Estimation and exclusion restrictions in clustered linear models*

Anna Mikusheva[†], Mikkel Sølvsten[‡], Baiyun Jing[§]

March 6, 2026

Abstract

We study linear regression models with clustered data, high-dimensional controls, and intricate exclusion restrictions. We propose a correctly centered internal instrument IV estimator that accommodates a broad class of exclusion restrictions and allows within-cluster dependence. The estimator admits a simple leave-out interpretation and is computationally tractable. We derive a central limit theorem for the associated quadratic form and propose a robust variance estimator. We also develop identification-robust inference procedures. Our framework extends dynamic panel methods to general clustered settings. We illustrate the approach in a large-scale fiscal intervention in rural Kenya, where spatial interference generates the exclusion-restriction pattern.

Keywords: weak exogeneity, many controls, interference, OLS inconsistency, internal instrument, leave-out approach

JEL Codes: C13, C22

*We are grateful to Isaiah Andrews, Patrik Guggenberger, Lihua Lei, and Morten Ø. Nielsen for helpful and insightful discussions. We thank the authors of Egger et al. (2022) for sharing access to their data. This research was supported by grants from the Danish National and Aarhus University Research Foundations (DNRF grant number DNRF186 and AUFF grant number AUFF-E-2022-7-3)

†Department of Economics, M.I.T., and Aarhus Center for Econometrics, 50 Memorial Drive, E52-526, Cambridge, MA, 02142, United States. E-mail: amikushe@mit.edu.; corresponding author

‡Department of Economics and Aarhus Center for Econometrics, Aarhus University, Fuglsangs Allé 4, Building 2621, B 13a, 8210 Aarhus V, Denmark. E-mail: miso@econ.au.dk.

§Department of Economics, Harvard University, Littauer Center, Cambridge, MA, 02138, United States. E-mail: bjing@fas.harvard.edu

1 Introduction

Clustered datasets—such as panel, network, spatial, or group-structured data in which individuals or firms are observed repeatedly or nested within groups—have become increasingly common in empirical research. These settings present several methodological challenges. One is the need to accommodate heterogeneity, often addressed by including fixed effects and flexible time or group trends, which typically requires many control variables. A more fundamental difficulty is within-cluster dependence: observations in the same cluster may be correlated due to, e.g., spatial and network interference, spillover effects, and time-series dependence. Such dependencies complicate statistical inference and limit the effective aggregation of information across observations.

The concept of exogeneity becomes more nuanced in the presence of clustered data. As we demonstrate, assuming only a per-observation exclusion restriction—uncorrelatedness of the error with the regressor within an observation—may deliver no identifying variation for consistent estimation of structural parameters. By contrast, strict exogeneity—each error term uncorrelated with all regressors in the cluster—is often implausible in many empirical contexts.

To bridge strict exogeneity and unrestricted dependence, we assume the error term is uncorrelated with a subset of within-cluster regressors; the subset is application-specific. In panel data, it is common to assume errors are mean-zero conditional on current and past (but not future) regressors, allowing feedback from shocks to future policies. In spatial contexts, localized spillovers motivate analogous restrictions: regressors corresponding to sufficiently distant units within a cluster may plausibly be uncorrelated with a given error term, even if nearby regressors are not. Examples include spatial leakage in deforestation (Jayachandran et al., 2017), spillovers in policing (Blattman et al., 2021), and fiscal intervention experiments in rural Kenya (Egger et al., 2022). In network settings, treatment effects may propagate along social ties, so regressors associated with nodes that are unconnected—or sufficiently distant in the network—may be uncorrelated with a given error term (Paluck et al., 2016). These partial exogeneity assumptions more accurately reflect the dependence structures present in many empirical applications.

When only a subset of exclusion restrictions holds, it becomes inappropriate to treat regressors as fixed or to condition on them. Their stochastic variation may correlate with the errors, undermining standard arguments that rely on conditioning on regressors. As a result, even estimators like OLS must be interpreted as ratios of stochastic quadratic forms, which introduces several challenges.

The most well-known is *Nickell bias* (Nickell, 1981), which occurs when the expected value of the OLS error numerator is nonzero, leading to asymptotic bias. While prominent in dynamic panels with fixed effects and lagged outcomes, analogous bias arises more broadly under clustered dependence. A second challenge is inference: standard variance estimators often fail to account for dependence among cross-cluster terms in the quadratic form. Finally, the denominator may remain asymptotically random, unlike in classical settings, leading to *weak identification* and further complicating inference.

