mathcal{A}_n^*)^+\|_{\mathrm{op}} = s_n$.

Applied to $\mathcal{A}_n^{\mathrm{diag}}$, LSQR therefore has iteration count and attainable accuracy governed by $\kappa(\mathcal{A}_n)^2$ rather than $\kappa(\mathcal{A}_n)$, and the bracket of Lemma 5.1(iv) involves s_n^2 in place of s_n; the estimator is the same (Section H.7.2 gives the benchmarks).

Proof. By Proposition 7.1, $\mathcal{A}_n^{\mathrm{diag}} = \mathcal{A}_n^*\mathcal{A}_n$ is self-adjoint and positive semidefinite, so its singular values coincide with its eigenvalues. If $\mathcal{A}_n = \sum_j \sigma_j \langle \cdot, e_j \rangle f_j$ is a singular value decomposition with $\{e_j\} \subset \mathcal{H}_n$ and $\{f_j\} \subset \mathcal{K}_n$ orthonormal, then $\mathcal{A}_n^*\mathcal{A}_n = \sum_j \sigma_j^2 \langle \cdot, e_j \rangle e_j$, so $\lambda_j = \sigma_j^2$. Restricted to $\mathrm{Ker}(\mathcal{A}_n)^\perp = \mathrm{span}\{e_j\}$, $\kappa(\mathcal{A}_n^{\mathrm{diag}}) = \sigma_1^2/\sigma_m^2 = \kappa(\mathcal{A}_n)^2$, which exceeds $\kappa(\mathcal{A}_n) \geq 1$ unless $\kappa(\mathcal{A}_n) = 1$. Finally $\|(\mathcal{A}_n^{\mathrm{diag}})^+\|_{\mathrm{op}} = \lambda_m^{-1} = \sigma_m^{-2} = s_n^2$, while $\|(\mathcal{A}_n^*)^+\|_{\mathrm{op}} = \sigma_m^{-1} = s_n$. Chen (2026, Theorem 5.6) gives the general case. □

E.2 Classical estimators as approximate inversions

A cluster-robust variance *matrix* performs a further compression, to the $k \times k$ aggregate $\sum_g \widetilde{X}_g' \cdot \widetilde{X}_g$, but that compression is lossless for V, which by (15) depends on $\{\Sigma_g\}$ only through it. All loss relative to the Riesz estimator therefore enters through the approximate inversion.

Scalar clusters. With $N_g = 1$ the restricted system is $(M \odot M)\, d = r$ with $r_i = \tilde{x}_i' Q_n \tilde{x}_i$. Its exact solution is the estimator of Cattaneo et al. (2018), up to the replacement of $M = I_n - \Pi_{[X\ W]}$ by M_W in the Hadamard system. That replacement is first-order negligible but not exact: the residuals come from the full regression, so their second moments are governed by $M \odot M$, and CJN is consistent but not exactly unbiased, the difference being $O_p(s_n^4 n^{-3/2})$ (Lemma E.2). CJN is thus, to first order, a scalar-cluster restricted Riesz solution, and under a diagonal covariance model the exact solution is the *exact* inversion; by Proposition 7.1 no procedure reading residual second moments can improve on it in identification terms.[22]

HC2 and HC3 are approximations to that inversion rather than alternatives to it. Under the conditions of Lemma E.3, $\max_i h_{ii} = O_p(n^{-1})$, the inverse-Hadamard operator is asymptotically diagonally dominant and the exact weight expands as $d_i = r_i(1-h_{ii})^{-2} + O_p(n^{-1}\|r\|_\infty)$ (Lemma E.3). HC3 matches the leading term exactly, HC2 differs from it by $O_p(n^{-1})$, and both satisfy (E.1) below: their weights differ from d by $O(\bar{h}\|r\|_\infty)$ entrywise (Lemma E.3), and $\mathcal{A}_n^{\mathrm{diag}}$ is a contraction. The regime is the one Section 2.1 identifies as favourable; when $p/n \to \alpha > 0$ it fails and the remainder is first order.

Lemma E.2 (Replacing M by M_W)**.** *Let $N_g = 1$ for every g, let $k = \dim(\beta)$ be fixed, and suppose Assumptions A1, A2 and A4 hold with $\max_i \|\tilde{x}_i\|^2 = O_p(1)$. Write $\Pi := \Pi_{\widetilde{X}} = \widetilde{X}(\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'$ for the projection onto $\mathrm{col}(\widetilde{X})$.*

(i) $M_W - M = \Pi$*, an orthogonal projection of rank k with $\Pi M = M\Pi = 0$; in particular $\|M_W - M\|_{\mathrm{op}} = 1$.*

(ii) $\big\|(M_W \odot M_W) - (M \odot M)\big\|_F^2 \leq 9\, k \max_i \Pi_{ii} = O_p(n^{-1})$.

[22] The qualifier "residual second moments" is not idle. Leave-out moments $\mathbb{E}[y_i \hat{u}_i \mid X] = (1 - h_{ii})\sigma_i^2$, of the kind underlying Kline et al. (2020), Jochmans (2022) and Anatolyev and Ng (2026), are β-free in expectation but are not quadratic forms in the residual vector, and they identify every σ_i^2 with $h_{ii} < 1$ even when $M \odot M$ is singular (Chen, 2026, Corollary 6.3), at the price of sampling variability that depends on the regression mean.

(iii) Suppose $M \odot M$ is nonsingular with probability approaching one, so that $\|(M \odot M)^{-1}\|_{\mathrm{op}} = s_n^2$ by Proposition E.1, and that $s_n^2/\sqrt{n} \to_p 0$. Let d and d^W solve $(M \odot M)\, d = r$ and $(M_W \odot M_W)\, d^W = r$, with $r_i = \tilde{x}_i' Q_n \tilde{x}_i$, and set $\widehat{V}_{\mathrm{P}} := n^{-2} \sum_i d_i \hat{u}_i^2$ and $\widehat{V}_{\mathrm{CJN}} := n^{-2} \sum_i d_i^W \hat{u}_i^2$. Then d^W is well defined with probability approaching one, and

$$\left|\widehat{V}_{\mathrm{CJN}} - \widehat{V}_{\mathrm{P}}\right| \;=\; O_p\big(s_n^4\, n^{-3/2}\big).$$

Hence $\mu_n(\widehat{V}_{\mathrm{CJN}} - \widehat{V}_{\mathrm{P}}) = o_p(1)$ whenever $\mu_n s_n^4 n^{-3/2} \to_p 0$; with $\mu_n \asymp n$, as in the heteroskedastic case, this is $s_n^4/\sqrt{n} \to_p 0$.

Under a diagonal covariance model $\mathbb{E}[\widehat{V}_{\mathrm{P}} \mid X, W] = V$ exactly, whereas $\mathbb{E}[\widehat{V}_{\mathrm{CJN}} \mid X, W] - V = n^{-2} \sum_i (d_i^W - d_i)\mathbb{E}[\hat{u}_i^2 \mid X, W]$ is in general nonzero; part (iii) shows that this bias, like the sampling difference, is of smaller order than V whenever $\mu_n s_n^4 n^{-3/2} \to_p 0$.

Proof. Part (i). By the Frisch–Waugh–Lovell decomposition of Section 3, $\Pi_{[X\ W]} = \Pi_W + \Pi$ with $\widetilde{X} = M_W X$; hence $M = M_W - \Pi$. Since $\mathrm{col}(\widetilde{X}) \subseteq \mathrm{col}([X\ W])$, which M annihilates, $M\Pi = 0$, and $\Pi M = (M\Pi)' = 0$. By Assumption A4, $\widetilde{X}$ has full column rank k with probability approaching one, so Π has rank k, and as a nonzero orthogonal projection $\|\Pi\|_{\mathrm{op}} = 1$.

Part (ii). From $M_W = M + \Pi$, $M_W \odot M_W - M \odot M = 2\, M \odot \Pi + \Pi \odot \Pi$. Three facts suffice: $\Pi_{ij}^2 \le \Pi_{ii}\Pi_{jj}$ (Cauchy–Schwarz for the positive semidefinite Π), $\sum_j M_{ij}^2 = M_{ii} \le 1$ and $\sum_j \Pi_{ij}^2 = \Pi_{ii}$ (idempotency), and $\sum_i \Pi_{ii} = k$. Hence

$$\|M \odot \Pi\|_F^2 = \sum_{i,j} M_{ij}^2 \Pi_{ij}^2 \le \max_j \Pi_{jj} \sum_i \Pi_{ii} \sum_j M_{ij}^2 \le k \max_i \Pi_{ii},$$
$$\|\Pi \odot \Pi\|_F^2 = \sum_{i,j} \Pi_{ij}^4 \le \max_{i,j} \Pi_{ij}^2 \sum_{i,j} \Pi_{ij}^2 \le k \max_i \Pi_{ii},$$

the last step using $\max_{i,j} \Pi_{ij}^2 \le (\max_i \Pi_{ii})^2 \le \max_i \Pi_{ii} \le 1$. By the triangle inequality,

$$\left\|M_W \odot M_W - M \odot M\right\|_F \le 2\sqrt{k \max_i \Pi_{ii}} + \sqrt{k \max_i \Pi_{ii}} = 3\sqrt{k \max_i \Pi_{ii}}.$$

Finally $\Pi_{ii} = n^{-1} \tilde{x}_i' \Gamma_n^{-1} \tilde{x}_i \le n^{-1} \|\Gamma_n^{-1}\|_{\mathrm{op}} \max_i \|\tilde{x}_i\|^2 = O_p(n^{-1})$ uniformly in i by Assumption A4, and k is fixed.

Part (iii). With $N_g = 1$, the matrix B of (42) is $M \odot M$, and $d = B^{-1} r$. Write $B_W := M_W \odot M_W$ and $\Delta := B_W - B$. By part (ii), $\|\Delta\|_{\mathrm{op}} \le \|\Delta\|_F = O_p(n^{-1/2})$, so $\|B^{-1}\|_{\mathrm{op}} \|\Delta\|_{\mathrm{op}} = O_p(s_n^2/\sqrt{n}) = o_p(1)$. On the event $\{\|B^{-1}\|_{\mathrm{op}} \|\Delta\|_{\mathrm{op}} \le 1/2\}$, whose probability tends to one, the Neumann series shows that $B_W = B(I + B^{-1}\Delta)$ is nonsingular with $\|B_W^{-1}\|_{\mathrm{op}} \le 2 s_n^2$, so d^W is well defined, and

$$d - d^W = B^{-1} r - B_W^{-1} r = B_W^{-1}(B_W - B) B^{-1} r = B_W^{-1} \Delta\, d.$$

Since $|r_i| \le \|Q_n\|_{\mathrm{op}} \|\tilde{x}_i\|^2 = O_p(1)$ uniformly, $\|r\|_2 = O_p(\sqrt{n})$ and $\|d\|_2 \le s_n^2 \|r\|_2 = O_p(s_n^2 \sqrt{n})$. Therefore

$$\|d - d^W\|_2 \le 2 s_n^2 \cdot O_p(n^{-1/2}) \cdot O_p(s_n^2 \sqrt{n}) = O_p(s_n^4).$$

For the residuals, $\hat{u}_i = \sum_j M_{ij} u_j$, and the u_j are independent by Assumption A1; Assumption A2, $\sum_j M_{ij}^2 \le 1$ and Rosenthal's inequality give $\mathbb{E}[\hat{u}_i^4 \mid X, W] \le C$ uniformly in i, so $\mathbb{E}[\sum_i \hat{u}_i^4 \mid X, W] \le$

Cn and, by Markov's inequality, $\big(\sum_i \hat{u}_i^4\big)^{1/2} = O_p(\sqrt{n})$. By Cauchy–Schwarz,

$$\big|\widehat{V}_{\mathrm{CJN}} - \widehat{V}_{\mathrm{P}}\big| = n^{-2}\Big|\sum_i (d_i^W - d_i)\hat{u}_i^2\Big| \le n^{-2}\|d - d^W\|_2\Big(\sum_i \hat{u}_i^4\Big)^{1/2} = O_p\big(s_n^4\, n^{-3/2}\big).$$

The same bound with $\mathbb{E}[\hat{u}_i^2 \mid X, W] \le C$ in place of $\hat{u}_i^2$ controls the conditional bias. Finally, with scalar clusters $\mathcal{A}_n^{\mathrm{diag}}$ acts on diagonal arrays as B, so $\widehat{V}_{\mathrm{P}}$ is $\widehat{V}_{\mathrm{Riesz}}$ by Corollary 7.2; since B is non-singular, $\mathrm{Ker}(\mathcal{A}_n) = \{0\}$, and Theorem 4.6 gives exact unbiasedness under a diagonal covariance model. □

Lemma E.3 (Diagonal expansion of the inverse-Hadamard weights)**.** *Let $N_g = 1$ for all g, let d solve $(M \odot M)d = r$, and suppose $\bar{h} := \max_i h_{ii} \to 0$. Write $M \odot M = \Lambda + F$ with $\Lambda = \mathrm{diag}\{(1-h_{ii})^2\}$ and $F_{ik} = h_{ik}^2$ for $i \neq k$, $F_{ii} = 0$. Then*

$$d_i = \frac{r_i}{(1-h_{ii})^2} + O\big(\bar{h}\,\|r\|_\infty\big).$$

If $[X\ W]$ has fixed dimension, bounded rows, and $\lambda_{\min}(n^{-1}[X\ W]'[X\ W])$ bounded away from zero, then $\bar{h} = O_p(n^{-1})$; HC3 reproduces the leading term exactly, and HC2 differs from it by $h_{ii}r_i/(1-h_{ii})^2 = O_p(n^{-1}\|r\|_\infty)$.

Proof. The entries of $M \odot M$ are M_{ik}^2, and since M is symmetric and idempotent, $\sum_k M_{ik}^2 = (M^2)_{ii} = 1 - h_{ii}$; removing the diagonal term $(1-h_{ii})^2$ gives $\sum_{k\neq i} F_{ik} = h_{ii}(1-h_{ii})$, which is (4). Hence $q := \|\Lambda^{-1}F\|_\infty = \max_i h_{ii}/(1-h_{ii}) = \bar{h}/(1-\bar{h})$, which is below one once $\bar{h} < 1/2$, as $\bar{h} \to 0$ ensures for large n. The Neumann series

$$(\Lambda + F)^{-1} = \big(I + \Lambda^{-1}F\big)^{-1}\Lambda^{-1} = \sum_{m\ge 0}(-\Lambda^{-1}F)^m\,\Lambda^{-1}$$

converges in the induced ℓ_∞ norm. Its first term gives $\Lambda^{-1}r$, with entries $r_i/(1-h_{ii})^2$, and the remaining terms are bounded by $q(1-q)^{-1}\|\Lambda^{-1}\|_\infty\|r\|_\infty = O(\bar{h}\,\|r\|_\infty)$, since $q = O(\bar{h})$ and $\|\Lambda^{-1}\|_\infty = (1-\bar{h})^{-2} \to 1$. Finally, with z_i the ith row of $Z = [X\ W]$, $h_{ii} = z_i'(Z'Z)^{-1}z_i \le \|z_i\|^2/\lambda_{\min}(Z'Z) = O_p(n^{-1})$ uniformly in i under the stated conditions, and $|h_{ii}r_i|/(1-h_{ii})^2 \le \bar{h}\|r\|_\infty/(1-\bar{h})^2$. □

General clusters. With general N_g the classical corrections use the block weights

$$D_{gg} = M_{gg}^{-\delta}(\widetilde{X}_g Q_n \widetilde{X}_g')M_{gg}^{-\delta},$$

with $\delta = 0$ for CR0 and CR1, $1/2$ for CR2 and 1 for CR3; CR3 inverts the decoupled system B_0 of Section 7, and the others approximate it (Lemma E.4).

Lemma E.4 (Leverage and cross-cluster mass)**.** *Suppose every M_{gg} is nonsingular and let $\varepsilon_n := \max_g \|M_{gg}^{-1} - I\|_{\mathrm{op}}$.*

(i) For every g, $\sum_{h\neq g}(M_{gg}^{-1}M_{gh})(M_{gg}^{-1}M_{gh})' = M_{gg}^{-1} - I \succeq 0$, and hence $\varepsilon_n \le \varrho_n^{\mathrm{sp}}$, with ϱ_n^{sp} as in (Proj-Sparsity).

(ii) For $\delta \in [0,1]$ let $D_{gg}^{(\delta)} := M_{gg}^{-\delta}R_{n,g}^\star M_{gg}^{-\delta}$. Then $\|D^{(\delta)} - D^{(1)}\|_{\mathcal{H}_n} \le 2(2\varepsilon_n + \varepsilon_n^2)\|R_n^\star\|_{\mathcal{H}_n}$.

(iii) $p := \operatorname{rank}([X\ W]) \le n\,\varepsilon_n$.

Proof. *(i).* M is symmetric and idempotent, so the (g,g) block of $M = M^2$ reads

$$M_{gg} = \sum_h M_{gh}M_{hg} = M_{gg}^2 + \sum_{h\neq g} M_{gh}M_{gh}'.$$

Pre- and post-multiplying by M_{gg}^{-1} gives the identity. Each summand is positive semidefinite, which gives $M_{gg}^{-1} - I \succeq 0$, and the triangle inequality with $\|AA'\|_{\mathrm{op}} = \|A\|_{\mathrm{op}}^2$ gives

$$\|M_{gg}^{-1} - I\|_{\mathrm{op}} \le \sum_{h\neq g} \|M_{gg}^{-1}M_{gh}\|_{\mathrm{op}}^2 \le \varrho_n^{\mathrm{sp}}.$$

(ii). The eigenvalues m of M_{gg} lie in $(0,1]$, and $m^{-\delta}-1 \le m^{-1}-1$ for $\delta \in [0,1]$, so $\|M_{gg}^{-\delta} - I\|_{\mathrm{op}} \le \varepsilon_n$. For any R and any A with $\|A-I\|_{\mathrm{op}} \le \varepsilon$, so that $\|A\|_{\mathrm{op}} \le 1+\varepsilon$, $\|ARA - R\|_F \le \|(A-I)RA\|_F + \|R(A-I)\|_F \le (2\varepsilon+\varepsilon^2)\|R\|_F$. Applying this with $A = M_{gg}^{-\delta}$ and $A = M_{gg}^{-1}$, using the triangle inequality and summing the squares over g gives the bound.

(iii). $p = \operatorname{tr}(I-M) = \sum_g \operatorname{tr}(I - M_{gg}) \le \sum_g N_g \|I - M_{gg}\|_{\mathrm{op}}$, and $1 - m \le m^{-1} - 1$ for $m \in (0,1]$, so $\|I - M_{gg}\|_{\mathrm{op}} \le \varepsilon_n$. □

First-order equivalence and leakage. A restricted weighting scheme that does not solve (41) exactly is first-order equivalent to the Riesz estimator when it reproduces $R_n^\star$ at the scale governing studentization.

Assumption A10 (Partial approximation)**.** The restricted weights D_n^{diag} satisfy

$$\big\|(\mathcal{A}_n^{\mathrm{diag}})^* D_n^{\mathrm{diag}} - R_n^\star\big\|_{\mathcal{H}_n} = o_p\Big(\mu_n^{-1}\sqrt{\tfrac{n}{\sup_g N_g}}\Big), \qquad \mu_n \frac{\sup_g N_g}{n} \big\|M D_n^{\mathrm{diag}} M\big\|_{\mathcal{K}_n} \xrightarrow{p} 0. \tag{E.1}$$

The first bound is at the μ_n^{-1} scale of the studentized statistic; the second is the norm condition (D.2) of Proposition D.1, stated for $MD_n^{\mathrm{diag}}M$ because $\widehat{U} = M\widehat{U}$.

Proposition E.5 (First-order equivalence)**.** *Let $D_n^\star$ be the minimum-norm solution to $\mathcal{A}_n^* D = R_n^\star$ and let D_n^{diag} satisfy Assumption A10. With $\widehat{V}_{\mathrm{Riesz},n} := \langle \widehat{C}_n, D_n^\star\rangle_{\mathcal{K}_n}$ and $\widehat{V}_{\partial,n} := \langle P_{\mathrm{diag}}\widehat{C}_n, D_n^{\mathrm{diag}}\rangle_{\mathcal{K}_n^{\mathrm{diag}}}$, under Assumptions A1 to A5 and Assumption A10, $\mu_n(\widehat{V}_{\partial,n} - \widehat{V}_{\mathrm{Riesz},n}) = o_p(1)$.*

Proof. Since P_{diag} is the orthogonal projection of $\mathcal{K}_n$ onto $\mathcal{K}_n^{\mathrm{diag}}$ and D_n^{diag} is block diagonal,

$$\langle P_{\mathrm{diag}}\widehat{C}_n, D_n^{\mathrm{diag}}\rangle_{\mathcal{K}_n^{\mathrm{diag}}} = \langle \widehat{C}_n, D_n^{\mathrm{diag}}\rangle_{\mathcal{K}_n},$$

so $\widehat{V}_{\partial,n} = \widehat{V}(D_n^{\mathrm{diag}})$ in the notation of Proposition D.1. For a block-diagonal array the full and restricted adjoints agree,

$$[\mathcal{A}_n^*(D_n^{\mathrm{diag}})]_g = \sum_h M_{gh} D_{hh}^{\mathrm{diag}} M_{hg} = [(\mathcal{A}_n^{\mathrm{diag}})^* D_n^{\mathrm{diag}}]_g,$$

so by Lemma A.4 and the first bound of (E.1),

$$\mu_n \|\Sigma_n\|_{\mathcal{H}_n} \big\|\mathcal{A}_n^*(D_n^{\mathrm{diag}}) - R_n^\star\big\|_{\mathcal{H}_n} = \mu_n \cdot O_p\Big(\sqrt{\tfrac{\sup_g N_g}{n}}\Big) \cdot o_p\Big(\mu_n^{-1}\sqrt{\tfrac{n}{\sup_g N_g}}\Big) = o_p(1),$$

which is (D.1); the second bound of (E.1) is (D.2). Proposition D.1 gives $\mu_n(\widehat{V}_{\partial,n} - V) = o_p(1)$, and (D.5) gives $\mu_n(\widehat{V}_{\mathrm{Riesz},n} - V) = o_p(1)$. Subtracting yields the claim. □

Write the matrix B of (42) as $B = B_0 + E$, where B_0 keeps the diagonal blocks $M_{gg} \otimes M_{gg}$ and E collects the cross-cluster blocks $M_{gh} \otimes M_{gh}$, $h \neq g$. Vectors d are identified with block-diagonal arrays $\{D_{gg}\}$, and all norms are those of $\mathcal{H}_n$, i.e. n^{-1} times the Euclidean norm of the vectorization.

Lemma E.6 shows the block-wise inverse satisfies Assumption A10 under the leakage condition

$$\|E\|_{\mathrm{op}} \max_g \|M_{gg}^{-1}\|_{\mathrm{op}}^2 = o_p\Big(\frac{n}{\mu_n \sup_g N_g}\Big) \qquad \text{and} \qquad \max_g \|M_{gg}^{-1}\|_{\mathrm{op}} = O_p(1), \tag{E.2}$$

whose computable sufficient form is (Proj-Sparsity). Normalizing by M_{gg}^{-1} ties the row sums to within-cluster leverage, $\|M_{gg}^{-1} - I\|_{\mathrm{op}} \leq \varrho_n^{\mathrm{sp}}$ (Lemma E.4).

Lemma E.6 (Block-wise inversion error)**.** *Suppose every M_{gg} is nonsingular, and let $d^{\mathrm{CR}} := B_0^{-1} r$, the CR3 weights $D^{(1)}$ of Lemma E.4. Then*

$$\big\|\mathcal{A}_n^{\mathrm{diag}}(d^{\mathrm{CR}}) - R_n^\star\big\|_{\mathcal{H}_n} \;\leq\; \|E\|_{\mathrm{op}} \max_g \|M_{gg}^{-1}\|_{\mathrm{op}}^2 \, \|R_n^\star\|_{\mathcal{H}_n}, \qquad \big\|d^{\mathrm{CR}}\big\|_{\mathcal{H}_n} \;\leq\; \max_g \|M_{gg}^{-1}\|_{\mathrm{op}}^2 \, \|R_n^\star\|_{\mathcal{H}_n}, \tag{E.3}$$

and $\|E\|_{\mathrm{op}} \leq \varrho_n^{\mathrm{sp}}$, $\max_g \|M_{gg}^{-1}\|_{\mathrm{op}} \leq 1 + \varrho_n^{\mathrm{sp}}$. Consequently, under Assumption A3, (E.2) implies Assumption A10 for the CR3 weights, and so does (Proj-Sparsity). By Lemma E.4(ii) the same holds for $D^{(\delta)}$ with any $\delta \in [0,1]$, which covers CR0 and CR2, and for CR1, whose degrees-of-freedom factor $c_n = \frac{G}{G-1}\frac{n-1}{n-p}$ satisfies $|c_n - 1| = O_p(\varepsilon_n + \sup_g N_g/n)$.