This paper studies the estimation of structural parameters in linear regression models with clustered data, high-dimensional controls, and researcher-specified exclusion restrictions. A key feature of the framework is the grouping of observations into disjoint clusters, permitting within-cluster dependence and independence across clusters. The setting accommodates unbalanced panels with multiple high-dimensional fixed effects, as well as spatial, network, and group-structured data. Our approach extends dynamic panel methods to a broader class of models, addressing the central challenges of bias, inference, and identification.

Our first contribution is to characterize a class of *correctly centered* internal instrument estimators that remove the leading asymptotic bias of OLS. The approach accommodates high-dimensional exogenous controls and adapts to the exclusion structure specified in the application. The resulting estimator is an internal instrument IV estimator, defined as the one closest to OLS in a specified norm. The estimator is easy to implement and asymptotically efficient under some conditions.

The procedure has a simple interpretation: for each observation, controls are partialled out using only those observations whose errors are uncorrelated with the observation's regressor, yielding an observation-specific leave-out projection. A just-identified IV regression is then performed on the transformed equation, using the original regressor as the instrument.

We further characterize the efficiency loss from varying the strength of exclusion restrictions. In particular, assuming exogeneity only at the observation level eliminates all identifying variation when fixed effects are present. This observation underscores the importance of carefully specifying the exogeneity structure in applied work.

Our second contribution addresses inference. We show that, outside special cases such as (i) strict exogeneity and (ii) block-diagonal residualization (e.g., models with only cluster fixed effects), the estimator's numerator is a nontrivial quadratic form in the errors. In these more general settings, standard cluster-robust variance estimators may fail, and valid inference requires central limit theorems for clustered quadratic forms. These issues are especially acute in models with many high-dimensional controls, such as two-way fixed effects (Verdier, 2020). We propose a new central limit theorem and variance estimator that account for this structure.

Our third contribution addresses weak identification, which arises when the internal instrument—though valid—captures little identifying variation due to many controls or weak exclusion restrictions. As in the classical dynamic panel literature (Blundell and Bond, 1998; Bun and Windmeijer, 2010), weak identification can compromise standard inference. We develop inference procedures that remain valid even when the estimator's denominator exhibits substantial sampling variability.

We propose a comprehensive empirical strategy that combines a correctly centered estimator with identification-robust inference and confidence sets. We apply it to a prominent randomized evaluation of a large-scale fiscal intervention in rural Kenya (Egger et al., 2022). A key challenge in this setting is spatial interference: treatment in one village affects outcomes in neighboring villages, complicating the choice of plausible exclusion restrictions. We show that weakening the exclusion restrictions yields less precise estimates and wider confidence sets.

This paper contributes to several strands of the econometrics literature. It relates to the extensive work on linear dynamic panel data models—a mature field with too many important contributions to list exhaustively. For recent overviews on correcting Nickell bias, see the surveys by Okui (2021) and Bun and Sarafidis (2015). Our estimation approach is more closely aligned with the internal instrument framework developed in Anderson and Hsiao (1981), Arellano and Bond (1991), Ahn and Schmidt (1995), Alvarez and Arellano (2003), and Blundell and Bond (1998), and we extend the framework to accommodate more general models. In addition, we directly address weak identification concerns noted in the dynamic panel literature, particularly in Bun and Windmeijer (2010). We also contribute to the emerging literature on quadratic central limit theorems for clustered data structures. These results have recently been employed in many-instrument panel settings, such as Ligtenberg

(2023), and are crucial for understanding the distribution of estimators that are quadratic functionals of the errors.

The remainder of the paper is organized as follows. Section 2 describes the data structure, a plausible set of exclusion restrictions, and two modeling perspectives—outcome-based and design-based. Section 3 characterizes correctly centered internal instrument estimators, introduces our proposed estimator as the solution to an efficiency optimization problem, and interprets it as a leave-out internal instrument procedure. Section 4 addresses uncertainty quantification, and Section 5 establishes a central limit theorem for quadratic forms with clustered data. Section 6 develops inference procedures robust to weak identification and a corresponding variance estimator. Section 7 illustrates the full strategy in an empirical application, showing how both the estimator and, especially, its uncertainty depend on the maintained set of assumptions.