Proof. Bounds. Since $\mathcal{A}_n^{\mathrm{diag}}$ is represented by B and $B_0 d^{\mathrm{CR}} = r$, $\mathcal{A}_n^{\mathrm{diag}}(d^{\mathrm{CR}}) - R_n^\star = B d^{\mathrm{CR}} - r = E d^{\mathrm{CR}}$, and $\|d^{\mathrm{CR}}\| \leq \|B_0^{-1}\|_{\mathrm{op}} \|r\| = \max_g \|M_{gg}^{-1}\|_{\mathrm{op}}^2 \|R_n^\star\|$ because $\|(M_{gg} \otimes M_{gg})^{-1}\|_{\mathrm{op}} = \|M_{gg}^{-1}\|_{\mathrm{op}}^2$; this gives (E.3). E is symmetric, since the transpose of $M_{gh} \otimes M_{gh}$ is $M_{hg} \otimes M_{hg}$, so its operator norm is bounded by its maximal block row sum, $\|E\|_{\mathrm{op}} \leq \max_g \sum_{h \neq g} \|M_{gh}\|_{\mathrm{op}}^2 \leq \max_g \sum_{h \neq g} \|M_{gg}\|_{\mathrm{op}}^2 \|M_{gg}^{-1} M_{gh}\|_{\mathrm{op}}^2 \leq \varrho_n^{\mathrm{sp}}$, using $\|M_{gg}\|_{\mathrm{op}} \leq 1$. Finally $\|M_{gg}^{-1}\|_{\mathrm{op}} \leq 1 + \varepsilon_n \leq 1 + \varrho_n^{\mathrm{sp}}$ by Lemma E.4(i).

Assumption A10 for CR3. With $\|R_n^\star\|_{\mathcal{H}_n} = O_p(\sqrt{\sup_g N_g/n})$ (Lemma A.3), the first bound of (E.3) is $o_p\big(\mu_n^{-1}\sqrt{n/\sup_g N_g}\big)$ under the first part of (E.2), which is the first condition of Assumption A10. For the second, $\|M d^{\mathrm{CR}} M\| \leq \|d^{\mathrm{CR}}\| = O_p(\sqrt{\sup_g N_g/n})$ under the second part of (E.2), so $\mu_n(\sup_g N_g/n)\|M d^{\mathrm{CR}} M\| = O_p\big((\mu_n^{2/3} \sup_g N_g/n)^{3/2}\big) = o_p(1)$, since $2/3 \leq (2+\lambda)/(2+2\lambda)$ for $\lambda \leq 2$, so that $\mu_n^{2/3} \sup_g N_g/n \to 0$ by Assumption A3. (Proj-Sparsity) implies (E.2) by the bounds just shown.

Other leverage corrections. By Lemma E.4(i)–(ii) and (Proj-Sparsity),

$$\|D^{(\delta)} - D^{(1)}\| = O_p\big(\varrho_n^{\mathrm{sp}} \sqrt{\sup_g N_g/n}\big) = o_p\big(\mu_n^{-1}\sqrt{n/\sup_g N_g}\big).$$

Since $\mathcal{A}_n^{\mathrm{diag}}$ is a contraction, both conditions of Assumption A10 transfer from $D^{(1)}$ to $D^{(\delta)}$. For CR1, $c_n - 1 = \frac{1}{G-1}\frac{n-1}{n-p} + \frac{p-1}{n-p}$, and $1/G \leq \sup_g N_g/n$ while $p/n \leq \varepsilon_n$ by Lemma E.4(iii), so

$$\|(c_n - 1) D^{(0)}\| = O_p\big((\varrho_n^{\mathrm{sp}} + \sup_g N_g/n)\sqrt{\sup_g N_g/n}\big).$$

The first term is handled as before, and the second is $o_p\big(\mu_n^{-1}\sqrt{n/\sup_g N_g}\big)$ because $\mu_n(\sup_g N_g/n)^2 \le \mu_n(\sup_g N_g/n)^{3/2} \to 0$. □

E.3 The spillover decomposition and the level factor

The decomposition (46) of Section 7.3 is the following lemma.

Lemma E.7 (Block spillover decomposition)**.** *Partition $\{1,\dots,n\}$ into clusters $g = 1,\dots,G$ and let M_{gh} denote the corresponding blocks of $M = I - H$. Then idempotence and symmetry give*

$$\sum_{h} M_{gh}M_{gh}' = M_{gg}, \qquad \sum_{h\neq g} M_{gh}M_{gh}' = M_{gg} - M_{gg}^2 = M_{gg}H_{gg}. \tag{E.4}$$

Suppose further that $M_{gg} \succ 0$, let P_g denote the orthogonal projector onto $\mathrm{Range}(H_{gg})$ and $H_{gg}^{+1/2}$ the square root of the Moore–Penrose inverse of H_{gg}, and set

$$L_{gk} \;=\; M_{gg}^{-1/2}M_{gk}, \qquad \tilde L_{gk} \;=\; H_{gg}^{+1/2}L_{gk}\ \ (k\neq g), \qquad \tilde\Sigma_g \;=\; \sum_{k\neq g}\tilde L_{gk}\Sigma_k\tilde L_{gk}'.$$

Then the weights resolve the identity,

$$\sum_{k} L_{gk}L_{gk}' = I_{N_g}, \qquad L_{gg}L_{gg}' = M_{gg}, \qquad \sum_{k\neq g} L_{gk}L_{gk}' = H_{gg}, \tag{E.5}$$

with $\sum_{k\neq g}\tilde L_{gk}\tilde L_{gk}' = P_g$, which is I_{N_g} when H_{gg} is nonsingular; the normalized block moment is

$$M_{gg}^{-1/2}\,\mathbb{E}[\widehat U_g\widehat U_g' \mid X,W]\,M_{gg}^{-1/2} \;=\; M_{gg}^{1/2}\Sigma_g M_{gg}^{1/2} \;+\; H_{gg}^{1/2}\tilde\Sigma_g H_{gg}^{1/2}; \tag{E.6}$$

and the bias decomposes as

$$\mathfrak{B}_g \;:=\; M_{gg}^{-1/2}\,\mathbb{E}[\widehat U_g\widehat U_g' \mid X,W]\,M_{gg}^{-1/2} - \Sigma_g \;=\; H_{gg}^{1/2}\big(\tilde\Sigma_g - \Sigma_g\big)H_{gg}^{1/2} \;+\; R_g, \tag{E.7}$$

where, writing $[\,C,D\,] = CD - DC$,

$$R_g \;=\; M_{gg}^{1/2}\Sigma_g M_{gg}^{1/2} + H_{gg}^{1/2}\Sigma_g H_{gg}^{1/2} - \Sigma_g \;=\; M_{gg}^{1/2}\big[\Sigma_g, M_{gg}^{1/2}\big] + H_{gg}^{1/2}\big[\Sigma_g, H_{gg}^{1/2}\big]. \tag{E.8}$$

$R_g = 0$ whenever Σ_g commutes with H_{gg}, and $R_g \equiv 0$ when $N_g = 1$.

Proof. Idempotence gives $(M^2)_{gg} = M_{gg}$, that is $\sum_h M_{gh}M_{hg} = M_{gg}$, and symmetry gives $M_{hg} = M_{gh}'$, which is the first identity in (E.4); separating $h = g$ and using $I_{N_g} - M_{gg} = H_{gg}$ gives the second.

Conjugating the first identity by $M_{gg}^{-1/2}$ yields $\sum_k L_{gk}L_{gk}' = I_{N_g}$. For the second identity in (E.5), $L_{gg}L_{gg}' = M_{gg}^{-1/2}M_{gg}^2M_{gg}^{-1/2} = M_{gg}$, and the third follows by subtraction. Since $L_{gk}L_{gk}' \preceq H_{gg}$ for each $k \neq g$, $\mathrm{Range}(L_{gk}) \subseteq \mathrm{Range}(H_{gg})$, so $H_{gg}^{1/2}\tilde L_{gk} = P_gL_{gk} = L_{gk}$; conjugating the third identity by $H_{gg}^{+1/2}$ gives $\sum_{k\neq g}\tilde L_{gk}\tilde L_{gk}' = H_{gg}^{+1/2}H_{gg}H_{gg}^{+1/2} = P_g$.

For (E.6), conjugate (7) by $M_{gg}^{-1/2}$: the first term becomes

$$M_{gg}^{-1/2} M_{gg} \Sigma_g M_{gg} M_{gg}^{-1/2} = M_{gg}^{1/2} \Sigma_g M_{gg}^{1/2},$$

and the second becomes

$$\sum_{k \neq g} L_{gk} \Sigma_k L_{gk}' = \sum_{k \neq g} H_{gg}^{1/2} \tilde{L}_{gk} \Sigma_k \tilde{L}_{gk}' H_{gg}^{1/2} = H_{gg}^{1/2} \tilde{\Sigma}_g H_{gg}^{1/2},$$

using $L_{gk} = H_{gg}^{1/2} \tilde{L}_{gk}$.

For (E.7), subtract Σ_g from (E.6) and add and subtract $H_{gg}^{1/2} \Sigma_g H_{gg}^{1/2}$. For the commutator form, write $S = M_{gg}^{1/2}$ and $T = H_{gg}^{1/2}$, so that $S^2 + T^2 = I_{N_g}$ by $M_{gg} + H_{gg} = I_{N_g}$. Substituting $\Sigma_g = (S^2 + T^2)\Sigma_g$ and factoring S and T on the left,

$$R_g = S\Sigma_g S + T\Sigma_g T - S^2\Sigma_g - T^2\Sigma_g = S(\Sigma_g S - S\Sigma_g) + T(\Sigma_g T - T\Sigma_g) = S[\Sigma_g, S] + T[\Sigma_g, T].$$

If Σ_g commutes with H_{gg} it commutes with S and T, both functions of H_{gg}, so both commutators vanish. When $N_g = 1$ all quantities are scalars and $R_g = (1 - h_{ii})\sigma_i^2 + h_{ii}\sigma_i^2 - \sigma_i^2 = 0$. □

At $N_g = 1$, (E.7) reduces to (5) with $R_g \equiv 0$: $M_{gg}^{-1/2}$ plays the role of dividing by M_{ii}, and $\tilde{\Sigma}_g$ that of $\tilde{\sigma}_i^2$. The two features without a scalar counterpart are the rotation in the contrast and the remainder.

Remark E.8 (Congruence averaging). The scalar average $\tilde{\sigma}_i^2 = \sum_{j \neq i} w_{ij} \sigma_j^2$ is a convex combination, so equal variances give a zero contrast. Its block counterpart $\tilde{\Sigma}_g = \sum_{k \neq g} \tilde{L}_{gk} \Sigma_k \tilde{L}_{gk}'$ is a sum of congruences with $\sum_{k \neq g} \tilde{L}_{gk} \tilde{L}_{gk}' = P_g$, which preserves positive semidefiniteness and total weight but not eigenbasis. $\Sigma_k = \sigma^2 I$ still gives a zero contrast, since $\tilde{\Sigma}_g = \sigma^2 P_g$ and $H_{gg}^{1/2}(P_g - I)H_{gg}^{1/2} = 0$, but equal blocks $\Sigma_k \equiv \Sigma$ do not: for illustration, with $N_g = 2$, H_{gg} nonsingular and two neighbouring clusters whose $\tilde{L}_{gk}$ are the two coordinate projections, $\tilde{\Sigma}_g = \operatorname{diag}(\Sigma_{11}, \Sigma_{22}) \neq \Sigma$ whenever $\Sigma_{12} \neq 0$.

Remark E.9 (Closed form of the remainder). Let $H_{gg} = O \operatorname{diag}(\eta_1, \ldots, \eta_{N_g}) O'$ with $\eta_i = \sin^2 \theta_i$, $\theta_i \in [0, \pi/2]$. Then $M_{gg}^{1/2}$ and $H_{gg}^{1/2}$ have eigenvalues $\cos\theta_i$ and $\sin\theta_i$ in the basis O, and with $\check{\Sigma} := O'\Sigma_g O$, the first form of (E.8) gives

$$\big(O' R_g O\big)_{ij} = \big(\cos\theta_i \cos\theta_j + \sin\theta_i \sin\theta_j - 1\big)\, \check{\Sigma}_{ij} = \big(\cos(\theta_i - \theta_j) - 1\big)\, \check{\Sigma}_{ij}. \tag{E.9}$$

Every factor lies in $[-1, 0]$ and vanishes exactly when $\eta_i = \eta_j$. So R_g only shrinks, and only the entries of Σ_g linking directions the projection absorbs to different degrees; and $R_g = 0$ if and only if Σ_g is block diagonal with respect to the eigenspaces of H_{gg}, which is commutation in a checkable form.

Proof of Proposition 7.3. For part (i), with $\Sigma = \sigma^2 I_n$, $\mathbb{E}[\widehat{U}_g \widehat{U}_g' \mid X, W] = \sigma^2 \sum_k M_{gk} M_{kg} = \sigma^2 M_{gg}$ by symmetry and idempotence of M. Hence

$$\mathbb{E}[\widehat{V}_{\mathrm{CR0}} \mid X, W] = \sigma^2 n^{-2} \sum_g z_g' M_{gg} z_g = \sigma^2 n^{-2} \Big(z'z - \sum_g z_g' H_{gg} z_g \Big),$$

while $V = \sigma^2 n^{-2} z'z$; the ratio is $1-\kappa_n$, and CR1 multiplies CR0 by c_n. CR2 and CR3 use the weights $M_{gg}^{-\delta} z_g z_g' M_{gg}^{-\delta}$ with $\delta = 1/2$ and $\delta = 1$, so

$$\mathbb{E}[\widehat{V}_{\mathrm{CR2}} \mid X, W] = \sigma^2 n^{-2} \sum_g z_g' M_{gg}^{-1/2} M_{gg} M_{gg}^{-1/2} z_g = \sigma^2 n^{-2} z'z = V,$$

and, since $M_{gg}^{-1} = I + H_{gg} M_{gg}^{-1}$, $\mathbb{E}[\widehat{V}_{\mathrm{CR3}} \mid X, W] = \sigma^2 n^{-2} \sum_g z_g' M_{gg}^{-1} z_g = (1+\tilde{\kappa}_n) V$. For the bounds, let $\eta_{gj} \in [0,1)$ be the eigenvalues of H_{gg}, below one because M_{gg} is nonsingular, and c_{gj} the coordinates of z_g in a common eigenbasis of H_{gg} and M_{gg}. With weights $\pi_{gj} := c_{gj}^2 / \sum_{g,j} c_{gj}^2$, $\kappa_n = \sum \pi_{gj} \eta_{gj}$ and $\tilde{\kappa}_n = \sum \pi_{gj} \eta_{gj} / (1-\eta_{gj})$. Jensen's inequality for the convex map $\eta \mapsto \eta/(1-\eta)$ gives $\tilde{\kappa}_n \geq \kappa_n/(1-\kappa_n)$, and $\eta_{gj} \leq \bar{\eta}$ gives $\tilde{\kappa}_n \leq \kappa_n/(1-\bar{\eta})$.

For part (ii), Assumption A1 gives (7) with every $\Sigma_k \succeq 0$, and conjugating it by M_{gg}^{-1} gives

$$M_{gg}^{-1} \mathbb{E}[\widehat{U}_g \widehat{U}_g' \mid X, W] M_{gg}^{-1} = \Sigma_g + M_{gg}^{-1} \Big(\sum_{k \neq g} M_{gk} \Sigma_k M_{kg} \Big) M_{gg}^{-1},$$

and the second term is positive semidefinite, so

$$\mathbb{E}[\widehat{V}_{\mathrm{CR3}} \mid X, W] - V = \frac{1}{n^2} \sum_g z_g' M_{gg}^{-1} \Big(\sum_{k \neq g} M_{gk} \Sigma_k M_{kg} \Big) M_{gg}^{-1} z_g \;\geq\; 0.$$

For part (iii), when the columns of $[X\ W]$ are constant within clusters, $H = R(R'R)^{+}R'$ with R stacking $\iota_g r_g'$, so

$$H_{gg} = \iota_g r_g' \Big(\sum_k N_k r_k r_k' \Big)^{+} r_g \iota_g' = \frac{h_g^c \iota_g \iota_g'}{N_g}, \qquad \sum_g h_g^c = \operatorname{tr}\Big(\Big(\sum_k N_k r_k r_k' \Big)^{+} \sum_g N_g r_g r_g' \Big) = p.$$

Since $z = \widetilde{X} c \in \operatorname{col}([X\ W])$ it is constant within clusters, $z_g = \zeta_g \iota_g$, and $z_g' H_{gg} z_g = \zeta_g^2 N_g h_g^c$, $z_g' z_g = N_g \zeta_g^2$, so $\kappa_n = \sum_g N_g \zeta_g^2 h_g^c / \sum_g N_g \zeta_g^2$. □

Off-block mass. The regimes in which κ_n is large are those of Section 2; it remains to note that off-block mass itself is the rule, not a feature of dyadic designs, non-nested fixed effects or spatial controls. Every M_{gk} with $k \neq g$ vanishes if and only if $\operatorname{col}([X\ W])$ is the direct sum of its intersections with the cluster coordinate subspaces, as with cluster fixed effects alone or interacted with within-cluster variables; an intercept or a cluster-level treatment already makes H dense. What sets the level of the bias is not whether clusters are linked but κ_n: in Experiment N, node fixed effects and dyad-level controls give the same off-block mass but different κ_n and different biases (Section 10.2).

Proof of (47). For each g, the triangle inequality and $\|AA'\|_{\mathrm{op}} = \|A\|_{\mathrm{op}}^2$ give $\sum_{h \neq g} \|M_{gg}^{-1} M_{gh}\|_{\mathrm{op}}^2 \geq \|M_{gg}^{-1} \big(\sum_{h \neq g} M_{gh} M_{hg} \big) M_{gg}^{-1}\|_{\mathrm{op}} = \|H_{gg} M_{gg}^{-1}\|_{\mathrm{op}}$, by (E.4) and because M_{gg} and $H_{gg} = I - M_{gg}$ commute. In their common eigenbasis $\|H_{gg} M_{gg}^{-1}\|_{\mathrm{op}} = \max_j \eta_j/(1-\eta_j) \geq \max_j \eta_j = \|H_{gg}\|_{\mathrm{op}}$, with $\eta_j \in [0,1)$ the eigenvalues of H_{gg}. Finally $\kappa_n = \sum_g z_g' H_{gg} z_g / \sum_g \|z_g\|^2 \leq \max_g \|H_{gg}\|_{\mathrm{op}}$. □

E.4 Existence

Throughout, $E_g \in \mathbb{R}^{n\times N_g}$ selects the observations of cluster g, so that $\mathcal{W}_g = \mathrm{col}(E_g)$ and $M_{gg} = E_g' M E_g$.

Proof of Proposition 7.4. Since M is symmetric and idempotent, every $v \in \mathbb{R}^{N_g}$ satisfies

$$v' M_{gg} v = v' E_g' M' M E_g v = \|M E_g v\|^2.$$

Hence $M_{gg} v = 0$ if and only if $M E_g v = 0$ (the converse is immediate from $M_{gg} = E_g' M E_g$), that is, if and only if $E_g v \in \mathrm{Ker}(M) = \mathrm{col}([X\ W])$. Since $v \mapsto E_g v$ is a bijection onto $\mathcal{W}_g$, M_{gg} is singular exactly when $\mathcal{W}_g \cap \mathrm{col}([X\ W]) \neq \{0\}$. □

Proof of Proposition 7.5. $\mathcal{D}_{\mathrm{part}} = \mathcal{D}_{\mathrm{Riesz}}$ is the solvability equivalence of Proposition 7.1, which holds at every n. The inclusions $\mathcal{D}_{\mathrm{CR}}^1 \subseteq \mathcal{D}_{\mathrm{CR}} \subseteq \mathcal{D}_{\mathrm{part}}$ hold by definition, since (d) adds a condition to (c) and (c) adds one to (a). It remains to exhibit two sequences.

$\mathcal{D}_{\mathrm{part}} \setminus \mathcal{D}_{\mathrm{CR}}$. Take G clusters of size two, W the G cluster indicators, and a scalar regressor with $x_{g1} = 1$, $x_{g2} = -1$ in every cluster, so $\widetilde{X} = X$, $\Gamma_n = 1$ and Assumption A4 holds. Every M_{gg} is singular by Proposition 7.4. Write $e := (1,-1)'/\sqrt{2}$ and $\bar{e}_g := E_g e$, so that $M_W = \sum_g \bar{e}_g \bar{e}_g'$ and $M = M_W - \bar{w}\bar{w}'$ with $\bar{w} := G^{-1/2}\sum_g \bar{e}_g$.Since $M = M M_W$ and $M_W E_g = \bar{e}_g e'$, for $\Sigma \in \mathcal{H}_n$,

$$\mathcal{A}_n(\Sigma) = \sum_g t_g\, M \bar{e}_g \bar{e}_g' M, \qquad t_g := e'\Sigma_g e.$$

In the orthonormal basis $\{\bar{e}_g\}$ of $\mathrm{col}(M_W)$ this is $(I_G - ww')\,\mathrm{diag}(t)(I_G - ww')$ with $w = G^{-1/2}\mathbf{1}_G$. If it vanishes, then $\mathrm{diag}(t) = wb' + bw'$ for some $b \in \mathbb{R}^G$, and the off-diagonal entries give $b_g + b_h = 0$ for all $g \neq h$, which for $G \geq 3$ forces $b = 0$ and hence $t = 0$. So $\mathrm{Ker}(\mathcal{A}_n) = \{\Delta : e'\Delta_g e = 0\ \forall g\}$. Since $z_g = \widetilde{X}_g \Gamma_n^{-1} a = \sqrt{2}\, a\, e$, every $\Delta \in \mathrm{Ker}(\mathcal{A}_n)$ has $\sum_g z_g' \Delta_g z_g = 2a^2 \sum_g e'\Delta_g e = 0$, so (38) and hence Assumption A5(i) hold for every $G \geq 3$. The sequence therefore lies in $\mathcal{D}_{\mathrm{part}}$ but not in $\mathcal{D}_{\mathrm{CR}}$.

$\mathcal{D}_{\mathrm{CR}} \setminus \mathcal{D}_{\mathrm{CR}}^1$. Take scalar clusters ($N_g = 1$) and designs with $p = \mathrm{rank}([X\ W])$ satisfying $p/n \to \alpha \in (0, 1/2)$ and equal leverage $h_{ii} = p/n$ for all i (for example, $[X\ W]$ with orthonormal columns drawn from a Hadamard or Fourier basis), with X chosen so that $\Gamma_n \to \Gamma_0 \succ 0$. Then each $M_{gg} = 1 - p/n > 0$ is nonsingular. The inverse-Hadamard system $M \odot M$ is strictly diagonally dominant, since $(1 - h_{ii})^2 > h_{ii}(1 - h_{ii}) = \sum_{j\neq i} M_{ij}^2$ whenever $h_{ii} < 1/2$, so it is nonsingular, $\mathrm{Ker}(\mathcal{A}_n) = \{0\}$ and Assumption A5(i) holds for every contrast. But $\varrho_n^{\mathrm{sp}} = \max_i \sum_{j\neq i} M_{ij}^2 / M_{ii}^2 = (p/n)/(1 - p/n) \to \alpha/(1-\alpha) > 0$, so (Proj-Sparsity) fails. The failure is substantive: with equal leverage $\kappa_n = p/n \to \alpha$, so within-cluster leverage stays first order and CR0 is biased by $-\kappa_n \to -\alpha$ even under homoskedasticity (Proposition 7.3). □

A singular M_{gg} is immaterial to computing the Riesz estimator but not to identification: it places entire within-cluster directions in $\mathrm{Ker}(\mathcal{A}_n)$, and Proposition 6.3 says which contrasts survive.

Proposition E.10 (Confined directions and the range condition)**.** *Let $0 \neq v \in \mathrm{Ker}(M_{gg})$ and decompose $E_g v = W\alpha + \widetilde{X} b$, possible and unique in b by Proposition 7.4. Then every array Δ with $\Delta_g = vw' + wv'$ for some $w \in \mathbb{R}^{N_g}$ and $\Delta_h = 0$ for $h \neq g$ lies in $\mathrm{Ker}(\mathcal{A}_n)$, and (38) holds on these arrays if and only if $a'b = 0$. In particular a confined direction lying in $\mathrm{col}(W)$ ($b = 0$), such as a cluster*

fixed effect, is harmless for every contrast; a confined direction with a component in $\mathrm{col}(\widetilde{X})$ *violates Assumption A5(i) for every contrast with* $a'b \neq 0$.