Notation. We use $0 < c < 1$ and $C > 0$ to denote generic finite constants that may vary across equations but do not depend on the sample size. For a matrix A, we let A^+ denote its Moore–Penrose generalized inverse, $\operatorname{tr}(A)$ its trace, $\|A\|_F^2 = \operatorname{tr}(A'A)$ its Frobenius norm squared, and $\|A\|$ its operator (spectral) norm.

2 Exclusion restrictions with clustered data

2.1 Data structure

We consider the estimation of a structural parameter β in the linear regression model

$$y_\ell = x_\ell \beta + w_\ell' \delta + e_\ell, \quad \text{for } \ell = 1, \ldots, n, \tag{1}$$

where ℓ indexes the n observations in the dataset. The parameter of interest, β, is a scalar, though most of our results extend to vectors of fixed dimension. The vector of strictly exogenous controls w_ℓ has dimension K, which may be large but is assumed to satisfy $K < n$. We denote by $W = [w_1, \ldots, w_n]'$ the matrix of control variables. For simplicity, we treat W as non-random; equivalently, all results may be interpreted conditional on W. Throughout, we assume that W has full rank and define the projection matrix $M = I_n - W(W'W)^{-1}W'$, which projects out the variation associated with the controls W.

In contrast to standard cross-sectional settings with independent observations, we allow for dependence within clusters. Specifically, we assume the data can be partitioned into N disjoint clusters, with independence across clusters and arbitrary dependence within them. Formally, let $\{S_i\}_{i=1}^N$ be a partition of the index set $\{1, \ldots, n\}$, so that $\cup_{i=1}^N S_i = \{1, \ldots, n\}$ and $S_i \cap S_j = \emptyset$ for $i \neq j$. Let $i(\ell)$ denote the index of the cluster containing observation ℓ, and define $T_i = |S_i|$ as the size of cluster i, so that $n = \sum_{i=1}^N T_i$. This cluster structure reflects the sampling design and is essential for the application of laws of large numbers and central limit theorems.

Example 1 (Unbalanced panel data)**.** Consider a regression with individual and time fixed effects: $y_{it} = \alpha_i + \mu_t + \beta x_{it} + e_{it}$, where y_{it} and x_{it} denote the outcome and treatment for student i at time t. Let $\ell = \ell(i,t)$ index the observations, and define clusters as $S_i = \{\ell(i,t) : t \text{ observed for individual } i\}$. Students may be observed across different, potentially non-overlapping periods. Here, following the standard panel-data setup, clusters coincide with individuals indexed by i. In other applications, however, clusters may be defined more broadly—for example, at the family level when students are siblings. □

Example 2 (Spatial data)**.** Suppose we study the causal effect of a municipality-level policy x_ℓ on outcome y_ℓ. Assume that municipalities are nested within metropolitan areas indexed by i, which serve as the clusters. A rich set of controls is included—such as cluster fixed effects and federally influenced covariates—so that $\{(x_\ell, e_\ell) : \ell \in S_i\}$ is independent across clusters, but arbitrary dependence is allowed within clusters. □

Example 3 (Network data)**.** Consider treatment and outcome data for individuals connected via a social or informational network. The data may be drawn either from distinct clusters within a large network or from multiple independent networks. Conditional on a comprehensive set of controls, we assume independence across clusters, while allowing for complex, unrestricted dependence within each cluster. □

2.2 Exclusion restrictions

Assume that random regressors x_ℓ and errors e_ℓ have uniformly bounded second moments. There are a variety of exclusion assumptions one may make in regression (1). The weakest of these is the contemporaneous exclusion restriction, $\mathbb{E}[x_\ell e_\ell] = 0 = \mathbb{E}[e_\ell]$ for all ℓ, which defines the parameter β. When all observations are independent, this restriction (under mild regularity conditions) ensures that ordinary least squares (OLS) yields a consistent estimator of β. However, as we show later, in the presence of clustered data, when the regression includes cluster fixed effects, the contemporaneous exclusion assumption alone fails to yield a consistent estimator of β.