Proof. By Proposition 7.4, $v \in \mathrm{Ker}(M_{gg})$ gives $ME_gv = 0$, and the decomposition $E_gv = W\alpha + \widetilde{X}b$ is possible because $E_gv \in \mathrm{col}([X\ W]) = \mathrm{col}(W) \oplus \mathrm{col}(\widetilde{X})$, with b unique by the directness of the sum.

For the first claim, the array Δ with $\Delta_g = vw' + wv'$ and $\Delta_h = 0$ otherwise has $\widetilde{\Delta} = E_g(vw' + wv')E_g'$, so

$$M\widetilde{\Delta}M = (ME_gv)(ME_gw)' + (ME_gw)(ME_gv)' = 0,$$

and $\Delta \in \mathrm{Ker}(\mathcal{A}_n)$.

For the second, $\sum_h z_h'\Delta_h z_h = 2(z_g'v)(w'z_g)$, which vanishes for every $w \in \mathbb{R}^{N_g}$ if and only if $z_g'v = 0$ — take $w = z_g$ for necessity. Using $\widetilde{X}'W = 0$ and $\widetilde{X}'\widetilde{X} = n\Gamma_n$,

$$z_g'v = z'E_gv = a'\Gamma_n^{-1}\widetilde{X}'\big(W\alpha + \widetilde{X}b\big) = n\,a'b,$$

so (38) holds on these arrays if and only if $a'b = 0$. The two special cases follow by setting $b = 0$ and $b \neq 0$ respectively. □

The condition $a'b = 0$ is necessary, not sufficient: the kernel can contain directions spread across several clusters in which the contributions offset, and these must be checked against (38) directly (Chen, 2026, Example 3.6). Its content is that absorbing a cluster effect destroys information about cluster levels, on which a within-cluster coefficient does not depend; a contrast that loads on a regressor confined to one cluster is itself asking about that cluster's level.

Remark E.11 (Where the hierarchy bottoms out). If $R_n^\star \notin \mathrm{Range}(\mathcal{A}_n^*)$, no weighting of the residual moment array, full or restricted (Proposition 7.1), reproduces the variance functional. By Proposition 6.4 the estimator nonetheless remains first-order valid wherever (40) holds, so the hierarchy bottoms out at a defect large enough to matter at the μ_n^{-1} scale. This is a limit of the information, not of the estimator: V is then not a function of the population residual second moments (Chen, 2026). A stronger covariance model shrinks $\mathrm{Ker}(\mathcal{A}_n)$; leave-out constructions, which are β-free but not translation invariant, reach the cross block $\Pi_{[X\ W]}\Sigma M$; and procedures that do not estimate the variance, such as a bootstrap, lie outside the class these results constrain.

F Proofs for Section 8

Throughout, $\langle\cdot,\cdot\rangle_{\mathcal{H}_n}$ and $\langle\cdot,\cdot\rangle_{\mathcal{K}_n}$ carry the factor n^{-2} of Section 4, $\Pi_{\mathcal{D}}$ is self-adjoint on $\mathcal{K}_n$, and $\Delta := \Sigma_n - \Omega_n \in \mathcal{D}_n$.

Three auxiliary results come first. Lemma 8.1 is stated in Section 8.1.

Proof of Lemma 8.1

Proof. All moments are conditional on (X, W), so K is fixed. Since $U'KU$ is a scalar, $U'KU = U'K'U = U'K_sU$ with $K_s = (K + K')/2$, and $\|K_s\|_F \leq \|K\|_F$ by the triangle inequality; we may therefore take K symmetric.

Variance decomposition. Write σ_{ij} for the entries of Σ_n and κ_{ijkl} for the joint fourth cumulant of (u_i, u_j, u_k, u_l). Since the means are zero, $\mathbb{E}[u_iu_ju_ku_l] = \sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk} + \kappa_{ijkl}$, so $\mathrm{Cov}(u_iu_j, u_ku_l) = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk} + \kappa_{ijkl}$ and, K being symmetric,

$$\mathrm{Var}(U'KU) = \sum_{i,j,k,l} K_{ij}K_{kl}\,\mathrm{Cov}(u_iu_j, u_ku_l) = 2\,\mathrm{tr}(K\Sigma_nK\Sigma_n) + \sum_{i,j,k,l} K_{ij}K_{kl}\kappa_{ijkl}.$$

Two moment bounds are used below. By Lyapunov's inequality $\sigma_{ii} \le (\mathbb{E}u_i^4)^{1/2} \le m_4^{1/2}$, so Cauchy–Schwarz gives $|\sigma_{ij}| \le (\sigma_{ii}\sigma_{jj})^{1/2} \le m_4^{1/2}$; and by Hölder's inequality $|\mathbb{E}[u_iu_ju_ku_l]| \le m_4$, so $|\kappa_{ijkl}| \le m_4 + 3m_4 = 4m_4$.

Gaussian part. Since $\Sigma_n \succeq 0$, $\mathrm{tr}(K\Sigma_nK\Sigma_n) = \|\Sigma_n^{1/2}K\Sigma_n^{1/2}\|_F^2 \le \|\Sigma_n\|_{\mathrm{op}}^2\|K\|_F^2$. By Assumption A7, $\sigma_{ij} = 0$ for $(i,j) \notin \mathcal{S}_n$, so each row of Σ_n has at most d_n nonzero entries, and the Schur test gives $\|\Sigma_n\|_{\mathrm{op}} \le \max_i \sum_j |\sigma_{ij}| \le d_n m_4^{1/2}$. This part is therefore at most $2m_4d_n^2\|K\|_F^2$.

Which cumulants vanish. Call (i,j,k,l) *connected* if the graph on the four positions is connected, where two positions are joined when their values form a pair in $\mathcal{S}_n$; equal values are joined, since $\mathcal{S}_n$ contains the diagonal. If (i,j,k,l) is not connected, the positions split into two nonempty groups whose value sets are disjoint and share no pair in $\mathcal{S}_n$. By Assumption A7 the two groups are then independent, and a joint cumulant of variables that split into independent groups is zero, so $\kappa_{ijkl} = 0$.

Counting the connected terms. Let $A_{ij} := \mathbb{1}\{(i,j) \in \mathcal{S}_n\}$, whose row sums are at most d_n. A connected graph on four positions contains a spanning tree, and by Cayley's formula there are $4^{4-2} = 16$ labelled trees τ on four vertices, so $\mathbb{1}\{\text{connected}\} \le \sum_\tau \prod_{pq\in\tau} A_{v_pv_q}$, with v_p the value at position p. Fix τ and use $|K_{ij}||K_{kl}| \le \frac{1}{2}(K_{ij}^2 + K_{kl}^2)$. It suffices to show that, for fixed values at two positions, the sum of $\prod_{pq\in\tau} A_{v_pv_q}$ over the other two is at most d_n^2. Because τ spans, some free position p is joined in τ to a fixed position r, and the other free position p' is joined to some position by an edge distinct from pr. Bounding every other factor by one, the sum over $v_{p'}$ is at most d_n for each v_p, and the sum over v_p of $A_{v_rv_p}$ is at most d_n. Each tree therefore contributes at most $d_n^2\sum_{i,j}K_{ij}^2 = d_n^2\|K\|_F^2$, and the cumulant part is at most $4m_4 \cdot 16\, d_n^2\|K\|_F^2$.

Adding the two parts, $\mathrm{Var}(U'KU) \le (2 + 64)\, m_4d_n^2\|K\|_F^2$, so $C_0 = 66$. □

Lemma F.1 (Size of the target and of the covariance)**.** *Under Assumptions A7 and A8:*

(i) $\|\Sigma_n\|_{\mathcal{H}_n} \le m_4^{1/2}\sqrt{d_n/n}$*;*

(ii) $\|R_D^T\|_{\mathcal{H}_n} \le n\big(\sum_{(i,j)\in\mathcal{S}_n}\omega_i^2\omega_j^2\big)^{1/2}$*, which is* $O_p(\sqrt{d_n/n})$ *under Assumption A9(iii);*

(iii) $\sum_{(i,j)\in\mathcal{S}_n}\omega_i^2\omega_j^2 \le d_n\sum_i\omega_i^4$.

Proof. All moments are conditional on (X, W). For every $B \in \mathcal{K}_n$, $\|B\|_{\mathcal{H}_n} = n^{-1}\|B\|_F$.

(i) For $(i,j) \notin \mathcal{S}_n$ we have $i \ne j$, and Assumption A7 with $I = \{i\}$, $J = \{j\}$ makes u_i and u_j independent with mean zero, so $\sigma_{ij} = 0$. For every pair, $|\sigma_{ij}| \le m_4^{1/2}$, as in the proof of Lemma 8.1. Each row of $\mathcal{S}_n$ has at most d_n entries, so $|\mathcal{S}_n| \le nd_n$ and

$$\|\Sigma_n\|_F^2 = \sum_{(i,j)\in\mathcal{S}_n}\sigma_{ij}^2 \le nd_nm_4.$$

Dividing by n^2 and taking square roots gives the bound.

(ii) The factor n^{-2} in the inner product does not affect orthogonality, so $\Pi_{\mathcal{D}}$ is also the orthogonal projection in the Frobenius norm. Let $\Pi_{\mathcal{S}}$ be the orthogonal projection onto $\mathcal{H}_n^{\mathcal{S}}$, which sets the entries outside $\mathcal{S}_n$ to zero; the result is symmetric because $\mathcal{S}_n$ is. Since $\mathcal{D}_n \subseteq \mathcal{H}_n^{\mathcal{S}}$ by Assumption A8, $\Pi_{\mathcal{D}} = \Pi_{\mathcal{D}}\Pi_{\mathcal{S}}$, and since orthogonal projections are contractions,

$$\|R_{\mathcal{D}}^T\|_{\mathcal{H}_n} = n\,\|\Pi_{\mathcal{D}}(\omega\omega')\|_F \le n\,\|\Pi_{\mathcal{S}}(\omega\omega')\|_F = n\Big(\sum_{(i,j)\in\mathcal{S}_n} \omega_i^2\omega_j^2\Big)^{1/2}.$$

Under Assumption A9(iii) the sum is $O_p(d_n/n^3)$, so the bound is $O_p(\sqrt{d_n/n})$.

(iii) By $ab \le \frac{1}{2}(a^2+b^2)$, $\omega_i^2\omega_j^2 \le \frac{1}{2}(\omega_i^4+\omega_j^4)$. Summing over $\mathcal{S}_n$, the two halves are equal because $\mathcal{S}_n$ is symmetric, so

$$\sum_{(i,j)\in\mathcal{S}_n} \omega_i^2\omega_j^2 \le \sum_{(i,j)\in\mathcal{S}_n} \omega_i^4 = \sum_i |\mathcal{S}(i)|\,\omega_i^4 \le d_n \sum_i \omega_i^4. \qquad \square$$

Lemma F.2 (A central limit theorem for dependency graphs). *Suppose Assumption A7 holds, $V > 0$, $n\sum_i \omega_i^2 = O_p(1)$, $n^3\sum_i \omega_i^4 = O_p(1)$, and Assumption A9(iv) holds. Then the Wasserstein distance between the conditional law of $V^{-1/2}\omega'U$ and $\mathcal{N}(0,1)$ is*

$$O_p\Big(\big(d_n\mu_n^{3/4}/n\big)^2 + \big(d_n\mu_n^{2/3}/n\big)^{3/2}\Big),$$

and Condition 1 holds whenever $\mu_n^{3/4} d_n/n \to 0$.

Proof. All moments and laws are conditional on (X, W), given which ω, $\mathcal{S}_n$ and V are fixed. Let $\xi_i := \omega_i u_i$, so that $\omega'U = \sum_i \xi_i$, $\mathbb{E}\xi_i = 0$ and $\mathrm{Var}(\sum_i \xi_i) = \omega'\Sigma_n\omega = V$.

Dependency neighbourhoods. Take $N_i := \mathcal{S}(i)$, which contains i and has at most d_n elements. The sets $I = \{i\}$ and $J = \{1,\dots,n\} \setminus \mathcal{S}(i)$ are disjoint with $(I \times J) \cap \mathcal{S}_n = \emptyset$, so by Assumption A7 ξ_i is independent of $\{\xi_j\}_{j\notin\mathcal{S}(i)}$. This is the definition of Ross (2011). His proof also uses that (ξ_i, ξ_j) is independent of (ξ_k, ξ_l) whenever $k, l \notin \mathcal{S}(i) \cup \mathcal{S}(j)$, which is Assumption A7 with $I = \{i, j\}$ and $J = \{k, l\}$.

Stein bound. Ross (2011, Theorem 3.6), with $D = d_n$ and $\sigma^2 = V$, gives

$$d_{\mathrm{W}} \le \frac{d_n^2}{V^{3/2}} \sum_i |\omega_i|^3 \mathbb{E}|u_i|^3 + \frac{\sqrt{28}\, d_n^{3/2}}{\sqrt{\pi}\, V}\Big(\sum_i \omega_i^4 \mathbb{E}u_i^4\Big)^{1/2}.$$

Rates. By Lyapunov's inequality $\mathbb{E}|u_i|^3 \le (\mathbb{E}u_i^4)^{3/4} \le m_4^{3/4}$. Writing $|\omega_i|^3 = |\omega_i|\,\omega_i^2$, Cauchy–Schwarz gives

$$\textstyle\sum_i |\omega_i|^3 \le (\sum_i \omega_i^2)^{1/2}(\sum_i \omega_i^4)^{1/2} = O_p(n^{-1/2}n^{-3/2}) = O_p(n^{-2}),$$

and $\sum_i \omega_i^4 = O_p(n^{-3})$. Also $V^{-1} = O_p(\mu_n)$, since $\mu_n V \to_p \nu_a > 0$. The first term is therefore $O_p(d_n^2\mu_n^{3/2}n^{-2})$ and the second $O_p(d_n^{3/2}\mu_n n^{-3/2})$, which is the stated rate.

The condition. Since $\mu_n \to \infty$, eventually $\mu_n \ge 1$ and $\mu_n^{2/3} \le \mu_n^{3/4}$, so both terms vanish when $\mu_n^{3/4} d_n/n \to 0$, and the conditional $d_{\mathrm{W}} \to_p 0$. For h bounded by one with Lipschitz constant at most one, $|\mathbb{E}[h(V^{-1/2}\omega'U) \mid X, W] - \mathbb{E}h(Z)|$ is at most $\min\{2, d_{\mathrm{W}}\}$, with $Z \sim \mathcal{N}(0,1)$. It therefore tends

to zero in probability while bounded, and dominated convergence gives $\mathbb{E}h(V^{-1/2}\omega' U) \to \mathbb{E}h(Z)$. Convergence for all such h is convergence in distribution. □

Proof of Theorem 8.2

Proof. All moments are conditional on (X, W), given which T, ω, Ω_n, $D_{\mathcal{D}}$ and the LSQR iterates are fixed. As in Section C.1, LSQR is analysed in exact arithmetic. For an (X, W)-measurable $D \in \mathcal{K}_n$ write $\widehat{V}(D) := \omega'\Omega_n\omega + \langle \check{u}\check{u}' - T\Omega_n T', D\rangle_{\mathcal{K}_n}$, so that $\widehat{V}_{\mathcal{D}} = \widehat{V}(D_{\mathcal{D}})$.

(i) Exactness. By Assumption A6, $\check{u} = TU$, so $\mathbb{E}[\check{u}\check{u}' \mid X, W] = T\Sigma_n T'$ and $\mathbb{E}[\widehat{V}(D) \mid X, W] = \omega'\Omega_n\omega + \langle T\Delta T', D\rangle_{\mathcal{K}_n}$. By cyclicity of the trace, and since $\Delta \in \mathcal{D}_n$ and $\Pi_{\mathcal{D}}$ is self-adjoint,

$$\langle T\Delta T', D\rangle_{\mathcal{K}_n} = \tfrac{1}{n^2}\operatorname{tr}(\Delta\, T'DT) = \big\langle \Delta, \Pi_{\mathcal{D}}(T'DT)\big\rangle_{\mathcal{H}_n} = \big\langle \Delta, (\mathcal{A}_{\mathcal{D}}^T)^* D\big\rangle_{\mathcal{H}_n}.$$

Also $\langle \Delta, R_{\mathcal{D}}^T\rangle_{\mathcal{H}_n} = \operatorname{tr}(\Delta\omega\omega') = \omega'\Delta\omega$, so $V = \omega'\Omega_n\omega + \langle \Delta, R_{\mathcal{D}}^T\rangle_{\mathcal{H}_n}$. Subtracting,

$$\mathbb{E}[\widehat{V}(D) \mid X, W] - V = \big\langle \Delta, (\mathcal{A}_{\mathcal{D}}^T)^* D - R_{\mathcal{D}}^T\big\rangle_{\mathcal{H}_n}, \tag{F.1}$$

the bias identity of Proposition 4.5 for the pair $(T, \mathcal{D}_n)$. For $D = D_{\mathcal{D}}$, $(\mathcal{A}_{\mathcal{D}}^T)^* D_{\mathcal{D}}$ is the orthogonal projection of $R_{\mathcal{D}}^T$ onto $\operatorname{Range}((\mathcal{A}_{\mathcal{D}}^T)^*)$, which is $R_{\mathcal{D}}^T$ itself under the range condition. The right side is then zero.

(ii) Consistency. Work on the event of Assumption A9(i), whose probability tends to one. There, by (i),

$$\widehat{V}_{\mathcal{D}} - V = \big\langle T(UU' - \Sigma_n)T', D_{\mathcal{D}}\big\rangle_{\mathcal{K}_n} = \operatorname{tr}\big(\Psi_n(UU' - \Sigma_n)\big) = U'\Psi_n U - \mathbb{E}[U'\Psi_n U \mid X, W],$$

with $\Psi_n := n^{-2}T'D_{\mathcal{D}}T$, which is (X, W)-measurable. Lemma 8.1 gives

$$\mathbb{E}[\mu_n^2(\widehat{V}_{\mathcal{D}} - V)^2 \mid X, W] \le C_0\, m_4\, \mu_n^2 d_n^2 \|\Psi_n\|_F^2.$$

Since $\|AB\|_F \le \|A\|_{\mathrm{op}}\|B\|_F$ and $\|D\|_{\mathcal{K}_n} = n^{-1}\|D\|_F$, we have $\|\Psi_n\|_F \le n^{-2}\|T\|_{\mathrm{op}}^2\|D_{\mathcal{D}}\|_F = n^{-1}\|T\|_{\mathrm{op}}^2\|D_{\mathcal{D}}\|_{\mathcal{K}_n}$. As in Lemma A.5, the pseudo-inverse has norm $s_n^{T,\mathcal{D}}$, so $\|D_{\mathcal{D}}\|_{\mathcal{K}_n} \le s_n^{T,\mathcal{D}}\|R_{\mathcal{D}}^T\|_{\mathcal{H}_n}$, and $\|R_{\mathcal{D}}^T\|_{\mathcal{H}_n} = O_p(\sqrt{d_n/n})$ by Lemma F.1(ii). Combining,

$$\mu_n^2 d_n^2\|\Psi_n\|_F^2 \le \Big(\frac{\mu_n d_n}{n}\Big)^2 \|T\|_{\mathrm{op}}^4 \big(s_n^{T,\mathcal{D}}\big)^2 \|R_{\mathcal{D}}^T\|_{\mathcal{H}_n}^2 = O_p\Big(\big[\|T\|_{\mathrm{op}}^2 s_n^{T,\mathcal{D}}\mu_n (d_n/n)^{3/2}\big]^2\Big) = o_p(1)$$

by Assumption A9(ii). By the conditional Chebyshev inequality, $\Pr(\mu_n|\widehat{V}_{\mathcal{D}} - V| > \varepsilon \mid X, W) \to_p 0$ for every $\varepsilon > 0$, and since these probabilities are bounded, dominated convergence gives $\mu_n(\widehat{V}_{\mathcal{D}} - V) \to_p 0$. Finally $\widehat{V}_{\mathcal{D}}/V - 1 = \mu_n(\widehat{V}_{\mathcal{D}} - V)/(\mu_n V) \to_p 0$, since $\mu_n V \to_p \nu_a > 0$ by Assumption A9(iv).

(iii) Inference. By Assumption A6, $a'\widehat{\beta} - a'\beta = \omega' U$, so $\widehat{V}_{\mathcal{D}}^{-1/2}(a'\widehat{\beta} - a'\beta) = (V/\widehat{V}_{\mathcal{D}})^{1/2}\, V^{-1/2}\omega' U$. The first factor tends to one in probability by (ii), and the second to $\mathcal{N}(0, 1)$ in distribution by Condition 1. Slutsky's theorem completes the proof.

(iv) LSQR. Let $\widehat{D}_n := D_{t_n}$. The iterates and the stopping time are computed from $(\mathcal{A}_{\mathcal{D}}^T, R_{\mathcal{D}}^T)$, which depend on (X, W) alone, so $\widehat{D}_n$ is (X, W)-measurable and (F.1) applies to it. Work on the event $\{R_{\mathcal{D}}^T \in \operatorname{Range}((\mathcal{A}_{\mathcal{D}}^T)^*)\} \cap \{\tau_n s_n^{T,\mathcal{D}} < 1\}$, whose probability tends to one by Assumption A9(i) and $\tau_n s_n^{T,\mathcal{D}} \to_p 0$.

Which criterion stops the iteration. On this event $R_{\mathcal{D}}^T$ and every $(\mathcal{A}_{\mathcal{D}}^T)^* D_t$ lie in $\mathrm{Range}((\mathcal{A}_{\mathcal{D}}^T)^*) = \mathrm{Ker}(\mathcal{A}_{\mathcal{D}}^T)^\perp$, and so does every residual r_t. There $\mathcal{A}_{\mathcal{D}}^T$ is bounded below by $1/s_n^{T,\mathcal{D}}$, so $\psi_t \geq \phi_t / s_n^{T,\mathcal{D}} > \tau_n \phi_t$ whenever $\phi_t > 0$, and the dual criterion never fires. The system is solvable, so $\phi_t \to 0$, and the primal criterion stops the iteration with $\phi_{t_n} \leq \tau_n \|R_{\mathcal{D}}^T\|_{\mathcal{H}_n}$.

Bias. By (F.1) and Cauchy–Schwarz,

$$\mu_n \left| \mathbb{E}[\widehat{V}(\widehat{D}_n) \mid X, W] - V \right| \leq \mu_n \|\Delta\|_{\mathcal{H}_n} \phi_{t_n} \leq \mu_n \tau_n \|\Delta\|_{\mathcal{H}_n} \|R_{\mathcal{D}}^T\|_{\mathcal{H}_n} = O_p(\mu_n \tau_n d_n / n) = o_p(1).$$

Here $\|\Delta\|_{\mathcal{H}_n} \leq \|\Sigma_n\|_{\mathcal{H}_n} + \|\Omega_n\|_{\mathcal{H}_n} = O_p(\sqrt{d_n/n})$ by Lemma F.1(i) and the hypothesis on Ω_n, and $\|R_{\mathcal{D}}^T\|_{\mathcal{H}_n} = O_p(\sqrt{d_n/n})$ by Lemma F.1(ii).

Fluctuation. The argument of (ii) applies with $\widehat{D}_n$ in place of $D_{\mathcal{D}}$, since $\|\widehat{D}_n\|_{\mathcal{K}_n} \leq \|D_{\mathcal{D}}\|_{\mathcal{K}_n}$ by Lemma 5.1(iii). That lemma is stated for $\mathcal{A}_n$, but its proof uses only that LSQR is applied to a linear map between finite-dimensional inner-product spaces, and $D_{\mathcal{D}}$ is the minimum-norm solution.

Adding the two parts, $\mu_n(\widehat{V}(\widehat{D}_n) - V) \to_p 0$, and the ratio and the t-statistic follow as in (ii) and (iii). □

Proof of Corollary 8.3

Proof. Work on the event that Γ_n is nonsingular, whose probability tends to one by Assumption A4.

The objects coincide. For $\mathcal{S}_n = \mathcal{S}_{1W}$, $\mathcal{H}_n^{\mathcal{S}}$ is the space $\mathcal{H}_n$ of block-diagonal symmetric arrays, and $\Pi_{\mathcal{H}}$ keeps the diagonal blocks. With $T = M$ and $\mathcal{D}_n = \mathcal{H}_n$, $\mathcal{A}_{\mathcal{H}}^M(\Delta) = M\Delta M = \mathcal{A}_n(\Delta)$ and $(\mathcal{A}_{\mathcal{H}}^M)^*(D) = \{[MDM]_{gg}\}_g = \mathcal{A}_n^*(D)$. Since $n\omega_g = \widetilde{X}_g \Gamma_n^{-1} a$, $R_{\mathcal{H}}^M = \{n^2 \omega_g \omega_g'\}_g = \{\widetilde{X}_g Q_n \widetilde{X}_g'\}_g = R_n^\star$. Hence $s_n^{M,\mathcal{H}} = s_n$, $D_{\mathcal{D}} = D_n^\star$, the LSQR iterates coincide, and with $\Omega_n = 0$, $\widehat{V}_{\mathcal{D}} = \widehat{V}_{\mathrm{Riesz}}$. Once the hypotheses are verified, parts (ii)–(iv) of Theorem 8.2 are therefore Theorems 6.1 and 6.2.

Assumption A7. $\mathcal{S}_{1W}$ is fixed by the cluster partition, and $\mathbb{E}[u_i \mid X, W] = 0$ by Assumption A1. If I, J are disjoint with $(I \times J) \cap \mathcal{S}_{1W} = \emptyset$, no cluster meets both I and J, so U_I and U_J are functions of disjoint collections of the U_g, which are independent by Assumption A1. By Lyapunov's inequality $\mathbb{E}[u_i^4 \mid X, W] \leq (\mathbb{E}[|u_i|^{4+2\lambda} \mid X, W])^{4/(4+2\lambda)}$, which Assumption A2 bounds by a finite constant m_4 almost surely.

Assumption A6. By (11), $a'\widehat{\beta} - a'\beta = n^{-1} a' \Gamma_n^{-1} \widetilde{X}' U = \omega' U$ with $\omega = \widetilde{X} \Gamma_n^{-1} a / n$, a function of (X, W). The residual is $\breve{u} = MY$, and M annihilates $[X\ W]$.

Assumption A8. By Assumption A1, $\Sigma_n = \mathrm{diag}(\Sigma_g)$ lies in $\mathcal{H}_n$, which is linear; take $\Omega_n = 0$.

Assumption A9. Part (i) is Assumption A5(i). Part (ii) is Assumption A5(ii) with $d_n = \sup_g N_g$, and $\|M\|_{\mathrm{op}} \leq 1$ since M is an orthogonal projection. Part (iv) is the first half of Assumption A3. For part (iii), let $v := \Gamma_n^{-1} a$, so that $\omega_i = \tilde{x}_i' v / n$. By Cauchy–Schwarz,

$$n^3 \sum_i \omega_i^4 = \frac{1}{n} \sum_i (\tilde{x}_i' v)^4 \leq \|v\|^4 \, \frac{1}{n} \sum_i \|\tilde{x}_i\|^4.$$

Here $\|v\| = O_p(1)$ by Assumption A4. Since $\|\tilde{x}_i\|^4 \leq 1 + \|\tilde{x}_i\|^{4+2\lambda}$, Assumption A2 gives

$$\mathbb{E}\Big[\frac{1}{n} \sum_i \|\tilde{x}_i\|^4\Big] \leq 1 + \frac{1}{n} \sum_g N_g C = 1 + C,$$

with C the second supremum there. By Markov's inequality the average is $O_p(1)$, so $n^3 \sum_i \omega_i^4 =$

$O_p(1)$, and Lemma F.1(iii) gives

$$n^3 \sum_{(i,j)\in\mathcal{S}_{1W}} \omega_i^2\omega_j^2 \le d_n\, n^3 \sum_i \omega_i^4 = O_p(d_n).$$

Condition 1 is Step 1 of the proof of Theorem 6.1, which uses only Assumptions A1 to A4.

The conditions of Theorem 8.2(iv). $\|\Omega_n\|_{\mathcal{H}_n} = 0$, and with $d_n = \sup_g N_g$ and $s_n^{M,\mathcal{H}} = s_n$ the tolerance conditions are (35). □

Proof of Proposition 8.5

Proof. Work on the event that Γ_n is nonsingular and $V > 0$, whose probability tends to one by Assumption A4 and (c). We verify the hypotheses of Theorem 8.2.

Assumption A7 is (a). In the sufficient condition, let I, J be disjoint with $(I \times J) \cap \mathcal{S}_{2W} = \emptyset$. Then no $i \in I$ and $j \in J$ share a row or a column cluster, so U_I is a function of $\{\alpha_{g(i)}, \beta_{h(i)}, \varepsilon_i\}_{i\in I}$, U_J of the corresponding collection for J, and the two collections are disjoint. Mutual independence of the components makes U_I and U_J independent.

Assumption A6 holds as in the proof of Corollary 8.3, which used no partition: $a'\widehat{\beta} - a'\beta = \omega' U$ by (11), and M annihilates $[X\ W]$. *Assumption A8* holds by the assumptions on $\mathcal{D}_n$ and Ω_n.

Assumption A9. Parts (i)–(ii) are assumed, with $\|M\|_{\mathrm{op}} \le 1$, and part (iv) is (c). For part (iii), the Cauchy–Schwarz bound in the proof of Corollary 8.3 gives $n^3 \sum_i \omega_i^4 \le \|\Gamma_n^{-1}a\|^4\, n^{-1} \sum_i \|\tilde{x}_i\|^4 = O_p(1)$ by (b), and Lemma F.1(iii) gives

$$n^3 \sum_{(i,j)\in\mathcal{S}_{2W}} \omega_i^2\omega_j^2 \le d_n\, n^3 \sum_i \omega_i^4 = O_p(d_n).$$

Condition 1 follows from Lemma F.2, whose hypotheses hold: Assumption A7 is (a); Assumption A9(iv) and $\mu_n^{3/4} d_n/n \to 0$ are (c); $n^3 \sum_i \omega_i^4 = O_p(1)$ was shown above; and

$$n \sum_i \omega_i^2 = \frac{1}{n}\, a'\Gamma_n^{-1} \widetilde{X}'\widetilde{X} \Gamma_n^{-1} a = a'\Gamma_n^{-1} a = O_p(1)$$

by Assumption A4.

Theorem 8.2 now gives conclusions (i)–(iii), and conclusion (iv) under its additional condition on Ω_n. For the fully robust estimator, Assumption A7 gives $\sigma_{ij} = \mathbb{E}[u_i \mid X, W]\, \mathbb{E}[u_j \mid X, W] = 0$ for $(i,j) \notin \mathcal{S}_{2W}$, so $\Sigma_n \in \mathcal{H}_n^{\mathcal{S}_{2W}} = \Omega_n + \mathcal{D}_n$, and $\|\Omega_n\|_{\mathcal{H}_n} = 0$. □

G Proofs for Section 9

The proofs follow the order of Section 9, with the notation of Section F, with one exception: the proof of Remark 9.3 comes after that of Proposition 9.5, on which it relies. A lemma characterizing the unbiased quadratic forms in the residuals is stated where it is first needed. We begin with the example of information loss referred to in Section 9.1.

Example G.1 (Restricted recovery fails for an oblique projection)**.** Take the location model $Z = \iota_3$ with three singleton clusters and the positive definite weights

$$\Upsilon = \begin{pmatrix} 1 & -\frac{1}{3} & 0 \\ -\frac{1}{3} & 1 & -\frac{2}{3} \\ 0 & -\frac{2}{3} & 1 \end{pmatrix}.$$

Then $Z'\Upsilon Z = 1$, $\omega_\Upsilon = \Upsilon\iota_3 = (\frac{2}{3}, 0, \frac{1}{3})'$ and $T_\Upsilon = I_3 - \iota_3\omega_\Upsilon'$, so

$$T_\Upsilon = \begin{pmatrix} \frac{1}{3} & 0 & -\frac{1}{3} \\ -\frac{2}{3} & 1 & -\frac{1}{3} \\ -\frac{2}{3} & 0 & \frac{2}{3} \end{pmatrix}, \qquad T_\Upsilon \odot T_\Upsilon = \frac{1}{9}\begin{pmatrix} 1 & 0 & 1 \\ 4 & 9 & 1 \\ 4 & 0 & 4 \end{pmatrix}.$$

The restricted map has rank two and annihilates $\Delta = \mathrm{diag}(-1, \frac{1}{3}, 1)$, while $[T_\Upsilon \Delta T_\Upsilon']_{12} = \frac{1}{3} \neq 0$ and the target loads on Δ, $\omega_\Upsilon'\Delta\omega_\Upsilon = -\frac{1}{3}$. No weighting of the squared weighted residuals is therefore unbiased for V_Υ, while the full map $s \mapsto T_\Upsilon \,\mathrm{diag}(s) T_\Upsilon'$ is injective and identifies it. The failure is not generic: for the weights Υ with unit diagonal and off-diagonal entries $(\Upsilon_{12}, \Upsilon_{23}) = (a, b)$, the restricted map is singular exactly on the line $a + b = -1$, and nearby weights leave it identified but ill-conditioned. It belongs to the coordinates, not to the information: $M \odot M$ is nonsingular, so by Propositions 9.1 and 7.1 the squared OLS residuals identify V_Υ.

Proof of Proposition 9.1

Proof. Let $TZ = 0$. Then $T\Pi_Z = 0$, so $T = TM$, and $\check{u} = TY = TU = TMU = T\widehat{U}$. For the ranges, $(\mathcal{A}^T)^*(D) = \Pi_{\mathcal{H}}(T'DT)$, and $T'DT = M(T'DT)M$ because $T' = MT'$. So $T'DT$ lies in $\{MSM : S \in \mathcal{K}_n\}$, and applying $\Pi_{\mathcal{H}}$ gives $\mathrm{Range}((\mathcal{A}^T)^*) \subseteq \mathrm{Range}(\mathcal{A}_n^*)$. If $\mathrm{Ker}(T) = \mathrm{col}(Z)$ then $\mathrm{col}(T') = \mathrm{col}(Z)^\perp = \mathrm{col}(M)$, so T^+T is the orthogonal projector onto $\mathrm{col}(M)$. For any $S \in \mathcal{K}_n$, $D := (T^+)'(MSM)T^+$ gives $T'DT = (T^+T)'MSM(T^+T) = MSM$, so $T'DT$ ranges over all of $\{MSM\}$ and the ranges are equal.

For the weighted fit, $T_\Upsilon Z = Z - Z(Z'\Upsilon Z)^+ Z'\Upsilon Z = 0$, since $(Z'\Upsilon Z)^+ Z'\Upsilon Z$ is the projector onto the row space of Z; this holds whether or not Z has full column rank. Conversely, $T_\Upsilon v = 0$ gives $v = Z(Z'\Upsilon Z)^+ Z'\Upsilon v \in \mathrm{col}(Z)$. Hence $\mathrm{Ker}(T_\Upsilon) = \mathrm{col}(Z)$, and the equality case gives $\mathrm{Range}((\mathcal{A}^{T_\Upsilon})^*) = \mathrm{Range}(\mathcal{A}_n^*)$. The first part gives $T_\Upsilon M = T_\Upsilon$ and $\check{u} = T_\Upsilon\widehat{U}$; also $MT_\Upsilon = M - MZ(Z'\Upsilon Z)^+ Z'\Upsilon = M$ because $MZ = 0$, so $M\check{u} = MT_\Upsilon U = MU = \widehat{U}$.

Finally, a functional is recoverable from the residuals of a fit exactly when its representer lies in the range of that fit's adjoint. The inclusion of ranges therefore gives the recoverability claim, and Proposition 7.1, $\mathrm{Range}((\mathcal{A}_n^{\mathrm{diag}})^*) = \mathrm{Range}(\mathcal{A}_n^*)$, reduces it to the within-cluster moments of the OLS residuals. □

Lemma G.2 (Unbiased quadratic forms)**.** *Let T be a residual map with $TZ = 0$ and $D \in \mathbb{S}^n$. The quadratic form $n^{-2}\check{u}'D\check{u}$ in $\check{u} = TY$ equals $U'KU$ with $K = n^{-2}T'DT$, and $\mathrm{col}(K) \subseteq \mathrm{col}(M)$; conversely, every symmetric K with $\mathrm{col}(K) \subseteq \mathrm{col}(M)$ arises in this way with $T = M$. The form is unbiased for $V_\Upsilon = \mathrm{tr}(\omega_\Upsilon\omega_\Upsilon'\Sigma)$ at every $\Sigma \in \mathcal{H}_n$ if and only if K lies in the affine set*

$$\mathscr{K} := \big\{K \in \mathbb{S}^n : \mathrm{col}(K) \subseteq \mathrm{col}(M),\ \Pi_{\mathcal{H}}K = \Pi_{\mathcal{H}}(\omega_\Upsilon\omega_\Upsilon')\big\},$$

whose direction space is $\mathscr{K}_0 := \{K \in \mathbb{S}^n : \mathrm{col}(K) \subseteq \mathrm{col}(M),\ \Pi_{\mathcal{H}} K = 0\}$. *The set* $\mathscr{K}$ *is nonempty exactly under the range condition of Assumption A5(i) with target* $n^2\{\omega_{\Upsilon,g}\omega'_{\Upsilon,g}\}$.

Proof. Since $TZ = 0$, $\breve{u} = TU$, so $n^{-2}\breve{u}'D\breve{u} = U'KU$; and $T = TM$ (proof of Proposition 9.1) gives $K = n^{-2}MT'DTM$, whence $\mathrm{col}(K) \subseteq \mathrm{col}(M)$. Conversely, if $\mathrm{col}(K) \subseteq \mathrm{col}(M)$ then $K = MKM$, which is the kernel of $D = n^2K$ with $T = M$. For $\Sigma \in \mathcal{H}_n$ the conditional mean is $\mathrm{tr}(K\Sigma) = \langle \Pi_{\mathcal{H}} K, \Sigma\rangle_F$, and $V_\Upsilon = \langle \Pi_{\mathcal{H}}(\omega_\Upsilon\omega'_\Upsilon), \Sigma\rangle_F$. Both $\Pi_{\mathcal{H}} K$ and $\Pi_{\mathcal{H}}(\omega_\Upsilon\omega'_\Upsilon)$ lie in $\mathcal{H}_n$, so the two agree for every $\Sigma \in \mathcal{H}_n$ if and only if $\Pi_{\mathcal{H}} K = \Pi_{\mathcal{H}}(\omega_\Upsilon\omega'_\Upsilon)$. Finally, the symmetric K with $\mathrm{col}(K) \subseteq \mathrm{col}(M)$ are the matrices MSM, $S \in \mathbb{S}^n$, and $\Pi_{\mathcal{H}}(MSM) = \mathcal{A}_n^*(S)$, so $\mathscr{K} \neq \emptyset$ if and only if $\Pi_{\mathcal{H}}(\omega_\Upsilon\omega'_\Upsilon) \in \mathrm{Range}(\mathcal{A}_n^*)$. □

Proof of Proposition 9.2

Proof. The system (50) is the constraint $\Pi_{\mathcal{H}} K = \Pi_{\mathcal{H}}(\omega_\Upsilon\omega'_\Upsilon)$ on the kernel $K = n^{-2}T'_\Upsilon DT_\Upsilon$, so by Lemma G.2 the kernels of its solutions lie in the affine set $\mathscr{K}$. Conversely, every $K \in \mathscr{K}$ satisfies $K = MKM$, so the construction in the proof of Proposition 9.1, with $T = T_\Upsilon$ and $S = n^2K$, gives a D with $n^{-2}T'_\Upsilon DT_\Upsilon = K$.

Suppose (50) is solvable, so that $\mathscr{K} \neq \emptyset$ (Lemma G.2). By Lemma G.2 with $T = M$, the kernels $n^{-2}MDM$ of the solutions of the OLS adjoint system with target $n^2\{\omega_{\Upsilon,g}\omega'_{\Upsilon,g}\}$ are exactly $\mathscr{K}$. By Proposition 4.4, whose argument does not use the form of the target, the minimum-norm element of $\mathscr{K}$ is therefore the kernel of the minimum-norm OLS representer for this target, and by Corollary 7.2 this kernel is $K_M = n^{-2}MH^MM$ with $\{H_g^M\}$ solving the restricted (Gram) system in (51). Its estimate is $U'K_MU = n^{-2}\widehat{U}'H^M\widehat{U}$, and $\widehat{U} = M\breve{u}$ by Proposition 9.1. Unbiasedness under the range condition is Theorem 8.2(i) with $T = M$. Under $U \sim \mathcal{N}(0, \sigma^2 I_n)$, $\mathrm{Var}(U'KU) = 2\sigma^4\|K\|_F^2$, which K_M minimizes over $\mathscr{K}$. □

Proof of Proposition 9.4

Proof. With $L = \Upsilon^{1/2}$, $LT_\Upsilon L^{-1} = I_n - LZ(Z'L^2Z)^+Z'L^2L^{-1} = I_n - LZ\big((LZ)'LZ\big)^+(LZ)' = \widetilde{M}$. Because L is block-diagonal and conformable, $T_{\Upsilon,hg} = L_h^{-1}\widetilde{M}_{hg}L_g$ for every pair of blocks. Given any solution $\{d_h\}$ of the restricted weighted system, set $\widetilde{H}_h := L_h^{-1}d_hL_h^{-1}$, a one-to-one correspondence with inverse $d_h = L_h\widetilde{H}_hL_h$. Then

$$\sum_h T'_{\Upsilon,hg}d_hT_{\Upsilon,hg} = L_g\Big(\sum_h \widetilde{M}_{gh}\widetilde{H}_h\widetilde{M}_{hg}\Big)L_g,$$

using $\widetilde{M}' = \widetilde{M}$, so the constraint is equivalent to $\sum_h \widetilde{M}_{gh}\widetilde{H}_h\widetilde{M}_{hg} = n^2\widetilde{\omega}_g\widetilde{\omega}'_g$ with $\widetilde{\omega} = L^{-1}\omega_\Upsilon$: $\{\widetilde{H}_h\}$ solves the restricted (Gram) system (44) of the whitened model, and every solution of that system arises in this way. The estimate is $n^{-2}\sum_g \breve{u}'_gd_g\breve{u}_g = n^{-2}\sum_g(L_g\breve{u}_g)'\widetilde{H}_g(L_g\breve{u}_g)$ and $L\breve{u} = LT_\Upsilon U = \widetilde{M}LU = \widetilde{u}$, the OLS residual of the whitened model; its OLS estimator is $\widehat{\beta}_\Upsilon$, with $a'\widehat{\beta}_\Upsilon - a'\beta = \omega'_\Upsilon U = \widetilde{\omega}'(LU)$, so $\widetilde{\omega}$ is its influence vector, and its errors LU have the conformable covariance $L\Sigma L \in \mathcal{H}_n$. Since $\widetilde{M}$ is an orthogonal projection, Corollary 7.2 applies to the whitened model: every solution of its restricted system gives the same estimate in every realization, equal to its OLS Riesz estimator.

For the range condition, $\mathcal{A}_{\widetilde{M}}(L\Sigma L) = \widetilde{M}L\Sigma L\widetilde{M} = LT_\Upsilon\Sigma T'_\Upsilon L$, so $\mathrm{Ker}(\mathcal{A}_{\widetilde{M}}) = \{L\Delta L : \Delta \in \mathrm{Ker}(\mathcal{A}^{T_\Upsilon})\}$, and $\mathrm{Ker}(\mathcal{A}^{T_\Upsilon}) = \mathrm{Ker}(\mathcal{A}_n)$: since $T_\Upsilon = T_\Upsilon M$ and $MT_\Upsilon = M$, $T_\Upsilon\Delta T'_\Upsilon = T_\Upsilon(M\Delta M)T'_\Upsilon$

and $M\Delta M = M(T_\Upsilon \Delta T_\Upsilon')M$, so one vanishes if and only if the other does. For $\Delta \in \mathrm{Ker}(\mathcal{A}_n)$, $\langle \widetilde{\omega}\widetilde{\omega}', L\Delta L\rangle = \langle \omega_\Upsilon \omega_\Upsilon', \Delta\rangle$ by cyclicity of the trace, so the whitened target is orthogonal to the whitened kernel if and only if $\{\omega_{\Upsilon,g}\omega_{\Upsilon,g}'\}$ is orthogonal to $\mathrm{Ker}(\mathcal{A}_n)$, which is the range condition of $\mathcal{A}_n$ for that target (Proposition 6.3). □

Proof of Proposition 9.5

Proof. By Lemma G.2, the estimators of the stated form have kernels ranging exactly over the symmetric matrices with $\mathrm{col}(K) \subseteq \mathrm{col}(M)$, and such an estimator is unbiased for V_Υ at every $\Sigma \in \mathcal{H}_n$ if and only if $K \in \mathscr{K}$. Under $U \sim \mathcal{N}(0, \Omega_w)$, $\mathrm{Var}(U'KU) = 2\,\mathrm{tr}(K\Omega_w K\Omega_w) = 2\|\Omega_w^{1/2} K \Omega_w^{1/2}\|_F^2$.

General Ω_w. Let $\mathcal{V} := \{K \in \mathbb{S}^n : MKM = K\}$. The objective is strictly convex on $\mathcal{V}$ and the constraint $\Pi_{\mathcal{H}} K = \Pi_{\mathcal{H}}(\omega_\Upsilon \omega_\Upsilon')$ is affine, so when $\mathscr{K}$ is nonempty the minimizer K^* is unique. It is characterized by $\mathrm{tr}(\Omega_w K^* \Omega_w L) = 0$ for every $L \in \mathcal{V}$ with $\Pi_{\mathcal{H}} L = 0$. For $L \in \mathcal{V}$ and $\Lambda \in \mathcal{H}_n$, $\langle L, M\Lambda M\rangle = \langle L, \Lambda\rangle$. So the orthogonal complement in $\mathcal{V}$ of $\{L \in \mathcal{V} : \Pi_{\mathcal{H}} L = 0\}$ is $\{M\Lambda M : \Lambda \in \mathcal{H}_n\}$, and the condition reads $M\Omega_w K^* \Omega_w M = M\Lambda M$ for some $\Lambda \in \mathcal{H}_n$. Since $K^* = MK^*M$, this is $(M\Omega_w M)K^*(M\Omega_w M) = M\Lambda M$. Multiplying on both sides by $W_\Omega = (M\Omega_w M)^+$, which satisfies $W_\Omega M \Omega_w M = M$ because $M\Omega_w M$ is positive definite on $\mathrm{col}(M)$, gives $K^* = W_\Omega \Lambda W_\Omega$. Conversely, every $K = W_\Omega \Lambda W_\Omega$ with $\Lambda \in \mathcal{H}_n$ that satisfies the constraint lies in $\mathcal{V}$ and satisfies the condition, since $M\Omega_w M W_\Omega = M$; for this convex problem the condition is also sufficient, so $K = K^*$. The estimate is $U'K^*U = \widehat{U}'K^*\widehat{U} = (W_\Omega \widehat{U})'\Lambda(W_\Omega \widehat{U})$, since $W_\Omega = W_\Omega M$.

For $\Omega_w = \Upsilon^{-1}$, $T_\Upsilon' \Upsilon T_\Upsilon = \Upsilon - \Upsilon Z(Z'\Upsilon Z)^+ Z'\Upsilon$ by direct expansion. Its null space is $\mathrm{col}(Z)$, so its column space is $\mathrm{col}(M)$, and $(M\Upsilon^{-1}M)(T_\Upsilon' \Upsilon T_\Upsilon) = M\Upsilon^{-1} T_\Upsilon' \Upsilon T_\Upsilon = M\big(I - Z(Z'\Upsilon Z)^+ Z'\Upsilon\big) = MT_\Upsilon = M$. Two symmetric matrices with column space $\mathrm{col}(M)$ whose product is the projector onto it are pseudo-inverses of each other, so $W_\Omega = T_\Upsilon' \Upsilon T_\Upsilon$.