At the other extreme is the *strict exogeneity* condition: $\mathbb{E}[e_\ell \mid x] = 0$ where $x = [x_1, \ldots, x_n]'$, which implies $\mathbb{E}[x_{\tilde{\ell}} e_\ell] = 0$ for all $\tilde{\ell}$ and ℓ. Under strict exogeneity, OLS is unbiased conditional on x. In particular, by the Frisch-Waugh-Lovell theorem, we have:

$$\hat{\beta}^{\mathrm{LS}} = \frac{x'My}{x'Mx} = \beta + \frac{\tilde{x}'e}{\tilde{x}'\tilde{x}},$$

where $y = [y_1, \ldots, y_n]'$, $e = [e_1, \ldots, e_n]'$ and $\tilde{x} = Mx$ denotes the residual from regressing x on W. Conditionally on x, the regression errors $\{e_\ell\}$ are the only source of randomness. Since $\hat{\beta}^{\mathrm{LS}}$ is linear in the errors, standard results—such as unbiasedness, the law of large numbers, and the central limit theorem—hold. The only complication introduced by the clustered data structure is the need to adjust standard errors appropriately: cluster-robust standard errors are required because the relevant level of independence is at the cluster level.

Strict exogeneity is often too strong to hold in applications with clustered data, where unmodeled dependence between regressors and outcomes within clusters is likely. Researchers should tailor the exclusion restrictions to reflect institutional or structural features of the setting. We assume that the researcher can specify a matrix of exclusion restrictions, denoted by $\mathcal{E}$, which is an $n \times n$ indicator matrix encoding the moment conditions the researcher is willing to impose. Specifically, $\mathcal{E}_{\tilde{\ell}\ell} = 1$ encodes the assumption $\mathbb{E}[x_{\tilde{\ell}} e_\ell] = 0$, while $\mathcal{E}_{\tilde{\ell}\ell} = 0$ encodes the absence of a restriction and allows $\mathbb{E}[x_{\tilde{\ell}} e_\ell] \neq 0$.

Due to the assumed independence across clusters, we automatically have $\mathbb{E}[x_{\tilde{\ell}} e_\ell] = 0$ whenever $i(\ell) \neq i(\tilde{\ell})$. For notational simplicity, we assume $\mathcal{E}_{\tilde{\ell}\ell} = 1$ in such cases. Thus, zero entries in $\mathcal{E}$ may only occur within the blocks corresponding to clusters. Zeros within these blocks represent the absence of exclusion restrictions. We consider the problem of estimating β in regression (1) under the exclusion restrictions encoded in the indicator matrix $\mathcal{E}$.

Assumption 1. *Assume* $\mathbb{E}[e_\ell] = 0$, $\mathcal{E}_{\ell\ell} = 1$ *for all* ℓ, *and* $\mathbb{E}[x_{\tilde{\ell}} e_\ell] = 0$ *whenever* $\mathcal{E}_{\tilde{\ell}\ell} = 1$.

Assumption 1 ensures contemporaneous exogeneity, $\mathbb{E}[x_\ell e_\ell] = 0$, which underlies the desirable properties of OLS in standard cross-sectional settings.

Example 1 (continued)**.** Suppose a researcher is interested in estimating the effect of program participation x_{it} on student achievement y_{it}. It is plausible that current achievement y_{it} could influence future eligibility or desire to participate x_{is} for $s > t$, thereby inducing correlation between the current error term e_{it} and future regressors x_{is}, violating strict exogeneity. In such settings, it is common to assume *weak exogeneity*: $\mathbb{E}[x_{is}e_{it}] = 0$ for $s \leq t$, meaning the error is uncorrelated with current and past regressors but may influence future regressors. This structure arises naturally in dynamic panel models when, for instance, the lagged outcome is included as a regressor, $x_{it} = y_{i,t-1}$, or more generally when regressors and outcomes jointly evolve over time. In this case, the matrix $\mathcal{E}$ has triangular blocks of zeros if the observations are ordered by clusters and by time within each cluster.

If clustering is defined at the family level, the researcher may also be concerned about *sibling feedback* effects. For example, if an older child's experience with the program affects a younger sibling's participation. In this case, the researcher may refrain from assuming exclusion restrictions across such pairs, encoding this by zeros in the corresponding entries in $\mathcal{E}$, linking the older sibling's error term with the younger sibling's regressor. □