Conformable Ω_w. Let $L := \Omega_w^{-1/2}$, block-diagonal and conformable, and $\widetilde{K} := L^{-1} K L^{-1}$, so the variance is $2\|\widetilde{K}\|_F^2$. If $v \perp \mathrm{col}(Z)$ then $L^{-1}v \perp \mathrm{col}(LZ)$, since $(LZ)'L^{-1}v = Z'v = 0$, so $L^{-1}\,\mathrm{col}(M) \subseteq \mathrm{col}(\widetilde{M})$, with equality since both have dimension $n - \mathrm{rank}(Z)$; hence $\mathrm{col}(K) \subseteq \mathrm{col}(M)$ if and only if $\mathrm{col}(\widetilde{K}) \subseteq \mathrm{col}(\widetilde{M})$, $\widetilde{M} := I_n - \Pi_{LZ}$. By conformability $\Pi_{\mathcal{H}}(L\widetilde{K}L)$ has blocks $L_g \widetilde{K}_{gg} L_g$, so the unbiasedness constraint is $\widetilde{K}_{gg} = \widetilde{\omega}_g \widetilde{\omega}_g'$ with $\widetilde{\omega} = L^{-1}\omega_\Upsilon$. The problem is therefore to minimize $\|\widetilde{K}\|_F$ subject to $\mathrm{col}(\widetilde{K}) \subseteq \mathrm{col}(\widetilde{M})$ and $\Pi_{\mathcal{H}}\widetilde{K} = \{\widetilde{\omega}_g \widetilde{\omega}_g'\}$, which is the kernel form of the minimum-norm problem of Definition 4.3 for the model (LY, LZ) fitted by OLS with target $\{\widetilde{\omega}_g \widetilde{\omega}_g'\}$; by Proposition 4.4, whose argument does not use the form of the target, its solution is that model's Riesz representer. The resulting estimator is $\widetilde{u}'\widetilde{K}\widetilde{u}$ with $\widetilde{u} = \widetilde{M}LU$, computable from the data as $\widetilde{M}L\check{u}$. For $\Omega_w = I_n$, $L = I_n$ and $\widetilde{M} = M$, giving $\widehat{V}_\Upsilon^M$; for $\Omega_w = \Upsilon^{-1}$, $L = \Upsilon^{1/2}$ and the estimator is $\widehat{V}_\Upsilon^\Upsilon$ by Proposition 9.4. By uniqueness, this is $\widehat{V}_\Upsilon^\Omega$. Under $\Upsilon = \Sigma^{-1}$ the working and true covariances coincide, so the variance minimized is the actual Gaussian variance. □

Proof of Remark 9.3

Proof. Assume the range condition, so that (50) is solvable and $\mathscr{K} \neq \emptyset$ (Lemma G.2). Both K_F and K_M lie in the affine set $\mathscr{K}$ of Lemma G.2: K_F because D_F solves the full weighted system, and K_M by construction. K_M is the minimum-Frobenius-norm element of $\mathscr{K}$ (Proposition 9.2, or Proposition 9.5 with $\Omega_w = I_n$), so it is orthogonal to the direction space $\mathscr{K}_0$, and $K_F - K_M \in \mathscr{K}_0$. Hence $\langle K_M, K_F - K_M\rangle_F = 0$ and $\|K_F\|_F^2 = \|K_M\|_F^2 + \|K_F - K_M\|_F^2$. The difference $U'(K_F - K_M)U$

has conditional mean $\langle \Pi_{\mathcal{H}}(K_F - K_M), \Sigma\rangle_F = 0$ for every $\Sigma \in \mathcal{H}_n$, and under $U \sim \mathcal{N}(0, \sigma^2 I_n)$ its covariance with $U'K_M U$ is $2\sigma^4 \operatorname{tr}(K_M(K_F - K_M)) = 0$.

For the working-covariance reading, the minimum-norm solution D_F lies in the range of the adjoint of $D \mapsto \Pi_{\mathcal{H}}(T_\Upsilon' D T_\Upsilon)$, which is $\{T_\Upsilon H T_\Upsilon' : H \in \mathcal{H}_n\}$. Hence $K_F = n^{-2} NHN$ with $N := T_\Upsilon' T_\Upsilon$. Since $\operatorname{Ker}(T_\Upsilon) = \operatorname{col}(Z)$, $\operatorname{col}(N) = \operatorname{col}(M)$. So $\Omega^* = N^+ + \Pi_Z$ is positive definite, and $M\Omega^* M = N^+$ because $M\Pi_Z = 0$. Therefore $W_{\Omega^*} = N$, and $K_F = W_{\Omega^*} \Lambda W_{\Omega^*}$ with $\Lambda = n^{-2} H \in \mathcal{H}_n$. Since it satisfies the unbiasedness constraint ($K_F \in \mathscr{K}$), the converse in the proof of Proposition 9.5 makes it the Ω^*-optimal kernel. □

Proof of Proposition 9.6

Proof. (i) By Proposition 9.4, Riesz-P(T) is $\widehat{V}_\Upsilon^\Upsilon$, which by Proposition 9.5 has the smallest Gaussian variance under $\Omega_w = \Upsilon^{-1}$ among all unbiased quadratic estimators, including Riesz-F(T) (unbiased, since injectivity of $\mathcal{A}_n$ gives the range condition). When $\Sigma = c\,\Upsilon^{-1}$ the true variance is c^2 times the working variance, so the ranking is unchanged.

(ii) By Propositions 9.4 and 9.5 and Remark 9.3, Riesz-P(T), Riesz-F(T) and Riesz-M are the working-covariance estimators for Υ^{-1}, Ω^* and I_n. Write $K_0 := K_M$. Under $\Sigma = \sigma^2 I_n$, the orthogonal decomposition in the proof of Remark 9.3, which holds for every unbiased kernel K, gives $\operatorname{Var}(U'KU) - \operatorname{Var}(U'K_0U) = 2\sigma^4 \|K - K_0\|_F^2$. It remains to compare the distances of the two kernels from K_0.

The working covariance of Riesz-P(T). For conformable Υ, $M\Upsilon^{-1}M = M - \varepsilon MEM + O(\varepsilon^2)$, which departs from M at first order.

The working covariance of Riesz-F(T). Let $\Delta := T_\Upsilon - M$. Since $MT_\Upsilon = M$ and $T_\Upsilon M = T_\Upsilon$ (Proposition 9.1), $M\Delta = 0$ and $\Delta M = \Delta$. Hence, exactly,

$$T_\Upsilon' T_\Upsilon = (M + \Delta)'(M + \Delta) = M + \Delta'\Delta,$$

with $\Delta = O(\varepsilon)$ and $\Delta'\Delta = M\Delta'\Delta M$. On $\operatorname{col}(M)$, $M + \Delta'\Delta$ therefore acts as $I + \Delta'\Delta$, and a Neumann series gives $M\Omega^* M = (M + \Delta'\Delta)^+ = M - \Delta'\Delta + O(\varepsilon^4)$, which departs from M only at second order.

Distances. The kernel is $K_\Omega = W_\Omega \Lambda W_\Omega$, with $\Lambda \in \mathcal{H}_n$ solving

$$\Pi_{\mathcal{H}}(W_\Omega \Lambda W_\Omega) = \Pi_{\mathcal{H}}(\omega_\Upsilon \omega_\Upsilon').$$

At $\Omega = I_n$ the operator $\Lambda \mapsto \Pi_{\mathcal{H}}(W_\Omega \Lambda W_\Omega)$ is $\mathcal{A}_n^{\mathrm{diag}} = \mathcal{A}_n^* \mathcal{A}_n$ (Proposition 7.1), nonsingular because $\mathcal{A}_n$ is injective, so it stays nonsingular near I_n; and $W_\Omega = (M\Omega M)^+$ is smooth there, since $M\Omega M$ has constant rank. Hence $\Omega \mapsto K_\Omega$ is differentiable at I_n, and it depends on Ω only through $M\Omega M$. The target ω_Υ is common to all three estimators at a given ε, and $\omega_\Upsilon = \omega + O(\varepsilon)$, so evaluating the derivative at the OLS target ω changes it by $O(\varepsilon)$. Hence $\|K_P - K_0\|_F = \varepsilon \|\dot{K}[E]\|_F + O(\varepsilon^2)$ and $\|K_F - K_0\|_F = O(\varepsilon^2)$. Squaring and subtracting gives the result. □

Corollary G.3 (Asymptotics with known weights)**.** *Suppose Υ is conformable with eigenvalues in $[\underline{\lambda}, \overline{\lambda}] \subset (0, \infty)$ with probability approaching one, and Assumptions A1 to A4 hold with $\widetilde{X}_\Upsilon$, $\Gamma_\Upsilon := n^{-1}\widetilde{X}_\Upsilon' \Upsilon \widetilde{X}_\Upsilon$ and V_Υ in place of $\widetilde{X}$, Γ_n and V. If the range and stability conditions of Assumption A5*

hold for the OLS operator with target $n^2\{\omega_{\Upsilon,g}\omega'_{\Upsilon,g}\}$, then $\widehat{V}^M_\Upsilon$ satisfies the conclusions of Theorem 8.2; if they hold for the whitened operator $\mathcal{A}_{\widetilde{M}}$ with target $n^2\{\widetilde{\omega}_g\widetilde{\omega}'_g\}$ and the rows of $\Upsilon_g^{1/2}$ have bounded ℓ_1 norm, $c_\Upsilon := \max_{g,i}\sum_j|(\Upsilon_g^{1/2})_{ij}| = O_p(1)$, so does $\widehat{V}^\Upsilon_\Upsilon$; and if they hold for $\mathcal{A}^{T_\Upsilon}$ with the same target, so does Riesz-F(T). By Propositions 9.1 and 9.4 the range conditions of the three are the same.

Proof of Corollary G.3

Proof. For $\widehat{V}^M_\Upsilon$ take $T = M$, $\omega = \omega_\Upsilon$, $\mathcal{D}_n = \mathcal{H}_n$, $\Omega_n = 0$ in Theorem 8.2, with $\mathcal{S}_n = \mathcal{S}_{1W}$ and $d_n = \sup_g N_g$. Assumptions A6 to A8 hold as in the proof of Corollary 8.3, with Assumption A6 because $a'\widehat{\beta}_\Upsilon - a'\beta = \omega'_\Upsilon U$ and $MZ = 0$, and $\|M\|_{\mathrm{op}} = 1$. For Assumption A9(iii), $\omega_\Upsilon = n^{-1}\Upsilon\widetilde{X}_\Upsilon\Gamma_\Upsilon^{-1}a$ mixes observations within a cluster, so we bound it blockwise. On $\mathcal{S}_{1W}$, $\sum_{(i,j)\in\mathcal{S}_n}\omega_i^2\omega_j^2 = \sum_g\|\omega_{\Upsilon,g}\|^4$. With $v := \Gamma_\Upsilon^{-1}a$, $\|\omega_{\Upsilon,g}\|^2 \le n^{-2}\overline{\lambda}^2\|v\|^2\|\widetilde{X}'_{\Upsilon,g}\widetilde{X}_{\Upsilon,g}\|_F$, so by Lemma A.2(i) with $\theta = 2$, $n^3\sum_g\|\omega_{\Upsilon,g}\|^4 = O_p(n^{-1}\sum_g N_g^2) = O_p(\sup_g N_g)$. Condition 1 is the Lyapunov argument of Step 1 of the proof of Theorem 6.1 with ω_Υ in place of ω, using $\|\omega_{\Upsilon,g}\| \le n^{-1}\overline{\lambda}\,\|v\|\,\|\widetilde{X}_{\Upsilon,g}\|$. Parts (i), (ii) and (iv) of Assumption A9 are the maintained range, stability and scale conditions.

For Riesz-F(T) take $T = T_\Upsilon$ in the same argument. Only the bound on the residual map changes: $T_\Upsilon = \Upsilon^{-1/2}\big(I_n - \Pi_{\Upsilon^{1/2}Z}\big)\Upsilon^{1/2}$, so $\|T_\Upsilon\|_{\mathrm{op}} \le (\overline{\lambda}/\underline{\lambda})^{1/2}$, which is Assumption A9(ii)'s requirement.

For $\widehat{V}^\Upsilon_\Upsilon$ apply the same argument to the whitened model: its residual maker $\widetilde{M}$ is an orthogonal projection, and its errors LU satisfy Assumption A1 because L is block diagonal, and Assumption A2 with the error-moment bound multiplied by $c_\Upsilon^{4+2\lambda}$: each whitened error $(L_gU_g)_i = \sum_j(L_g)_{ij}u_{gj}$ satisfies, by Minkowski's inequality, $\|(L_gU_g)_i\|_{4+2\lambda} \le \sum_j|(L_g)_{ij}|\,\|u_{gj}\|_{4+2\lambda} \le c_\Upsilon\max_j\|u_{gj}\|_{4+2\lambda}$. The same step bounds the design moment of Assumption A2 for the whitened design $L_g\widetilde{X}_{\Upsilon,g}$: with $p = 4+2\lambda$, Minkowski's and Jensen's inequalities give $\mathbb{E}\|(L_g\widetilde{X}_{\Upsilon,g})_i\|^p \le c_\Upsilon^{p-1}\sum_j|(L_g)_{ij}|\,\mathbb{E}\|\tilde{x}_{\Upsilon,gj}\|^p$, and averaging over i, with the column sums of $|L_g|$ also at most c_Υ by symmetry, multiplies the bound by c_Υ^p. Assumption A4 is unchanged, since the whitened design has Gram matrix $n^{-1}\widetilde{X}'_\Upsilon L^2\widetilde{X}_\Upsilon = \Gamma_\Upsilon$. The condition on c_Υ holds for diagonal weights and for random-effects weights, for which $\Upsilon_g^{1/2} = aI + bJ$ with $b = O(N_g^{-1})$. Its target weights satisfy $\|\widetilde{\omega}_g\| \le \underline{\lambda}^{-1/2}\|\omega_{\Upsilon,g}\|$, and its variance is $\widetilde{\omega}'L\Sigma L\widetilde{\omega} = V_\Upsilon$, so Assumption A3 is unchanged. By Proposition 9.4 the estimate equals $\widehat{V}^\Upsilon_\Upsilon$ in every realization, and by its proof the two range conditions coincide with each other and with that of $\mathcal{A}_n$. □

H Supplementary Simulation Design and Results

This appendix completes Section 10 exhibit by exhibit. Section H.1 lists the estimators, the parameters, the numbers of designs and draws, and how the estimators are computed. Sections H.2 to H.6 follow Exhibits 1 to 5, giving for each the design details not in the main text and the supplementary results. Section H.7 reports the results that belong to no single exhibit: the conditioning of the restricted system, the accuracy and wall time of the computational routes, and a misspecified partition. Unless a note says otherwise, the conventions of Section 10.1 apply to every table and figure.

H.1 Estimators, parameters and computation

Table 14 lists the estimators compared; their critical values follow the suffix convention of Section 10.1.

Table 14: Estimators

Label	Exhibit	Weights	Requires
CR0, CR1	1, 2, 5	$H_g = \widetilde{X}_g Q \widetilde{X}_g'$, with and without the df factor	—
CR2, CR3	1, 2, 5	$M_{gg}^{-1/2}$, M_{gg}^{-1} conjugation	M_{gg} nonsingular
CR2-BM	1, 2, 5	CR2 with Satterthwaite t	M_{gg} nonsingular
CR2$^+$, CR3$^+$	2	Moore–Penrose powers of M_{gg} (Pustejovsky and Tipton, 2018)	—
HC0–HC3	1, 2	the $N_g = 1$ analogues	$h_{ii} < 1$ (HC2, HC3)
HCK	1, 2	exact inverse of $M_W \odot M_W$ (Cattaneo et al., 2018), ignoring the partition and regressor of interest X	$M_W \odot M_W$ nonsingular
LCO	1, 2	(H.1), from Anatolyev and Ng (2026)	M_{gg} nonsingular
CGM, Thompson	3	CR0 on each dimension, minus CR0 on their intersection (CGM; Cameron et al., 2011) or minus HC0 (Thompson, 2011)	—
CGM1	3	CGM with the factor $\frac{G_k}{G_k-1}\frac{n-1}{n-p}$ on each component, G_{12} the nonempty intersections	—
Riesz	1–3, 5	$\widehat{V}_{\text{Riesz}}$, Algorithm 1; on $\mathcal{H}_n^{\mathcal{S}_{2W}}$ in 3	—
Riesz-t_ν	1–5	the same with the t_ν critical value of (31)	—
Riesz[$\mathcal{D}^{\text{het}+\rho}$]	1, 3, 5	(48) on the class of Section 8.3; two levels in 3	—
Riesz-M, -Υ	4	(51); Proposition 9.4	—
WLS-CR0	4	the weighted cluster sandwich	—
Oracle	all	the exact conditional V	—

The leave-cluster-out estimator is the one estimator compared that is not a quadratic form in the residuals. With $\widehat{U}_g^{(-g)} = M_{gg}^{-1}\widehat{U}_g$, Anatolyev and Ng (2026, Definition 2.1) replace $\widehat{U}_g\widehat{U}_g'$ by $\operatorname{sym}\big(y_g(\widehat{U}_g^{(-g)})'\big)$, which is unbiased for Σ_g because the leave-cluster-out fit is unbiased for the regression mean. With the influence weights $z_g = \widetilde{X}_g\Gamma^{-1}a$ of the contrast, the symmetrization is vacuous inside the quadratic form and

$$\widehat{V}_{\text{LCO}} = \frac{1}{n^2}\sum_{g=1}^{G}\big(z_g'y_g\big)\big(z_g'M_{gg}^{-1}\widehat{U}_g\big); \tag{H.1}$$

with a single regressor of interest, as in the simulations, $z_g = n\widetilde{X}_g/(\widetilde{X}'\widetilde{X})$. It escapes the spillover of (7) rather than correcting it, at the cost of a sampling variance depending on the regression mean and of requiring every M_{gg} to be invertible.

H.1.1 Designs and draws

Table 15 explains the design parameters and Table 16 lists, for every experiment, the number of configurations of the design parameters, the number S of designs drawn for each, and the number R of error draws per design (Section 10.1).

Table 15: Parameters

Parameter	Symbol	Meaning
Sample size, clusters	(n, G)	Number of observations and of error clusters; $G = n$ is the scalar case, and G_1, G_2 count the clusters on each dimension with two-way clustering
Cluster-size imbalance	η_N	Cluster sizes $N_g \propto \exp(\eta_N g/G)$; $\eta_N = 0$ gives equal sizes
Control ratio	ℓ/n	Number of controls in W relative to n
Cluster component of controls	ρ_w	Share of each control's variance due to a cluster-level component
Cluster component of regressor	ρ_x	Share of the regressor's variance due to a cluster-level component
Common factor	φ	Weight of a factor with cluster-specific loadings shared by the regressor and the controls
Heteroskedasticity	ϑ	$\vartheta = 1$ makes the variance of the idiosyncratic error a function of x and w; $\vartheta = 0$ makes it constant
Within-cluster error correlation	ρ_u	Share of the error variance due to a cluster shock, so that the errors within a cluster are equicorrelated
Two-way error components	ρ_1, ρ_2	Shares of the error variance due to a shock on each clustering dimension
Absorbed share	π	Share of clusters whose indicators are added to W
Absorption defect	τ	Size of a perturbation of those indicators; $\tau = 0$ absorbs the clusters exactly

Table 16: Designs and draws behind each exhibit

	Exhibit	Design	Configurations	Designs (S)	Draws (R)
A	1, 5	Scalar	14	100	1,000
B	1, 5	One-way clusters, $\rho_u = 0.3$	28	50	1,000
B	1	One-way clusters, $\rho_u = 0.6$	28	50	1,000
B	1	Scalar, clustered regressors ($\rho_u = 0$)	14	50	1,000
B	1	One-way clusters, $G = 50$, $n = 1000$	7	50	2,000
B	1	One-way clusters, $G = 10$, $n = 1000$	7	50	2,000
B	1, 5	One-way clusters, $G = 10$, $n = 100$	28	50	2,000
C	2	Fixed effects†	200	1	2,000
C	2	Rare indicators	42	10	2,000
E	3	Two-way clusters, few controls†	320	1	2,000
E	3	Two-way clusters, counting bound†	360	1	2,000
F	4	Weighted fits†	150	1	2,000
G	4	Weighted fits, clusters†	200	1	2,000
H	4	Estimated weights†	90	1	1,000
I	1	Deviation classes†	240	1	2,000
K	4	Full and restricted weighted systems	504	10	5,000
L	5	Cluster sizes, treated clusters	25	20	2,000
N	5	Dyadic designs	12	20	2,000

Configurations: cells of the design grid. Designs (S): independent draws of (X, W, Σ) per configuration. Draws (R): error vectors per design. †The design parameters are drawn afresh for every design, so each design is its own configuration. The non-Gaussian runs of Section H.2.5 repeat A with $S = 100$ and B with $S = 50$ ($R = 1,000$), B at $G = 10$ with clustered regressors only, $S = 50$ and $R = 2,000$, under each law. Experiment D (Section H.7.3) uses 50 designs per value of $\varkappa$ and $R = 5,000$, and the conditioning results (Section H.7.1) the grid described there; the replication of Cattaneo et al. (2018) follows their scheme, 5,000 designs with one draw each.

H.1.2 Implementation

LSQR is run to relative tolerance 10^{-10} with iteration caps of $20\max(G, 100)$ for Full-LSQR, 20,000 for Restricted-LSQR and 10,000 in Experiment E, and we report the attained residual, the iteration count, which criterion of (28) terminated the run, and the fraction of designs reaching the cap. Before the experiments we verify the identities the design relies on: $\mathcal{A}_n^{\mathrm{diag}} = \mathcal{A}_n^* \mathcal{A}_n$ and $\lambda_j(\mathcal{A}_n^{\mathrm{diag}}) = \sigma_j(\mathcal{A}_n)^2$ to 10^{-14}; $\langle \mathcal{A}_n(\Sigma), D_n^\star \rangle = V$ to 10^{-14} when $\delta_n = 0$; the restricted and full systems giving equal estimates in every realization; $\mathcal{A}_n^{\mathrm{diag}} = M \odot M$ when $N_g = 1$; cluster fixed effects rendering every M_{gg} singular while $\delta_n / \|R_n^\star\|$ is at machine precision (of order 10^{-15}); and Example G.1 reproduced exactly. The first of these is what makes the conditioning comparison interpretable: with the two systems verified to share a kernel, a gap between variants cannot reflect a difference in what is identified.

Designs violating the range condition. The designs do not guarantee Assumption A5(i), which is a property of (X, W) and the contrast; the stopping rule checks it, since primal termination certifies $\delta_n \le \tau_n \|R_n^\star\|_{\mathcal{H}_n}$ (Section 7.5). In Experiments A and B the full-system solver terminated on the primal criterion in every design in which it converged, so $\delta_n / \|R_n^\star\|_{\mathcal{H}_n} \le 10^{-10}$ there; the runs that reached the iteration cap are high-$\kappa(\mathcal{A}_n)$ designs of Section H.7.1. The designs on which the range condition can fail are the designs of the identification part of Experiment E, which Exhibit 3 reports by side of the counting bound.

H.2 Exhibit 1: coverage as ℓ/n grows

The design of Exhibit 1 is given in full in Section 10.1. This subsection reports the spread over designs behind the medians of the main text, reruns the design of Cattaneo et al. (2018) with their replication scheme, reports coverage with few clusters, also with $n = 100$, evaluates the restricted classes on covariances outside them, and repeats Exhibit 1 under non-Gaussian errors.

H.2.1 Spread over designs

Table 17 gives, for the heteroskedastic designs of the main grid, the 10th and 90th percentiles over designs of the exact bias and of the standard deviation of $\widehat{V}/V$, around the medians reported in Section 10. The Riesz estimator is exactly unbiased in every design, so its bias has no spread. The classical estimators do have spread, and it is asymmetric: HC3 and CR3 have long right tails, CR0 and HC0 long left tails, and with clustered regressors CR1 and HCK understate the variance by a wide margin in a tenth of the designs. The medians are therefore representative for the Riesz estimator and favourable to the classical ones. Standard deviations vary much less across designs than biases do.

H.2.2 Replication of Cattaneo et al. (2018)

Table 18 reruns Model 1 of Cattaneo et al. (2018) with their design and their replication scheme: $n = 700$, $x_i \sim \mathcal{N}(0,1)$, K rare indicators $w_i = \mathbf{1}(v_i \ge 2.5)$ with $v_i \sim \mathcal{N}(0, I_K)$, homoskedastic errors or $\mathrm{Var}[u \mid x, w] = \kappa_u(1 + (t(x) + \iota' w)^2)$ and $\mathrm{Var}[x \mid w] = \kappa_v(1 + (\iota' w)^2)$ with population normalizations, and a new (X, W) in each of 5,000 replications per configuration, so that coverage is

Table 17: Exhibit 1: spread over designs of the bias and standard deviation of $\widehat{V}/V$

	Bias			Standard deviation		
	P10	Median	P90	P10	Median	P90
Scalar ($G = n$)						
HC0	-57.2%	-26.8%	-2.3%	0.06	0.11	0.15
HC1	-15.0%	-8.4%	-1.3%	0.11	0.13	0.16
HC2	-14.6%	-7.9%	-0.7%	0.11	0.13	0.16
HC3	1.1%	15.8%	69.7%	0.15	0.18	0.22
HCK	-1.4%	-0.9%	-0.6%	0.14	0.16	0.19
Riesz	0.0%	0.0%	0.0%	0.15	0.17	0.19
Clusters, errors only ($\rho_w = \rho_x = 0$, $\rho_u = 0.3$)						
CR0	-55.3%	-26.0%	-2.7%	0.07	0.12	0.16
CR1	-11.3%	-6.9%	-1.2%	0.13	0.15	0.17
CR2-BM	-10.7%	-6.3%	-0.5%	0.14	0.15	0.17
CR3	1.7%	18.6%	79.9%	0.17	0.20	0.27
HCK	-3.3%	-0.7%	1.9%	0.13	0.15	0.17
LCO	-1.0%	-0.1%	1.0%	0.20	0.23	0.28
Riesz	0.0%	0.0%	0.0%	0.16	0.19	0.24
Clusters, errors and regressors ($\rho_w = \rho_x = 0.6$, $\rho_u = 0.3$)						
CR0	-55.5%	-43.8%	-5.9%	0.07	0.09	0.17
CR1	-29.5%	-17.7%	-4.4%	0.11	0.13	0.17
CR2-BM	-18.2%	-13.8%	-1.3%	0.13	0.14	0.18
CR3	3.5%	42.5%	107.6%	0.19	0.25	0.33
HCK	-35.9%	-24.7%	-9.6%	0.07	0.10	0.16
LCO	-1.1%	0.0%	1.1%	0.23	0.27	0.33
Riesz	0.0%	0.0%	0.0%	0.19	0.23	0.28

Heteroskedastic designs ($\vartheta = 1$) of the main grid, all ℓ/n pooled. P10 and P90: 10th and 90th percentiles over designs; the median is the value in table 3. Riesz-t_ν uses the same variance estimator as Riesz and is omitted.

unconditional as in their table. The estimators are theirs as published, with M the residual maker of W alone, and ours, the minimum-norm solution of $(M_Z \odot M_Z)h = \omega \odot \omega$ with M_Z the residual maker of $[x\ W]$. The coverage of their eight estimators differs from their published values by a median of 0.006 and at most 0.091. Two features of the design matter for reading their table. First, an indicator active for exactly two units gives both $M_{ii} = 1/2$ and two identical rows of $M \odot M$, so the matrix HCK inverts is singular in at least 100% of the replications with $K > 1$; we use the Moore–Penrose inverse, which is what reproduces their HCK column. Second, an indicator active for a single unit absorbs it, and such units are dropped, since they contribute nothing to $\widehat{\beta}$ or to any estimator. In this design the Riesz estimator covers between 0.942 and 0.952, and HCK between 0.939 and 0.952: in the scalar case, where HCK applies, the two coincide up to the leverage of x.

Table 18: Replication of Cattaneo et al. (2018), Model 1: coverage of nominal 95% intervals

K/n	HO1	HC2	HC3	HCK	Riesz	Oracle	Singular
Homoskedastic							
0.001	0.952 (0.950)	0.951 (0.948)	0.951 (0.948)	0.950 (0.948)	0.951	0.953	0.12
0.101	0.949 (0.956)	0.946 (0.952)	0.957 (0.962)	0.946 (0.951)	0.947	0.947	1.00
0.201	0.953 (0.947)	0.952 (0.946)	0.970 (0.968)	0.951 (0.945)	0.951	0.953	1.00
0.301	0.953 (0.950)	0.952 (0.951)	0.981 (0.977)	0.952 (0.949)	0.952	0.954	1.00
0.401	0.951 (0.954)	0.949 (0.952)	0.987 (0.989)	0.947 (0.949)	0.949	0.954	1.00
Heteroskedastic							
0.001	0.855 (0.880)	0.947 (0.945)	0.947 (0.945)	0.947 (0.945)	0.947	0.948	0.11
0.101	0.841 (0.750)	0.935 (0.926)	0.963 (0.957)	0.940 (0.948)	0.942	0.947	1.00
0.201	0.839 (0.804)	0.936 (0.924)	0.981 (0.973)	0.947 (0.945)	0.949	0.955	1.00
0.301	0.843 (0.856)	0.927 (0.926)	0.987 (0.981)	0.943 (0.947)	0.944	0.950	1.00
0.401	0.848 (0.875)	0.926 (0.929)	0.992 (0.990)	0.939 (0.948)	0.942	0.952	1.00

Our coverage, with the value published in Cattaneo et al. (2018, Table 1(a), Gaussian approximation) in parentheses; $5,000$ replications per row, each drawing a new (X, W). HO1: homoskedastic with degrees-of-freedom correction. Singular: share of replications in which $M \odot M$ is singular.

H.2.3 Few clusters

Table 19 holds $n = 1000$ in the setting with clustered regressors and varies only the number of clusters; table 4 summarizes it. The first row of each panel, $\ell/n = 0.01$, has ten controls, as in Exhibit 3, and isolates the effect of few clusters. From two hundred to ten clusters, CR1-t_{G-1} falls from 0.944 to 0.788 and CR2-BM to 0.905, while Riesz-t_ν and the class fall only to 0.934 and 0.941. Few clusters raise the classical bias because regressors with a cluster component count against G rather than n: CR1 is biased by -4.3% with two hundred clusters, -18.2% with fifty and -63.1% with ten. The other rows add controls. With ten clusters the bias of CR1 then falls, because once ℓ exceeds G the controls' cluster components absorb the regressor's, whereas the fully robust estimator degrades: it is negative in 6.0% of draws, and with t_ν critical values its worst design covers 0.750.

With $n = 100$, table 20 repeats the ten-cluster panel, clusters of about ten observations, which also cuts the information within each cluster. The ordering is the same: the fully robust estimator, with t_ν critical values, covers 0.873, with median degrees of freedom 1.9 and $\widehat{V} \leq 0$ in 10.3% of draws, while the class estimator covers 0.951, CR2-BM 0.936 and CR1 with the t_{G-1} critical value 0.902.

Table 19: Exhibit 1: coverage with many and few clusters, $n = 1000$

ℓ/n	CR1-t_{G-1}	CR3-t_{G-1}	CR2-BM	Riesz	Riesz-t_{G-1}	Riesz-t_ν	Class-t_ν
$G = 200$							
0.01	0.944	0.953	0.950	0.948	0.949	0.951	0.952
0.1	0.912	0.968	0.937	0.945	0.946	0.950	0.950
0.3	0.904	0.986	0.923	0.942	0.944	0.949	0.950
0.5	0.935	0.996	0.932	0.942	0.943	0.952	0.953
All	0.920	0.976	0.934	0.944	0.945	0.950	0.951
$G = 50$							
0.01	0.919	0.956	0.941	0.935	0.942	0.948	0.949
0.1	0.861	0.975	0.906	0.935	0.941	0.950	0.949
0.3	0.931	0.982	0.932	0.933	0.939	0.951	0.950
0.5	0.937	0.993	0.936	0.915	0.921	0.948	0.950
All	0.906	0.977	0.926	0.931	0.937	0.950	0.949
$G = 10$							
0.01	0.788	0.976	0.905	0.855	0.889	0.934	0.941
0.1	0.921	0.968	0.931	0.886	0.920	0.946	0.949
0.3	0.936	0.979	0.942	0.837	0.869	0.919	0.949
0.5	0.935	0.992	0.940	0.711	0.732	0.795	0.949
All	0.901	0.978	0.928	0.834	0.864	0.909	0.948

Clustered regressors, $n = 1000$, $\rho_u = 0.3$, $\rho_w = \rho_x = 0.6$, $\vartheta = 1$; $S = 50$ designs per configuration and $R = 2,000$ draws at $G = 50$ and 10, and as in Exhibit 1 at $G = 200$. Riesz: the fully robust estimator. Class-t_ν: the class $\mathcal{D}^{\text{het}+\rho}$ of Section 8.3.

Table 20: Exhibit 1 with ten clusters and $n = 100$

	$G = 10$, $n = 100$						
ℓ/n	CR1-t_{G-1}	CR3-t_{G-1}	CR2-BM	Riesz	Riesz-t_{G-1}	Riesz-t_ν	Class-t_ν
0.01	0.905	0.949	0.945	0.864	0.900	0.944	0.950
0.1	0.879	0.969	0.931	0.834	0.868	0.921	0.950
0.3	0.905	0.983	0.929	0.764	0.792	0.862	0.950
0.5	0.925	0.993	0.939	0.664	0.683	0.735	0.951
All	0.902	0.974	0.936	0.785	0.814	0.873	0.951

Clustered regressors, $\rho_w = \rho_x = 0.6$, $\rho_u = 0.3$, $\vartheta = 1$; $S = 50$ designs per configuration and $R = 2,000$ draws. Columns as in table 19.

H.2.4 Deviation classes

Design (Experiment I). The target is the OLS variance, with (n, G) equal to $(150, 30)$ or $(300, 60)$ and errors, heteroskedastic in half the designs, whose within-cluster correlation takes one of three forms: a common equicorrelation $\rho_u \in \{0.3, 0.6\}$, which lies in $\mathcal{D}^{\text{het}+\rho}$; cluster-specific equicorrelations $\rho_g \sim \text{U}[\rho_u - 0.3, \rho_u + 0.3]$, which lie in $\mathcal{D}^{\text{het}+\rho_g}$ but not in $\mathcal{D}^{\text{het}+\rho}$; and a within-cluster AR(1) with parameter ρ_u, which lies in none of the restricted classes. ρ_u and $\vartheta \in \{0, 1\}$ are balanced across the three structures, and the other parameters are drawn for each design: $\rho_w, \rho_x \in \{0.3, 0.6\}$, $\ell/n \in \{0.05, 0.2, 0.4\}$, $\eta_N \in \{0, 1\}$, and normal or uniform controls. We compare the class estimators of Section 8.3, the class $\mathcal{D}^{\text{het}}$ around a plug-in equicorrelation, a model-based random-effects variance, which assumes $\Sigma = \sigma^2 I_n + \gamma(J - I_n)$ and plugs in the residual variance $\hat{\sigma}^2 = \widehat{U}'\widehat{U}/(n-p)$ and the average within-cluster residual cross-product $\hat{\gamma}$, CR1, CR2 and CR2-BM, and the fully robust estimator; each Riesz estimator is reported with and without the degrees-of-freedom correction of Section 5.2, and exact Gaussian moments are computed for every estimator with fixed weights.

Table 21 reports Experiment I, and it separates the two ingredients the remaining coverage gap depends on. Once the bias is removed, what is left is the tail of the studentized statistic, and at these cluster counts ($G = 30$ and 60) the degrees-of-freedom correction closes it: the Riesz estimators with t_ν critical values cover 0.944 to 0.952, against 0.935 to 0.937 for CR2-BM. The lowest, 0.944, is the class $\mathcal{D}^{\text{het}+\rho}$ under AR(1) errors, outside its class. The fully robust estimator pays for its exactness with intervals 7% to 8% longer than the oracle's; the class $\mathcal{D}^{\text{het}+\rho}$ pays 1% to 3%, because its representer is smaller (Remark 8.4) and its degrees of freedom larger (median 37, against 13 for the fully robust estimator and 24 for CR2-BM). It is also the only estimator here that is unbiased on its class and has a smaller relative root mean squared error than CR2, 0.26 to 0.27 against 0.33 to 0.34. Outside its class its median exact bias is small, 0.5% under cluster-specific correlation and -3.8% under AR(1), but the worst design reaches -12% and -21%: exactness is bought class by class. Under AR(1) the bias comes from the heteroskedastic designs: its median is -0.3% with homoskedastic errors and -7.1% with heteroskedastic ones. Estimating the correlation level outside the class and plugging it in ($\mathcal{D}^{\text{het}}$) brings back a bias of -13% to -16%, and the model-based variance, which ignores heteroskedasticity, is biased by -26% to -32%.

Table 21: Experiment I: deviation classes and degrees-of-freedom correction

	Coverage by true correlation			Length	Rel. RMSE
Estimator	common	cluster-specific	AR(1)	vs. oracle	of $\widehat{V}$
CR2	0.923	0.920	0.923	-7 to -5%	0.33–0.34
CR2-BM	0.937	0.935	0.937	-2 to 0%	0.33–0.34
Model-based	0.901	0.896	0.888	-18 to -15%	0.31–0.34
Riesz[$\mathcal{D}^{\text{het}}$], plug-in	0.918	0.913	0.917	-9 to -7%	0.27–0.30
Riesz[$\mathcal{D}^{\text{het}+\rho}$]	0.942	0.943	0.934	-3 to -1%	0.26–0.27
Riesz[$\mathcal{D}^{\text{het}+\rho}$]-$t_\nu$	0.952	0.952	0.944	$+1$–3%	0.26–0.27
Riesz[$\mathcal{D}^{\text{het}+\rho_g}$]-$t_\nu$	0.950	0.949	0.946	$+3$–4%	0.36–0.39
Riesz (fully robust)	0.921	0.921	0.919	-2%	0.43–0.45
Riesz-t_ν (fully robust)	0.948	0.947	0.946	$+7$–8%	0.43–0.45

80 designs per correlation structure, $(n, G) \in \{(150, 30), (300, 60)\}$, $R = 2{,}000$ draws. Length and relative RMSE: ranges across the three structures. The common equicorrelation lies in $\mathcal{D}^{\text{het}+\rho}$, the cluster-specific one in $\mathcal{D}^{\text{het}+\rho_g}$ only, and the AR(1) in no restricted class. The oracle covers 0.950.

H.2.5 Non-Gaussian errors

This section runs non-Gaussian versions of the main designs by replacing every component — the cluster shock and the idiosyncratic term — by a standardized t_5 draw (heavy tails, excess kurtosis 6) or a standardized χ^2_3 draw (skewness 1.63). The components keep mean zero and unit variance, so Σ, the target V and the exact mean bias of every quadratic-form estimator are unchanged: the Riesz estimator remains exactly unbiased under every law. What changes is the distribution of $\widehat{\beta}$ given the design, which is no longer exactly normal, so the oracle's departures from 0.95 now measure the normal approximation, and the variability of $\widehat{V}$, which the Satterthwaite degrees of freedom of (31), derived under Gaussian errors, do not account for.

Table 22 reports the two settings in which the degrees-of-freedom correction matters: clustered regressors with $G = 200$, and the same setting with ten clusters and $n = 100$ (Section H.2.3). None of the conclusions depends on normality. With clustered regressors and $G = 200$, the oracle covers 0.950 under t_5 errors and 0.950 under χ^2_3 errors, and Riesz-t_ν covers 0.950 and 0.950, against 0.950 with Gaussian errors; CR2-BM covers 0.935 and 0.934. With ten clusters the fully robust estimator is as fragile as under normality (Riesz-t_ν 0.870 and 0.871), and the class estimator, with t_ν critical values, remains at 0.952 and 0.951. $\widehat{\beta}$ and $\widehat{V}$ average many error components, so the tails of each matter little.

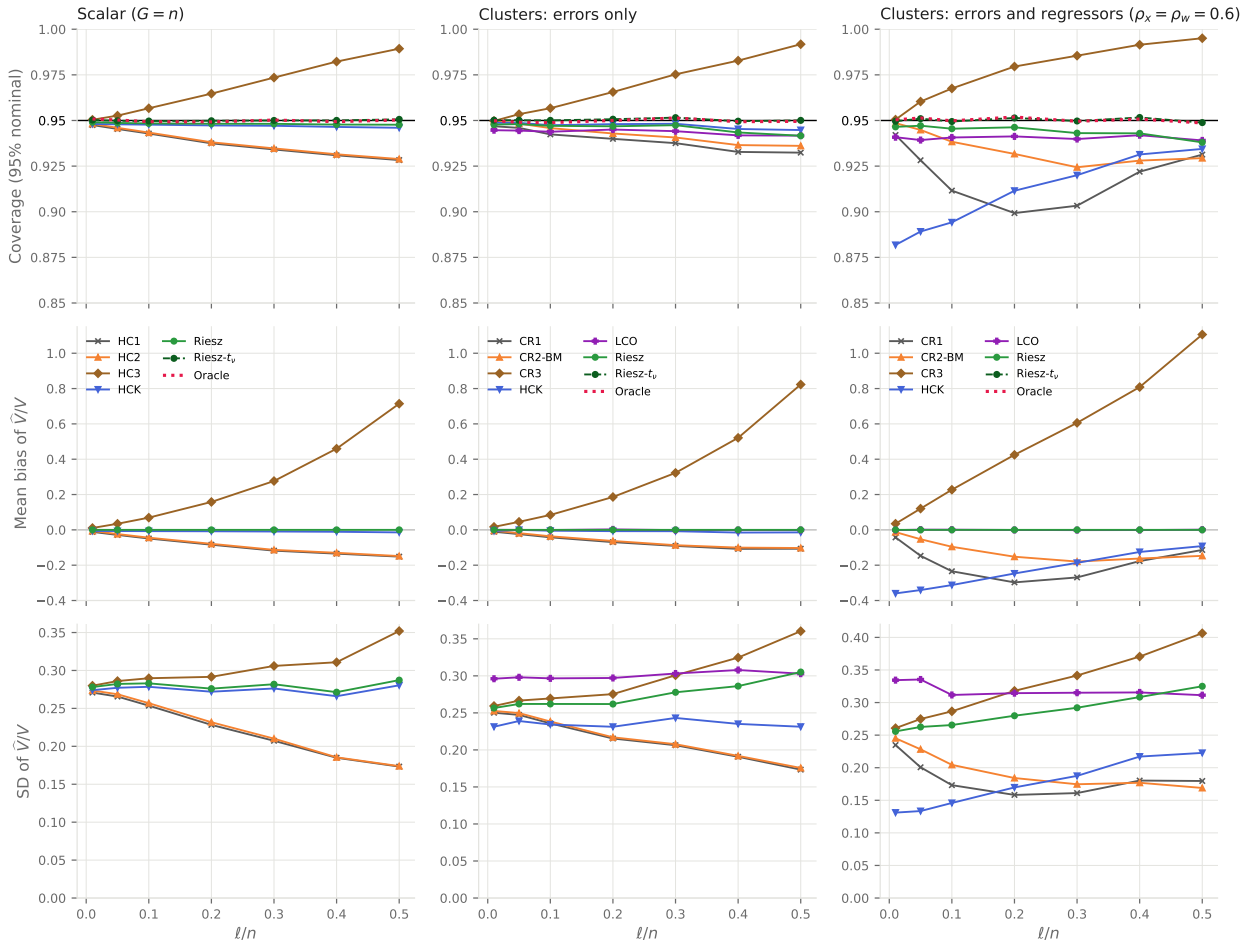


Figure 9: Exhibit 1 with every error component drawn from a standardized t_5 distribution. Columns: the scalar setting, errors-only clusters and clustered regressors ($n = 1000$, $G = 200$, $\rho_u = 0.3$).

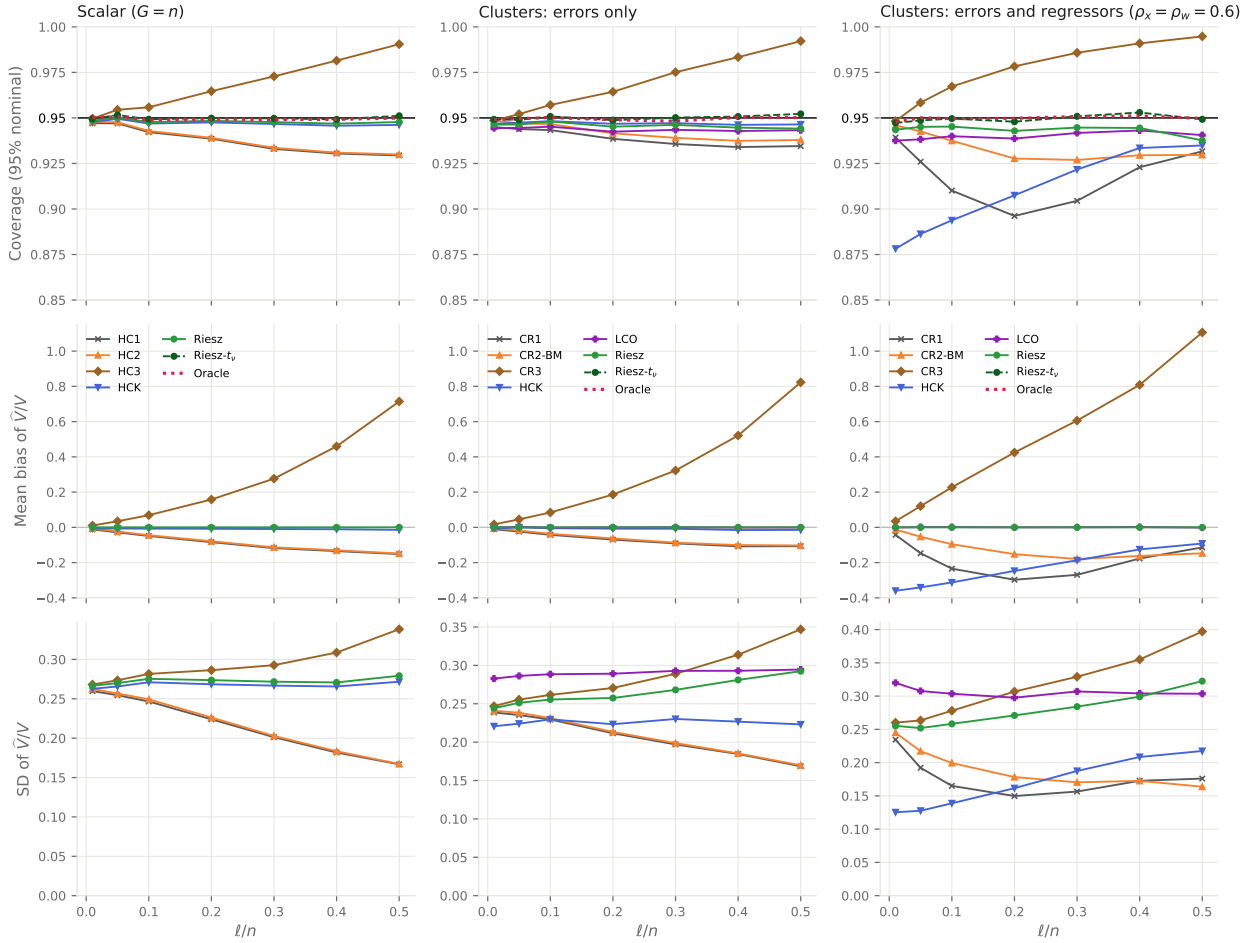


Figure 10: As figure 9, with standardized χ^2_3 errors.

Table 22: Coverage under non-Gaussian errors

Errors	Oracle	CR2-BM	Riesz	Riesz-t_ν	Class-t_ν	Neg.
Clusters with clustered regressors, $G = 200$						
Normal	0.951	0.934	0.944	0.950	0.951	0.000
t_5	0.950	0.935	0.944	0.950	0.951	0.000
χ^2_3	0.950	0.934	0.943	0.950	0.950	0.000
Clusters with clustered regressors, $G = 10$						
Normal	0.950	0.936	0.785	0.873	0.951	0.103
t_5	0.949	0.939	0.782	0.870	0.952	0.108
χ^2_3	0.948	0.940	0.783	0.871	0.951	0.107

$\vartheta = 1$. Riesz: the fully robust estimator; Class-t_ν: the class $\mathcal{D}^{\text{het}+\rho}$. Neg.: share of draws in which the fully robust Riesz estimator is negative.

H.3 Exhibit 2: existence

Coverage in Experiment C is compared against the estimators that remain defined: CR1 and the generalized-inverse CR2^{+} and CR3^{+} of Pustejovsky and Tipton (2018), which replace $M_{gg}^{-1/2}$ and M_{gg}^{-1} by Moore–Penrose powers, and HC1, HC2 and HC3 in the two-period panel.

Exhibit 2 augments W with dummies for every cluster or for a randomly chosen fraction π of them. Each such cluster's span then lies in $\mathrm{col}([X\ W])$, so M_{gg} is singular and the block corrections cannot be evaluated as written, while the absorbed direction lies in $\mathrm{Ker}(\mathcal{A}_n)$ and is harmless for the contrast by Proposition E.10. In every experiment π is capped so that $\ell + \pi G + 2 \le 0.9\,n$. The heteroskedasticity function σ_ε^2 is evaluated on the original controls, not on the added dummies. The two-period panel has independent errors ($G = n = 1000$) and 500 units observed twice; W adds a fixed effect for each unit and a period dummy to the controls, whose unit components the fixed effects absorb, and $\ell/n \le 0.3$ keeps $\ell + n/2 + 2 \le 0.9\,n$. Each M_{ii} is positive, so HC2 and HC3 are defined; why HCK is not, and why the range condition still holds, is explained with Exhibit 2 in Section 10.2.