Example 2 (continued)**.** Suppose a researcher is interested in estimating the average treatment effect of a policy using a randomized controlled trial, in which an economic resource is randomly assigned across municipalities ℓ. Spatial interference arises when outcomes in a given location may also be affected by treatment in neighboring locations—for example, because of resource sharing or price adjustments in shared local markets. Such spillover effects violate the exogeneity assumption across neighboring localities, since treatment in one location directly affects outcomes in the neighboring locality. If the researcher believes that interference occurs only when locations are within R kilometers of each other and is negligible at larger distances, then the matrix $\mathcal{E}$ encodes indicators of whether two locations lie more than R kilometers apart. Notable examples of such settings are spatial leakage in deforestation (Jayachandran et al., 2017), spillovers in policing (Blattman et al., 2021), or a fiscal intervention experiment in rural Kenya (Egger et al., 2022). □

Example 3 (continued)**.** Consider the randomized controlled trial studied in (Paluck et al., 2016), in which students are randomly assigned to participate in a program that teaches teenagers conflict-resolution skills. A researcher may believe that the treatment affects not only the outcomes of treated students, but also those of their close friends. In this case, it is reasonable to assume that an exclusion restriction holds only for pairs of observations that are not directly connected by friendship, so that the matrix $\mathcal{E}$ is given by the complement of the friendship adjacency matrix. Such spillovers of information or resources through social networks are increasingly common in modern randomized controlled trials. Prominent examples include Banerjee et al. (2013), Banerjee et al. (2019), and Banerjee et al. (2024). □

2.3 Asymptotic bias of OLS

When strict exogeneity fails, the OLS estimator may suffer from asymptotic bias and, in some cases, may even be inconsistent.

Lemma 1. *Suppose assumptions 1 and 6 (stated in the Appendix) hold. Then,*

$$\hat{\beta}^{\mathrm{LS}} = \beta + \frac{\frac{1}{n}\sum_{\ell=1}^{n}\sum_{\tilde{\ell}=1}^{n} M_{\tilde{\ell}\ell}\mathbb{E}[x_{\tilde{\ell}}e_{\ell}]}{Q} + o_p(1),$$

where $M_{\tilde{\ell}\ell}$ are entries of the $n \times n$ projection matrix M, and $Q = \operatorname{plim}_{n\to\infty} \frac{1}{n}\sum_{\ell=1}^{n} \tilde{x}_{\ell}^2$.

This lemma is a relatively straightforward extension of the well-known Nickell bias (Nickell, 1981). Nickell's original work focused on a more restricted setting commonly referred to as a "dynamic panel data" model.

Example 4 (Dynamic panel data)**.**

$$y_{it} = \alpha_i + \beta y_{i,t-1} + e_{it}. \tag{2}$$

This is a special case of our Example 1 with the weak exogeneity restriction $\mathbb{E}[e_{it} | \{y_{is} : s < t\}] = 0$ for all i, t. Consider the simplest case of a balanced panel with $T_i = T$ for all i. The projection matrix M corresponds to demeaning within units, and thus has a block-diagonal structure when observations are ordered by cluster. Under homoskedasticity with error variance σ^2, $\mathbb{E}[x_{\tilde{\ell}}e_{\ell}] = \mathbb{E}[y_{is}e_{it}] = \beta^{s-t}\sigma^2$ when $s \geq t$. Substituting these expressions into Lemma 1 yields the bias formula originally derived in Nickell (1981). It is well known that in this setting, OLS is inconsistent for fixed T, as the bias does not vanish. The leading term of the bias is $O(1/T)$, and if T grows slower than the number of clusters N, the bias may dominate the variance asymptotically, resulting in invalid inference. □

The source of Nickell's bias is well understood: differencing out fixed effects mixes observations from different periods, and future values of the regressor may correlate with current errors. Our framework is more general than the standard dynamic panel specification in (2). It is worth emphasizing that the result established by Nickell (1981)—and generalized in Lemma 1—is best understood as an asymptotic bias: a term which, when subtracted from the OLS estimator, yields a consistent (or asymptotically unbiased) estimator. Asymptotic bias is distinct from finite sample bias, as it ignores the randomness in the denominator of the OLS expression. While this randomness becomes negligible asymptotically, it is present in finite samples.