H.4 Exhibit 3: two-way clustering

Design (Experiment E). Experiment E has two parts, which share the error model and the regressors. Errors are the two-way analogue of the error of (52),

$$u_i = \sqrt{\rho_1}\, e_{g_1(i)} + \sqrt{\rho_2}\, e_{g_2(i)} + \sqrt{1-\rho_1-\rho_2}\, \varepsilon_i, \qquad \varepsilon_i \mid X, W \sim \mathcal{N}\big(0, \sigma_\varepsilon^2(x_i, w_i)\big),$$

with $\rho_1, \rho_2 \in \{0.15, 0.3\}$, whose covariance is supported exactly on $\mathcal{S}_{2W}$. The two cluster labels are assigned independently, so the cluster sizes vary. The cluster component of the controls is on the first dimension and that of the regressor on the second, so the projection couples observations along both. The remaining parameters ($\rho_w, \rho_x \in \{0, 0.3, 0.6\}$, $\vartheta \in \{0, 1\}$, ρ_1, ρ_2 and the distribution of the controls) are drawn afresh for each design.

The validity part has $n = 1000$, $\ell = 10$ and $G_1, G_2 \in \{10, 30, 50, 100\}$, with 320 designs, twenty for each of the 16 pairs (G_1, G_2), and $R = 2,000$ draws each; the counting bound holds in every design. The identification part has $n = 500$, $G_1 = G_2 = G \in \{10, 30, 50, 100\}$ and $\ell/n \in \{0.1, 0.2, \ldots, 0.9\}$, with ten designs at each of the 36 points (360 in all) and $R = 2,000$ draws each. With independent assignment, $|\mathcal{S}_{2W}| \approx n^2(2/G - 1/G^2)$, so the bound $\frac{1}{2}(|\mathcal{S}_{2W}| + n) \le \frac{1}{2}r(r+1)$ holds approximately when $(1 - p/n)^2 \ge 2/G - 1/G^2$, that is up to $\ell/n \approx 0.56$, 0.74, 0.80 and 0.86 for $G = 10$, 30, 50 and 100. For each design we record the dimension of $\mathcal{H}_n^{\mathcal{S}_{2W}}$, the bound, which side of it the design is on, and the criterion at which LSQR stops.

The bias of CGM1. Table 23 splits the bias of CGM1 in table 6 by the cluster components of the controls, ρ_w, and of the regressor, ρ_x. The regressor's cluster component drives it: with ten clusters per dimension the median bias is between -2% and -6% when $\rho_x = 0$, and between -17% and -22% when $\rho_x \ge 0.3$. The controls' cluster component matters little, because it lies on the other dimension.

Table 23: Exhibit 3: exact relative bias of CGM1 by the cluster components of the controls and the regressor

$\min(G_1, G_2)$	ρ_w	$\rho_x = 0$	$\rho_x = 0.3$	$\rho_x = 0.6$
10	0	-1.7%	-19.5%	-22.0%
	0.3	-3.6%	-16.5%	-22.0%
	0.6	-6.4%	-19.1%	-18.5%
30	0	-1.3%	-6.6%	-11.4%
	0.3	-3.3%	-5.9%	-10.7%
	0.6	-3.3%	-10.5%	-9.8%
50, 100	0	-1.0%	-4.8%	-5.6%
	0.3	-2.4%	-4.2%	-6.2%
	0.6	-2.3%	-4.2%	-7.0%

Validity part of Experiment E ($n = 1000$, $\ell = 10$). Median over designs, 5 to 24 per cell; $\min(G_1, G_2) = 50$ and 100 are pooled. ρ_w: cluster component of the controls, on the first dimension; ρ_x: of the regressor, on the second.

Results near the counting bound. Table 7 reports the identification part split by the counting bound. The residual at which LSQR stops separates the two sides (figure 5(a)): its median is $8.4e-11$ where the bound holds and 0.78 where it fails. Besides most estimators of table 6, the table includes one-way CR0 and the variant of Thompson (2011), which subtracts HC0 instead of the intersection term. CGM and Thompson do not correct for leverage, and CGM1 only through the factor $(n-1)/(n-p)$, as HC1 does. Where the bound holds, CGM is biased by -53.8% and covers 0.775, and CGM1-t_{G-1} covers 0.914. The fully robust Riesz estimator is unbiased but, as ℓ/n grows, rests on fewer residual dimensions: its median degrees of freedom are 4.4, and it is negative in 6.6% of draws, most often near the bound (figure 11). Counting these draws as failures, Riesz-t_ν covers 0.905, although it covers 0.970 when $\widehat{V} > 0$. The class covers 0.950. Where the bound fails, Riesz is biased by -30.2% and covers 0.750. The class remains unbiased and covers nominally up to $\ell/n = 0.8$; at $\ell/n = 0.9$, with fewer than fifty residual dimensions, it falls to 0.79–0.85, so that pooled over these designs it covers 0.890, about as CGM1-t_{G-1} (0.895) (figure 5(d)).

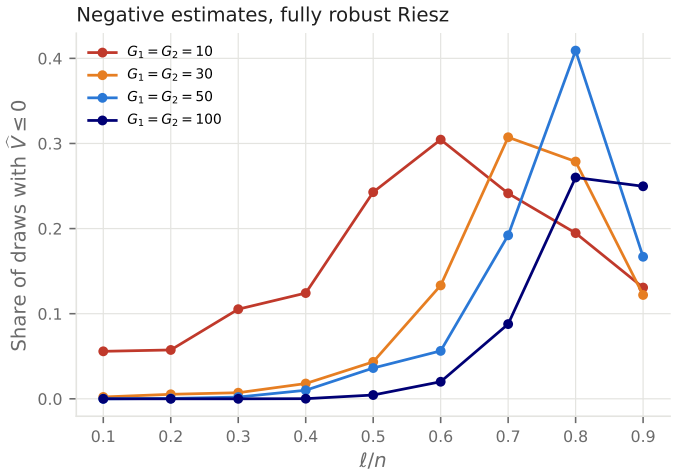


Figure 11: Exhibit 3, identification: share of draws with $\widehat{V} \leq 0$ for the fully robust Riesz estimator, by $G_1 = G_2$ and ℓ/n.

H.5 Exhibit 4: weighted and generalized least squares

Design (Experiments F, G and H). Experiment F refits the base model of Section 10.1 with $N_g = 1$ by weighted least squares with known $\Upsilon = \mathrm{diag}(v_i)$, $\log v_i \sim \mathcal{N}(0, \varsigma^2)$ and $\varsigma \in \{0, 0.25, 0.5, 1, 2\}$; the weights are unrelated to Σ, as sampling or design weights are, and $\varsigma = 0$ is OLS. Experiment G uses clusters, $(n, G) = (150, 30)$, and four weighting schemes: observation-level weights as in F, with $\varsigma \in \{0.5, 1, 2\}$ ("unit" in the tables); block weights $\Upsilon_g = [(1-r)I + rJ]^{-1}$ with a working correlation $r \in \{0.3, 0.6, 0.9\}$, as in random-effects GLS with a misspecified model; the two combined; and infeasible GLS, $\Upsilon = \Sigma^{-1}$. All are conformable with the partition. The target is $V_\Upsilon = \mathrm{Var}(a'\widehat{\beta}_\Upsilon \mid X, W)$, computed exactly. We compare the weighted cluster sandwich (WLS-CR0), the estimator $\widehat{V}_\Upsilon^M$ of (51) (Riesz-M), the whitened estimator $\widehat{V}_\Upsilon^\Upsilon$ of Proposition 9.4 (Riesz-Υ), and, in Section H.5.2, the minimum-norm solution of the full weighted system (Riesz-F(T)); because the target is known and every estimator is a quadratic form in U, the exact Gaussian standard deviation of each is also computed, free of Monte Carlo noise. Experiment H repeats G with heteroskedastic errors and weights estimated on every draw from the OLS residuals — a random-effects model with a moment estimate of the correlation, a log-variance model in the index of the design, or both — and computes Riesz-M and Riesz-Υ with the estimated weights plugged in, ignoring the first stage (Section 9.3). With estimated weights the point estimator is not linear in U, so the benchmark is its Monte Carlo variance within the design, and infeasible GLS and OLS are computed on the same draws for reference.

The quantity of interest is deterministic given the design. Let Π_{diag} and Π denote the orthogonal projections onto $\mathrm{Range}((\mathcal{A}^{T,\mathrm{diag}})^*)$ and $\mathrm{Range}((\mathcal{A}^T)^*)$, and define the projection defects

$$\zeta_{\mathrm{diag}} = \frac{\|(I - \Pi_{\mathrm{diag}})R_\Upsilon\|}{\|R_\Upsilon\|}, \qquad \zeta = \frac{\|(I - \Pi)R_\Upsilon\|}{\|R_\Upsilon\|}. \tag{H.2}$$

Under OLS Proposition 7.1 forces $\zeta_{\mathrm{diag}} = \zeta$; for conformable weights Proposition 9.4 forces $\zeta_{\mathrm{diag}} = \zeta$ as well, while with special weights that couple clusters Example G.1 has $\zeta_{\mathrm{diag}} > 0$ with $\zeta = 0$. A positive ζ_{diag} does not close in extended precision, which distinguishes a recoverability failure from a conditioning failure. Since Σ is known, the population bias implied by the defect is computed exactly rather than inferred from coverage. We also report $\kappa(\mathcal{A}_n^{T,\mathrm{diag}})/\kappa(\mathcal{A}_n^T)^2$, which Proposition E.1 fixes at one under OLS and which need not equal one in the original coordinates of a weighted fit, and $\|T'T - T\|_F$ as a design covariate.

H.5.1 Projection defects

Figure 12 shows, for Experiment F, the projection defects of the restricted and full weighted systems and the ratio of their condition numbers that bear on Exhibit 4. Both defects are at machine precision in every design (at most 3.7×10^{-15}), as Proposition 9.4 requires for weights conformable with the partition. The squaring identity of Proposition E.1 holds exactly at $\varsigma = 0$ and fails with weights: the median ratio is 1.14, 2.11 and 17 at $\varsigma = 0.25$, 1 and 2, with a maximum of 5.6×10^4.

H.5.2 The full weighted minimum-norm estimator

This subsection adds to table 8 the minimum-norm solution of the full weighted system, Riesz-F(T). By Remark 9.3 it is the efficient estimator for the working covariance $(T_\Upsilon' T_\Upsilon)^+ + \Pi_Z$. With

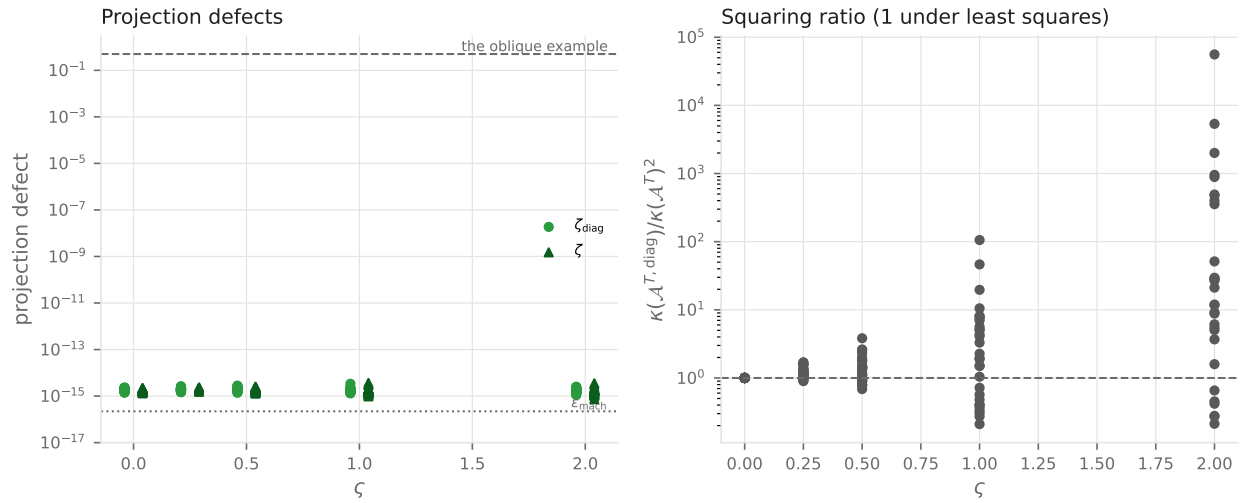


Figure 12: Experiment F: weighted least squares with $\log v_i \sim \mathcal{N}(0, \varsigma^2)$, 150 designs. Left: projection defects ζ_{diag} and ζ of (H.2), one point per design, shifted slightly apart horizontally, against the value 0.50 in Example G.1 and machine precision. Right: $\kappa(\mathcal{A}_n^{T,\text{diag}})/\kappa(\mathcal{A}_n^T)^2$, which equals one under OLS.

observation-level weights that vary little, that covariance is close to the identity, and Riesz-F(T) lies between Riesz-Υ and Riesz-M. With block and GLS weights it is far from both I_n and Υ^{-1}, and Riesz-F(T) is the least efficient of the three and covers least (table 24). Its coverage never exceeds that of the better of the other two. Its share of designs as the most efficient is at most 10%, attained with mildly dispersed observation-level weights, where its margin is below one percent.

Table 24: Experiments F and G, including the full weighted minimum-norm estimator

	Coverage				Share most efficient		
Weights	WLS-CR0	Riesz-Υ	Riesz-F(T)	Riesz-M	Riesz-Υ	Riesz-F(T)	Riesz-M
F, unit, $\varsigma = 0.5$	0.869	0.920	0.921	0.922	0.07	0.10	0.83
F, unit, $\varsigma = 2$	0.724	0.855	0.894	0.921	0.00	0.00	1.00
G, unit	0.777	0.877	0.904	0.914	0.00	0.04	0.96
G, unit + block	0.784	0.869	0.833	0.906	0.22	0.00	0.78
G, block	0.885	0.921	0.862	0.914	0.84	0.00	0.16
G, infeasible GLS	0.897	0.923	0.853	0.906	1.00	0.00	0.00

Coverage pooled over 30 (F) or 50 (G) designs per row, $R = 2{,}000$ draws, normal critical values. "Share most efficient" is the share of designs in which the estimator has the smallest exact Gaussian standard deviation among the three Riesz estimators. The oracle covers 0.949–0.951.

H.5.3 The full and the restricted system (Experiment K)

Design. Experiment K compares the full and the restricted system in the weighted residuals directly, with clusters of five, $G \in \{10, 20, 40\}$, $p \in \{3, 10\}$ regressors, and 10 designs in each of 504 cells. The weights are $\Upsilon = \exp(\varsigma\xi)$, with dispersion $\varsigma \in \{1, 2, 3\}$ and ξ at the observation or the cluster level; the errors are equicorrelated within clusters with $\rho \in \{0, 0.5\}$, and three of the four weight schemes do not model Σ: ξ standard normal and independent of the design (sampling weights); two strata, the smaller fifth of the sample oversampled by the factor e^{ς} (survey design weights); and ξ the normal scores of x (weights that move with the regressor). The fourth is correct GLS: the error variances are

$\exp(-\varsigma\xi)$, with the same equicorrelation, and $\Upsilon = \Sigma^{-1}$ blockwise, so that ς is again the dispersion of the log weights. In the first three, Σ is homoskedastic or has variances that rise with x.

The full system, whose implicit working covariance is homoskedastic when the weights are constant (Remark 9.3), is better by at least two points in 46% of the cells of the first three schemes, and the restricted system in 7%. The ordering is not uniform, which Proposition 9.6(ii), a local result under homoskedasticity, does not claim: with independent weights and heteroskedastic errors the restricted system is the better one in a few cells.

A numerical check of Proposition 9.6(ii). Table 25 computes, for one design with 20 clusters of 5, $\Sigma = I_n$ and weights $\exp(\varepsilon\xi)$, the exact excess Gaussian variance of each estimator over the homoskedastic-optimal kernel K_0, normalized by V^2. Halving ε divides the excess of Riesz-P(T) by about four and that of Riesz-F(T) by about sixteen, the rates ε^2 and ε^4 of the proof; the ratio grows from 4.5 to 261 as ε falls from 0.4 to 0.05.

Table 25: Excess variance over the homoskedastic-optimal kernel, $\Sigma = I_n$

ε	Riesz-F(T)	Riesz-P(T)	Ratio
0.4	4.7×10^{-4}	2.1×10^{-3}	4.5
0.2	2.2×10^{-5}	4.5×10^{-4}	20.5
0.1	1.4×10^{-6}	1.0×10^{-4}	73.5
0.05	9.6×10^{-8}	2.5×10^{-5}	261.2

H.5.4 Feasible GLS (Experiment H)

Table 26 reports Experiment H. The plug-in Riesz estimators remain close to their known-weight counterparts: coverage is 0.908 to 0.920, and the downward bias relative to the Monte Carlo variance of the point estimator is 1% to 8%, largest for the two-component variance model, which is the first-stage estimation the plug-in ignores. With t_ν critical values, Riesz-Υ covers 0.945 to 0.949. The weighted sandwich is biased by -22% to -34% and covers 0.866 to 0.891. Riesz-Υ edges out Riesz-M, as expected when the weights are an attempt to model Σ. Feasible GLS pays when its working model is close: the combined model reduces the variance of the point estimator by 35% relative to OLS and shortens the Riesz intervals by 21%, while the random-effects and the heteroskedastic model gain 23% and 7%.

Table 26: Experiment H: feasible GLS with plug-in variance estimation

	Coverage						Mean bias of $\widehat{V}$	
Working model	WLS-CR0	Riesz-Υ	Riesz-M	Riesz-Υ-t_ν	GLS known	OLS	Riesz-Υ	WLS-CR0
Random effects	0.871	0.920	0.914	0.949	0.954	0.950	-3.0%	-27.2%
Heteroskedastic	0.866	0.918	0.917	0.947	0.955	0.949	-1.3%	-33.7%
Both	0.891	0.920	0.908	0.945	0.955	0.949	-6.4%	-21.5%

30 designs per model, $(n, G) = (150, 30)$, $R = 1,000$ draws. Bias: relative to the Monte Carlo variance of the feasible-GLS point estimator, itself noisy by about $\pm 4.5\%$. "GLS known" is infeasible GLS with Riesz-Υ-t_ν and "OLS" the unweighted fit with Riesz-t_ν, on the same draws. The variance of the feasible-GLS point estimator relative to OLS is 0.77, 0.93 and 0.65.

H.5.5 Weights that couple clusters (Experiment J)

Experiments F and G use weights conformable with the partition. Proposition 9.5 also covers weights that couple the maintained dependence units, where whitening mixes clusters and the restricted system in the weighted residuals is no longer a change of coordinates of an orthogonal one. Experiment J checks that case.

Design. *J-spatial* places $n = 300$ points uniformly on the unit square, with 25 grid regions as clusters and the one-way class by region as the maintained model. The regressor combines a smooth spatial field with noise of heterogeneous scale. The controls are a quadratic in the coordinates ("poly"), region fixed effects plus the coordinates ("FE"), or the quadratic plus 90 random controls ("many", $\ell/n = 0.3$). The estimator is GLS with spatial-error-model weights $\Upsilon = (I - \lambda W_{\mathrm{sp}})'(I - \lambda W_{\mathrm{sp}})$, where W_{sp} is a symmetrized, row-standardized six-nearest-neighbour matrix and $\lambda \in \{0.4, 0.8, 0.95\}$. About a fifth of the absolute mass of Υ falls across region boundaries once $\lambda \geq 0.8$. The true covariance is block-diagonal by region, with exponential correlation within regions, either heteroskedastic in line with the regressor (het.) or homoskedastic (hom.).

J-serial is a single series with $n = 200$, a persistent regressor, an intercept and a trend ("few"), or these plus 30 random controls ("many"). The maintained class is the band $|t - s| \leq 4$. The estimator is Prais–Winsten GLS with working correlation $r \in \{0.5, 0.9\}$. The true errors are heteroskedastic MA(4), inside the class, or AR(1) with coefficient 0.5, outside it (†), which adds the truncation bias of Remark 8.7.

Each cell has 20 designs and $5{,}000$ Gaussian draws per design. The estimators are:

- the classical sandwich on the GLS residuals: WLS-CR0 by region in J-spatial, and the Bartlett HAC with bandwidth 4 in J-serial;
- Riesz-M (working identity);
- Riesz-Ω, the estimator $\widehat{V}^{\Omega}_{\Upsilon}$ of Proposition 9.5 with $\Omega_w = \Upsilon^{-1}$;
- Riesz-P(T), the restricted system in the weighted residuals;
- Riesz-F(T), the full weighted minimum-norm solution.

As in Section 10, a draw with $\widehat{V} \leq 0$ counts as a failure to cover; critical values are normal, and the invalid rate of Riesz-F(T) is reported separately.

Results. *Identification and exactness.* The restricted system in the weighted residuals never loses an identifying direction: every projection defect is below 8×10^{-13}. All four Riesz estimators are exactly unbiased on the maintained class, to 6×10^{-13}. The failure of Example G.1, which requires special weights, does not occur, even with weights that couple clusters.

Classical sandwich. The classical sandwich carries the projection spillover. With many controls it is biased by -29% to -54% in J-spatial at $\lambda \leq 0.8$ and by -34% in J-serial at $r = 0.5$, with coverage between 0.79 and 0.88 (table 27).

Efficiency. With weak coupling ($\lambda = 0.4$, $r = 0.5$) the four Riesz estimators nearly coincide: their standard deviations are within 18% of one another, and their coverage within about three percentage points. With strong coupling, the two estimators written in the weighted residuals degrade (table 28).

- Riesz-F(T) has up to 4.4 times the standard deviation of the most efficient estimator. It is negative in up to 28% of draws, and it has the lowest coverage in every strong-coupling cell.
- Riesz-P(T) reaches 2.8 times the smallest standard deviation at $\lambda = 0.95$.
- Riesz-M and Riesz-Ω, both functions of the OLS residuals, stay within 20% of the most efficient estimator in every cell, and cover 0.84 to 0.94.

Between these two, Riesz-Ω is the most efficient in every J-serial cell, where the AR(1) working covariance is close to the truth. In J-spatial neither dominates, since the spatial-error working covariance is not the block-diagonal truth. This is the pattern Proposition 9.5 predicts. An estimator is efficient to the extent that its working covariance is plausible, and the working covariance implicit in Riesz-F(T), $(T'T)^{+} + \Pi_Z$, becomes less plausible as the weights couple more strongly.

Table 27: Experiment J: exact bias of the classical sandwich, coverage, and invalid rate of Riesz-F(T)

			Bias	Coverage					Invalid
Controls	λ / r	Truth	classical	Classical	Riesz-M	Riesz-Ω	Riesz-P(T)	Riesz-F(T)	Riesz-F(T)
J-spatial									
poly	0.4	het.	-0.14	0.926	0.939	0.939	0.939	0.938	0.000
poly	0.4	hom.	-0.13	0.917	0.930	0.929	0.931	0.930	0.000
poly	0.8	het.	-0.13	0.931	0.939	0.940	0.940	0.939	0.000
poly	0.8	hom.	-0.07	0.928	0.930	0.926	0.934	0.914	0.002
poly	0.95	het.	-0.12	0.924	0.939	0.938	0.935	0.916	0.005
poly	0.95	hom.	$+0.12$	0.948	0.929	0.920	0.871	0.694	0.230
FE	0.4	het.	-0.11	0.923	0.936	0.936	0.936	0.936	0.000
FE	0.4	hom.	-0.09	0.925	0.934	0.932	0.933	0.933	0.000
FE	0.8	het.	-0.13	0.921	0.940	0.937	0.938	0.936	0.000
FE	0.8	hom.	-0.08	0.924	0.931	0.928	0.928	0.896	0.007
FE	0.95	het.	-0.14	0.922	0.940	0.939	0.937	0.893	0.016
FE	0.95	hom.	-0.07	0.925	0.931	0.920	0.897	0.660	0.274
many	0.4	het.	-0.54	0.788	0.929	0.932	0.930	0.929	0.000
many	0.4	hom.	-0.33	0.878	0.880	0.901	0.902	0.870	0.027
many	0.8	het.	-0.53	0.792	0.927	0.930	0.929	0.917	0.001
many	0.8	hom.	-0.29	0.884	0.849	0.883	0.892	0.780	0.116
many	0.95	het.	-0.49	0.815	0.925	0.923	0.896	0.862	0.034
many	0.95	hom.	-0.07	0.921	0.845	0.869	0.716	0.657	0.279
J-serial									
few	0.5	MA(4)	-0.23	0.900	0.929	0.935	0.928	0.929	0.000
few	0.5	AR(1)†	-0.04	0.938	0.930	0.930	0.931	0.929	0.000
few	0.9	MA(4)	$+0.07$	0.953	0.927	0.938	0.928	0.818	0.075
few	0.9	AR(1)†	$+0.17$	0.961	0.933	0.932	0.925	0.844	0.053
many	0.5	MA(4)	-0.34	0.870	0.921	0.931	0.924	0.919	0.000
many	0.5	AR(1)†	-0.16	0.917	0.920	0.924	0.923	0.916	0.001
many	0.9	MA(4)	-0.09	0.931	0.905	0.930	0.910	0.800	0.090
many	0.9	AR(1)†	$+0.00$	0.945	0.910	0.918	0.904	0.793	0.112

Medians over 20 designs, 5,000 draws each. Bias is the exact $\mathrm{tr}(K\Sigma)/V - 1$ of the classical sandwich. The Riesz estimators are exactly unbiased on the maintained class. Coverage counts draws with $\widehat{V} \leq 0$ as failures; normal critical values. †Outside the maintained class.