2.4 Design based modeling

Model (1) combined with Assumption 1 can be viewed as an instance of *outcome modeling*, since it specifies the outcome equation. In cross-sectional settings, however, OLS estimators perform well when either the outcome or the treatment equation is correctly specified. The alternative approach—based on the treatment equation—is referred to as *design-based* and is commonly used when researchers understand the random assignment of treatment, such as in randomized controlled trials (RCTs). In this paper, we define a design-based model by specifying the following treatment equation, along with Assumption 2:

$$x_\ell = w_\ell' \delta_x + v_\ell. \tag{3}$$

Assumption 2. *Assume* $\mathbb{E}[v_\ell] = 0$*,* $\mathcal{E}_{\ell\ell} = 1$ *for all* ℓ*, and* $\mathbb{E}[v_{\tilde{\ell}}(y_\ell - \beta x_\ell)] = 0$ *whenever* $\mathcal{E}_{\tilde{\ell}\ell} = 1$*.*

Example 3 (continued)**.** Suppose a researcher wishes to estimate the effect of access to resources on individual outcomes by conducting an RCT in which individuals ℓ are randomly provided with additional resources. The researcher assumes that the treatment assignment equation is correctly specified as $x_\ell = w_\ell' \delta_x + v_\ell$, where $w_\ell' \delta_x$ denotes the propensity score, which may depend on individual characteristics as well as cluster fixed effects.

Now suppose there are reasons to expect *interference*—that is, an individual's outcome may be influenced by the treatment assignment of their friends, for example, due to the sharing of resources within social networks. The researcher observes the friendship network and randomizes treatment independently of it. The outcome equation is posited as

$$y_\ell = \beta x_\ell + g(W, F_\ell' x) + e_\ell,$$

where F_ℓ encodes the friendship links relevant for individual ℓ, so that $F_\ell' x$ is the vector of treatment assignments for peers who may influence individual ℓ. The error term e_ℓ is assumed to be independent of all assignments x, but may be correlated within clusters to capture common shocks (e.g., village-level effects). In this case, the outcome modeling approach is complicated by the need to specify the functional form of the indirect effect. This effect might depend on whether *any* friends are treated, the *number* of treated friends, or the unobserved *intensity* of friendships. The quantity $g(W, F_\ell' x)$ may vary systematically within clusters, and this variation may not be well captured by fixed effects or observed covariates. Consequently, if the indirect effect is absorbed into the unmodeled regression error term, the outcome model (1) may be misspecified.

Nonetheless, design-based assumptions remain valid for pairs ℓ and $\tilde{\ell}$ who are not friends:

$$\mathbb{E}[v_{\tilde{\ell}}(y_\ell - \beta x_\ell)] = \mathbb{E}[v_{\tilde{\ell}}(g(W, F_\ell' x) + e_\ell)] = \mathbb{E}[v_{\tilde{\ell}}] \cdot \mathbb{E}[g(W, F_\ell' x) + e_\ell] = 0,$$

by the independence of treatment assignment. In this context, the exclusion restriction matrix $\mathcal{E}$ is the complement of the adjacency matrix of the friendship graph: it has zeros for direct friends and ones elsewhere. □

The asymptotic bias expression from Lemma 1 can equivalently be restated in the design-based framework:

$$\hat{\beta}^{\text{LS}} = \frac{x'My}{x'Mx} = \frac{v'My}{v'Mx} = \beta + \frac{v'M(y - \beta x)}{v'Mx} = \beta + \frac{\frac{1}{n}\sum_{\ell,\tilde{\ell}=1}^{n} M_{\tilde{\ell}\ell}\mathbb{E}[v_{\tilde{\ell}}(y_\ell - \beta x_\ell)]}{Q} + o_p(1).$$

Figure 1 illustrates this asymptotic bias in a simulation designed to mimic the network-interference setting in Example 3. We consider a DGP based on a stochastic block model where clusters represent school classes and within-cluster edges represent friendships. The re-

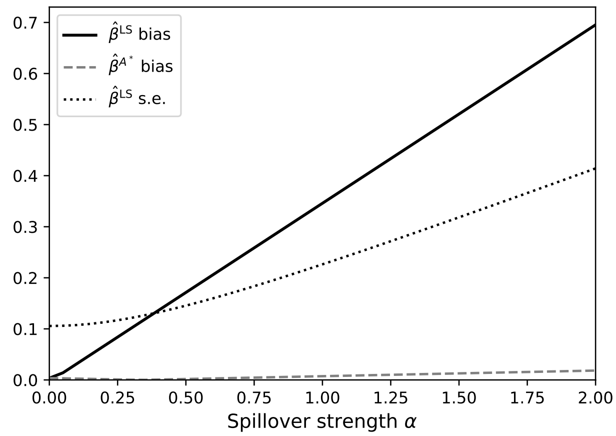

Figure 1: Bias of $\hat{\beta}^{\text{LS}}$ in simulations

Notes. This figure presents the absolute bias of $\hat{\beta}^{\text{LS}}$ (solid line) and its cluster-robust standard error (dotted line) in a design-based DGP with $n = 500$ and varying spillover strength α. The absolute bias of the proposed estimator $\hat{\beta}^{A^*}$ (described in Theorem 1, dashed line) is shown as a benchmark.