Table 28: Experiment J: exact Gaussian standard deviation relative to the most efficient Riesz estimator

Controls	λ / r	Truth	Riesz-M	Riesz-Ω	Riesz-P(T)	Riesz-F(T)
J-spatial						
poly	0.4	het.	1.00	1.00	1.00	1.00
poly	0.4	hom.	1.03	1.05	1.00	1.03
poly	0.8	het.	1.00	1.00	1.00	1.01
poly	0.8	hom.	1.07	1.18	1.00	1.31
poly	0.95	het.	1.00	1.00	1.00	1.06
poly	0.95	hom.	1.00	1.17	1.52	3.84
FE	0.4	het.	1.00	1.00	1.00	1.00
FE	0.4	hom.	1.01	1.03	1.00	1.01
FE	0.8	het.	1.00	1.00	1.00	1.01
FE	0.8	hom.	1.00	1.11	1.07	1.42
FE	0.95	het.	1.00	1.01	1.01	1.14
FE	0.95	hom.	1.00	1.19	1.43	4.42
many	0.4	het.	1.00	1.00	1.00	1.01
many	0.4	hom.	1.13	1.02	1.00	1.18
many	0.8	het.	1.00	1.00	1.00	1.04
many	0.8	hom.	1.20	1.07	1.00	1.60
many	0.95	het.	1.00	1.00	1.06	1.22
many	0.95	hom.	1.08	1.00	2.80	2.66
J-serial						
few	0.5	MA(4)	1.08	1.00	1.08	1.08
few	0.5	AR(1)†	1.08	1.00	1.08	1.12
few	0.9	MA(4)	1.11	1.00	1.08	1.99
few	0.9	AR(1)†	1.06	1.00	1.07	2.75
many	0.5	MA(4)	1.10	1.00	1.08	1.11
many	0.5	AR(1)†	1.13	1.00	1.08	1.18
many	0.9	MA(4)	1.17	1.00	1.12	2.01
many	0.9	AR(1)†	1.11	1.00	1.08	2.53

Medians over 20 designs of $\mathrm{SD}(\widehat{V})/\min \mathrm{SD}(\widehat{V})$ across the four Riesz estimators, computed exactly under the true covariance.

H.6 Exhibit 5: the level factor

Design (Experiments L and N). Both experiments use the error model of Exhibit 1, with $\rho_u = 0.3$ and $\vartheta = 1$. Experiment L ($n = 1000$) has two parts. At $G = 100$ it varies the size imbalance $\eta_N \in \{0, 2, 4\}$ at $\ell/n \in \{0.01, 0.05, 0.2\}$ with clustered regressors, $\rho_w = \rho_x = 0.3$, in three settings for the regressor's cluster component: random, as in Exhibit 1; scaled by $\bar{N}/N_g$, which loads it on the small clusters and equalizes the clusters' shares of the variation; and scaled by $N_g/\bar{N}$, which loads it on the large clusters. At $\eta_N = 4$ the smallest clusters are singletons. At $G = 200$, $\eta_N = 1$ and $\ell/n = 0.05$ it replaces the regressor by a cluster-level treatment on $T \in \{5, 10, 20, 50\}$ clusters. Experiment N has dyadic designs: $G = 240$ dyads of three observations linking two node sets, every side-B node in $k \in \{3, 6, 12\}$ dyads. In the *node* setting the controls are node fixed effects on both sides, each shared by the dyads of one node; in the *diffuse* setting they are the same number of dyad-level normal controls, which matches p, p/n and p/G. The two settings give the same off-block mass at every k, so they isolate the role of κ_n.

Further results. κ_n stays close to p/n when only the errors are clustered; clustered regressors, $\rho_w = \rho_x = 0.6$, at least double it when the controls are few, and the gap closes as ℓ/n grows, because the controls' own cluster components then absorb the cluster means. The median distance of CR0's exact bias from $-\kappa_n$ is 0.050, against 0.076 from $-p/n$. With unequal sizes and the regressor random, CR1 is biased by -24.0% at $\eta_N = 4$ and $\ell/n = 0.05$, between the equal-size and the large-cluster cases of Exhibit 5; loading the regressor on the small clusters equalizes the clusters' shares and lowers the bias to -6.3%. At $\ell/n = 0.2$, p/n dominates the level and κ_n no longer moves with the sizes. The dyadic designs at $k = 3$ have off-block masses 0.107 and 0.106, so the difference in their bias (Section 10.2) is carried by κ_n.

H.7 Further results

H.7.1 Conditioning of the restricted system

Grid. These results use a wider grid than the exhibits, which adds a common factor shared by the regressor and the controls: x_{gi} and each non-intercept control $w_{gi}^{(j)}$ are replaced by $\sqrt{\varphi}\, b_g f_{gi} + \sqrt{1-\varphi}$ times their base-model value, with $f_{gi} \sim \mathcal{N}(0,1)$ and cluster-specific Rademacher loadings b_g, drawn separately for x and each control; $\varphi = 0$ is the base model. The factor makes the regressor and the controls share directions, which concentrates the leverage and raises the median $\kappa(\mathcal{A}_n)$ from about 1.5 at $\varphi = 0$ to about 5 in Experiment A and 10 in Experiment B at $\varphi = 0.8$, so the grid spans a range of conditioning that the base designs do not. With $N_g = 1$ the parameters ρ_w, ρ_x, ρ_u and η_N have no effect, so Experiment A crosses ℓ/n, the control distribution (normal, uniform or rare indicators), $\varphi \in \{0, 0.2, \ldots, 0.8\}$ and $\vartheta \in \{0, 1\}$ (210 cells). Experiment B varies the parameters on which the predictions of table 2 turn, $\ell/n \times \varphi \times \vartheta \times \rho_u$ with $\rho_u \in \{0, 0.3, 0.6\}$, at normal controls, $\rho_w = \rho_x = 0.3$ and $\eta_N = 1$ (210 cells), and adds one-at-a-time sweeps of the control distribution, $\rho_w, \rho_x \in \{0, 0.3, 0.6\}$ and $\eta_N \in \{0, 1\}$, each crossed with ℓ/n (49 further cells). To let $\kappa(\mathcal{A}_n)$ vary more widely than the base designs allow, each of A and B adds 100 designs in which W is augmented with dummies for a fraction π of clusters, as in Section H.3, with $\pi \sim \mathrm{U}[0.1, 0.9]$ in B and $\pi \sim \mathrm{U}[0.01, 0.10]$ in A, where a dummy is an observation indicator, perturbed as $\Delta_g + \tau\zeta_g$ with $\zeta_{gi} \sim \mathcal{N}(0,1)$ and

$\tau = 10^{\mathrm{U}[-3,-0.5]}$; the remaining parameters are drawn from their grid values. At $\tau = 0$ the absorbed direction lies in $\mathrm{Ker}(\mathcal{A}_n)$ and does not raise $\kappa(\mathcal{A}_n)$, which is taken on the identified subspace. At small $\tau > 0$ the direction is retained but only weakly preserved, which moves $\kappa(\mathcal{A}_n)$ over several orders of magnitude (Section 2.2). These are numerical devices; the base designs use $\pi = 0$.

Three routes. The conditioning results compare three ways of computing the same estimator. By Corollary 7.2 they target the same $H_n^\star$, so any difference between computed values is numerical. Full-LSQR applies LSQR to the pair $(\mathcal{A}_n, \mathcal{A}_n^*)$; Restricted-LSQR applies the same LSQR implementation to $\mathcal{A}_n^{\mathrm{diag}}$, the restricted system on its own terms; and Restricted-direct solves it to machine accuracy by iterative refinement from the eigendecomposition of $\mathcal{A}_n^{\mathrm{diag}}$, with residuals evaluated in double-double arithmetic, certified against a 50-digit singular value decomposition on small designs. The gap between Restricted-LSQR and Restricted-direct measures the $\kappa(\mathcal{A}_n)^2$ cost predicted by Proposition E.1; the gap between Full-LSQR and Restricted-direct measures the stopping tolerance. All gaps are computed on $D = \mathcal{A}_n(H)$ rather than on H, since two solutions differing by an element of $\mathrm{Ker}(\mathcal{A}_n)$ give identical estimates, and on the weights rather than the estimates, so that none carries sampling noise.

Results. Table 29 compares the three routes to the Riesz estimator on the 669 designs of Experiments A and B, whose condition numbers range from 1.03 to 9.0×10^3. On the grid itself $\kappa(\mathcal{A}_n)$ is at most 83 in Experiment B; the higher values come from the absorption devices. The prediction of Proposition E.1 is borne out in iterations. Solved on its own terms, the restricted system needs 1.6 times as many iterations as the full system for $\kappa(\mathcal{A}_n) < 3$, 9.9 times as many for $\kappa \in [10, 30)$ and 28 times as many for $\kappa \in [30, 10^2)$; from $\kappa \approx 55$ it increasingly exhausts the cap of 20,000 iterations, which the full system first reaches at $\kappa \approx 2.5 \times 10^3$. In accuracy the prediction is not visible at the tolerance used. Wherever both solvers converge, the median ratio of Restricted-LSQR's distance from Restricted-direct to Full-LSQR's is between 1.1 and 1.4 in every bin, up to $\kappa \approx 8.3 \times 10^3$: both track the stopping tolerance rather than $\kappa(\mathcal{A}_n)^2 \varepsilon_{\mathrm{mach}}$. The larger gaps in the top bins of table 29 come from runs stopped at the cap.

A controlled comparison on a single design (Section H.7.2) separates the costs in iterations, accuracy and wall time.

Table 29: Experiment B: the three computational routes by condition number

		Gap to Restricted-direct		Iterations		Share at cap	
$\kappa(\mathcal{A}_n)$	Designs	Restr.	Full	Restr.	Full	Restr.	Full
$[1, 3)$	158	7.9×10^{-11}	5.4×10^{-11}	23	14	0	0
$[3, 10)$	87	1.9×10^{-10}	1.6×10^{-10}	148	42	0	0
$[10, 30)$	49	5.4×10^{-10}	5.0×10^{-10}	1,311	132	0	0
$[30, 10^2)$	25	1.7×10^{-9}	1.4×10^{-9}	10,389	370	0.12	0
$[10^2, 10^3)$	23	4.1×10^{-8}	1.0×10^{-8}	12,530	691	0.48	0
$[10^3, 10^4)$	17	8.5×10^{-7}	3.2×10^{-7}	20,000	979	0.53	0.29

Restr.: Restricted-LSQR; Full: Full-LSQR. Medians within each bin. Gaps are $\|D - D_{\mathrm{exact}}\|_F / \|D_{\mathrm{exact}}\|_F$ with $D = \mathcal{A}_n(H)$. Iteration caps are 20,000 (Restricted-LSQR) and 4,000 (Full-LSQR, $G = 200$). In Experiment A, $\kappa(\mathcal{A}_n) \leq 1.4 \times 10^3$ and neither solver reaches its cap.

H.7.2 Accuracy and wall time of the computational routes

To separate the costs in iterations and accuracy, and to measure time, which the grid does not isolate, table 30 holds one Experiment B design fixed ($n = 1000$, $G = 200$, $\eta_N = 0$, $\dim \mathcal{H}_n = 3{,}000$) and moves $\kappa(\mathcal{A}_n)$ from 3.6 to 7,748 through the absorption defect τ, solving at tolerances 10^{-10} and 10^{-14}. Three findings qualify the practical reading of Proposition E.1, while the proposition itself, $\kappa(\mathcal{A}_n^{\text{diag}}) = \kappa(\mathcal{A}_n)^2$, holds to twelve digits throughout.

1. *Iterations.* Restricted-LSQR needs 3 times as many iterations as Full-LSQR at $\kappa \approx 3.6$ and 14–15 times as many from $\kappa \approx 10^3$. This is the predicted cost of solving the squared system.

2. *Accuracy.* At 10^{-10} both solvers lose accuracy alike, roughly in proportion to the tolerance times κ. At 10^{-14} the squaring appears, but mildly: at $\kappa = 7{,}748$ Restricted-LSQR is 10 times less accurate than Full-LSQR (8×10^{-10} against 8×10^{-11}), 16 times better than the worst case $\kappa^2 \varepsilon_{\text{mach}}$, and at $\kappa = 862$ the two agree within a factor of 3. Every numerical gap is at least six orders of magnitude below the sampling error of $\widehat{V}$, so none of them affects inference.

3. *Time.* The ordering reverses. Once $\mathcal{A}_n^{\text{diag}}$ is assembled, an iteration of Restricted-LSQR is a $3{,}000 \times 3{,}000$ matrix–vector product, while an iteration of Full-LSQR applies $\mathcal{A}_n$ and $\mathcal{A}_n^*$, two dense $n \times n \times n$ products, about fifty times the cost. Restricted-LSQR is therefore 4 to 15 times faster in wall time, and the direct solve Restricted-direct, which assembles and diagonalizes $\mathcal{A}_n^{\text{diag}}$, is fastest of all at high κ, at a cost independent of κ.

The advantage of the matrix-free full route is thus not speed at these dimensions. Assembling $\mathcal{A}_n^{\text{diag}}$ costs memory quadratic in $\dim \mathcal{H}_n = \sum_g N_g(N_g+1)/2$ and a direct solve costs time cubic in it. That dimension grows with the square of the cluster sizes: it is already 10,881 in the most unbalanced designs of Experiment L, at $n = 1000$, and of order n^2/G for the two-way supports of Experiment E. For such designs the restricted system can still be solved without assembly, by applying $\mathcal{A}_n^{\text{diag}}$ as $\mathcal{A}_n^*$ after $\mathcal{A}_n$. What distinguishes the full route is then its conditioning: it works with $\kappa(\mathcal{A}_n)$ where the restricted route works with $\kappa(\mathcal{A}_n)^2$; Experiment E, with $\dim \mathcal{H}_n$ of order n^2/G, is run on it.

Table 30: Accuracy and wall time of the three computational routes against $\kappa(\mathcal{A}_n)$

		Full-LSQR			Restricted-LSQR			Restricted-direct
$\kappa(\mathcal{A}_n)$	Tol.	Gap	Iter.	Sec.	Gap	Iter.	Sec.	Sec.
3.6	10^{-10}	1×10^{-10}	33	1.1	2×10^{-10}	95	0.1	2.2
10.5	10^{-10}	3×10^{-10}	101	3.1	6×10^{-10}	761	0.4	2.1
78.8	10^{-10}	4×10^{-9}	293	9.6	4×10^{-9}	3,516	1.8	2.2
862	10^{-10}	5×10^{-8}	542	17.8	4×10^{-8}	7,763	4.7	2.3
7,748	10^{-10}	4×10^{-7}	775	25.4	4×10^{-7}	11,704	6.9	2.4
862	10^{-14}	5×10^{-12}	806	26.0	1×10^{-11}	11,751	6.9	2.3
7,748	10^{-14}	8×10^{-11}	1,197	39.2	8×10^{-10}	18,901	11.2	2.4

One Experiment B design with $\pi = 0.5$ and $\tau \in \{0.03, 0.01, 0.003, 0.001\}$ (first row: no absorption). Gap is $\|D - D_{\text{exact}}\|_F / \|D_{\text{exact}}\|_F$ with $D = \mathcal{A}_n(H)$. Seconds are single-process wall time on a laptop (Apple silicon, multithreaded BLAS); Restricted-LSQR excludes, and Restricted-direct includes, the assembly of $\mathcal{A}_n^{\text{diag}}$. All runs halted on the primal criterion.

H.7.3 Misspecified partition

Design (Experiment D). The estimator is given the $G = 30$ clusters, but the cluster shock is drawn on the coarser partition that pools adjacent clusters in pairs, $\varepsilon_g = \sqrt{1-\varkappa}\, e_g + \sqrt{\varkappa}\, e_{p(g)}$, so that $\varkappa = 0$ is the correctly specified control and $\varkappa = 1$ makes each pair one cluster. The cluster components $z_g^{(j)}$ and v_g of the controls and the regressor are drawn on the same pairs, for every $\varkappa$. This is what makes the experiment informative: with the regressor's cluster component on the fine partition, the influence weights of adjacent clusters are independent, and the omitted cross-boundary covariance contributes to V with either sign and on average nothing (a median of -0.6% over the designs with $\varkappa > 0$). With it on the pairs, it contributes a median of 8%, and its median magnitude grows from 8% of V at $\varkappa = 0.25$ to 21% at $\varkappa = 1$. Since the regressors do not change with $\varkappa$, any degradation is caused by the misspecified error partition alone. We draw $\rho_w, \rho_x, \rho_u \in \{0.3, 0.6\}$: $\varkappa$ has no effect when $\rho_u = 0$, and on average none when the regressors have no cluster component on the pairs.

Table 31 reports Experiment D. Every estimator degrades with $\varkappa$. Apart from CR3, whose large upward bias the omitted covariance absorbs, the Riesz estimator's bias moves the most, by -0.27 from $\varkappa = 0$ to 1, against -0.22 for CR0, -0.22 for CR1, -0.22 for CR2 and -0.23 for the leave-cluster-out estimator. This is the predicted exposure: an estimator that inverts the maintained structure exactly inherits all of its misspecification. The margin is small, however, and exposure is not level: the Riesz estimator starts exactly unbiased, so at $\varkappa = 1$ its bias (-27%) is still about half of CR0's (-51%), comparable to CR2's (-28%). Its coverage falls from 0.909 to 0.852, against 0.925 to 0.874 for CR2, and with t_ν critical values from 0.942 to 0.896. CR3 covers best at every $\varkappa$, but only because its upward bias, 34% at $\varkappa = 0$, offsets the omitted covariance as $\varkappa$ grows. CR2, CR3 and the leave-cluster-out estimator do not exist in 26% of the designs, those with discrete controls in which some M_{gg} is singular; their columns are computed on the others.

Table 31: Experiment D: degradation under a misspecified partition

	Coverage				Exact mean bias of $\widehat{V}/V$			
$\varkappa$	0	0.25	0.5	1	0	0.25	0.5	1
CR0	0.863	0.859	0.847	0.799	-28.9%	-30.4%	-36.6%	-51.1%
CR1	0.916	0.909	0.894	0.863	-10.6%	-16.1%	-23.0%	-32.6%
CR2	0.925	0.915	0.903	0.874	-6.2%	-11.9%	-20.1%	-28.3%
CR3	0.967	0.958	0.942	0.929	33.5%	44.7%	1.9%	3.3%
LCO	0.894	0.873	0.870	0.826	0.1%	-5.0%	-14.3%	-23.1%
Riesz	0.909	0.900	0.886	0.852	0.0%	-9.7%	-18.0%	-27.4%
Riesz-t_ν	0.942	0.934	0.920	0.896				

50 designs per value of $\varkappa$, each with $R = 5,000$ draws. The oracle covers 0.949–0.951.

I Supplement to the Applications

I.1 Generalized-inverse corrections in the governor application

Some M_{gg} are singular in both sets of regressions of Section 11.1, so CR2 and CR3 are not defined as written. Table 32 reports their generalized-inverse versions CR2$^+$ and CR3$^+$ (Pustejovsky and Tipton, 2018), which replace $M_{gg}^{-1/2}$ and M_{gg}^{-1} by Moore–Penrose powers, alongside CR1 and the

Riesz estimator. CR3$^+$ is the block form, without the jackknife factor $(G-1)/G$, which with $G = 1{,}518$ is immaterial.

Relative to the Riesz estimator, CR2$^+$ is smaller and CR3$^+$ larger in every salary regression. In the fiscal regressions CR2$^+$ is within 7% of the Riesz estimator and CR3$^+$ is 32% to 43% larger. The two corrections reach different conclusions at the 5% level in three regressions. For shared ancestry, CR2$^+$ gives $p = 0.047$ and CR3$^+$ gives 0.110, and the Riesz estimator lies between them at 0.069: significant at 10% but not at 5%. For total revenue and pre-1930 revenue, CR2$^+$ gives 0.040 and 0.018 and CR3$^+$ gives 0.131 and 0.080, and the Riesz estimator, at 0.041 and 0.020, agrees with CR2$^+$.

Table 32: Generalized-inverse corrections in the colonial-governor regressions

	Standard error				p-value			
Regression	CR1	CR2$^+$	CR3$^+$	Riesz	CR1	CR2$^+$	CR3$^+$	Riesz
Salary regressions (Table 10)								
(1) Shared ancestors	0.0470	0.0520	0.0647	0.0570	0.028	0.047	0.110	0.069
(2) Both aristocrats	0.1238	0.1615	0.2361	0.1838	0.083	0.183	0.363	0.242
(3) Both Eton	0.0772	0.0912	0.1196	0.1002	0.085	0.145	0.267	0.185
(4) Both Oxbridge	0.0468	0.0521	0.0646	0.0562	0.126	0.169	0.267	0.202
(6) Connected	0.0356	0.0389	0.0479	0.0414	0.006	0.012	0.042	0.019
Fiscal regressions (Table 11)								
(1) Revenue	0.0174	0.0195	0.0265	0.0196	0.021	0.040	0.131	0.041
(2) Revenue, pre-1930	0.0208	0.0232	0.0314	0.0236	0.008	0.018	0.080	0.020
(3) Customs revenue	0.0256	0.0291	0.0402	0.0304	0.040	0.071	0.191	0.084
(4) Internal revenue	0.0319	0.0367	0.0517	0.0370	0.175	0.239	0.402	0.243
(5) Expenditure	0.0189	0.0210	0.0283	0.0206	0.120	0.163	0.300	0.154
(6) Expenditure, pre-1930	0.0232	0.0257	0.0345	0.0254	0.069	0.101	0.221	0.096
(7) Revenue services	0.0535	0.0582	0.0763	0.0563	0.094	0.124	0.241	0.112
(8) Public works	0.0619	0.0695	0.0935	0.0654	0.085	0.125	0.255	0.103

p-values with normal critical values.

I.2 The weighted estimators in the housing wealth application

On a randomly drawn $\Sigma \in \mathcal{H}_n$ the exact relative bias of Riesz-M, Riesz-P(T) and Riesz-F(T) in table 12 is below 10^{-12}, and so is that of the class estimator on a member of its class. In the fitted class member of Section 11.2, $\hat{\rho}$ is between 0.04 and 0.11.

The fitted covariances of Section 11.2 have η between -0.45 and -0.41 and a within-state correlation of 0 or 0.1. Under them the exact Gaussian standard deviation of $\widehat{V}/V_\Upsilon$ is 1.07 to 2.07 for Riesz-M, 1.07 to 1.97 for Riesz-P(T) and 1.34 to 2.51 for Riesz-F(T); Riesz-M has the smaller standard deviation in every regression without within-state correlation, and Riesz-P(T) in six of seven with correlation 0.1. Under $\Sigma = \Upsilon^{-1}$ the three ranges are 2.3 to 5.4, 0.9 to 1.9 and 4.0 to 9.3. Under $\Sigma = I_n$ they are 1.05 to 1.24, 1.24 to 2.01 and 1.15 to 1.42, so Riesz-M is the most precise and Riesz-P(T) the least in every regression.

CR3, written as the quadratic form $\frac{G-1}{G}\sum_g (a_g' U)^2$ with a_g the change in the influence weights ω_Υ when state g is dropped ($a_g' Z = 0$), has expected value 1.24 to 1.87 times V_Υ under the fitted covariances, under $\Sigma = \Upsilon^{-1}$ and under $\Sigma = I_n$.