searcher observes the friendship network but not the intensity of each connection, denoted as weight $G_{\tilde{\ell}\ell} = G_{\ell\tilde{\ell}} \sim \text{Exp}(1)$ for connected pairs and zero otherwise. Treatment is binary and assigned independently across units with probability $\mu_{i(\ell)}$, which varies across clusters. The outcome y_ℓ includes a spillover component depending on the treatment of friends weighted by unobserved friendship intensities: $y_\ell = \beta\, x_\ell + \alpha \sum_{\tilde{\ell}} G_{\ell\tilde{\ell}}\, x_{\tilde{\ell}} + e_\ell$. The least-squares estimator controlling for cluster fixed effects exhibits substantial bias in this setting. Full details of the simulation design are provided in Appendix B. Despite random assignment of treatment, the OLS estimator can be highly misleading when interference is present. The bias arises from the correlation induced by spillovers within clusters, illustrating the Nickell-type bias in clustered regressions with network interference.

The simulation illustrates the usefulness of the design-based perspective in settings where the treatment assignment mechanism is credible but the outcome equation may be difficult to specify. The primary exposition in this paper is framed within the outcome modeling approach. However, all results can equivalently be reformulated under the design-based perspective with only minor modifications. We provide such restatements in the remarks following the main theorems. Additionally, where appropriate, we discuss the possibility of a *doubly robust* formulation, in which the researcher assumes that either the outcome equation or the treatment equation is correctly specified, without committing to which one.

3 Correctly centered estimators

3.1 Class of estimators

This subsection shows that unbiased estimation is generally impossible once regressors are random, even under standard orthogonality conditions, and motivates a weaker requirement that we call correct centering. Assume we have clustered data $\{y_\ell, x_\ell\}$ generated according to model (1) under Assumption 1. The set of controls W, the cluster assignments S_i, and the exclusion restriction matrix $\mathcal{E}$ are given to the econometrician and treated as fixed. We consider the problem of estimating the parameter β. Let $\mathcal{F}$ denote the class of all joint distributions F over (x, y) that satisfy model (1) under Assumption 1.

Lemma 2. *There exists no estimator $\hat{\beta} = u(x, y)$ that is unbiased for all $F \in \mathcal{F}$.*

In fact, the proof of Lemma 2 establishes something even stronger: in a cross-sectional setting with an independent sample $\{y_i, x_i\}_{i=1}^n$ from a correctly specified regression model $y_i = \beta x_i + e_i$, where the regressor x_i is random and $\mathbb{E}[e_i] = \mathbb{E}[x_i e_i] = 0$, no unbiased estimator of β exists. The OLS estimator is not unbiased in this setting because the randomness of the regressors induces randomness in the OLS denominator, making it impossible to pass the expectation operator through the ratio. This observation motivates the need to consider a broader and more appropriate class of estimators, in which the normalization is separated from the centering.

Definition 1. An estimator $\hat{\beta} = \frac{C_1(x,y)}{C_2(x)}$ is said to be *correctly centered* if for all $F \in \mathcal{F}$:

$$\mathbb{E}_F[C_1(x, y)] = \beta\, \mathbb{E}_F[C_2(x)].$$

Correct centering is a weaker condition than unbiasedness. The key distinction lies in the treatment of randomness in the denominator. When the denominator is non-random—so that $\mathbb{E}[C_2(x)] = C_2(x)$—the two concepts coincide. In a correctly specified cross-sectional regression with random regressors, OLS is correctly centered but not unbiased.

For many of the estimators we consider, we can establish a form of the law of large numbers (under broadly applicable technical conditions) that ensures $C_2(x) - \mathbb{E}[C_2(x)] \xrightarrow{p} 0$

and $C_1(x,y) - \mathbb{E}[C_1(x,y)] \xrightarrow{p} 0$—with any necessary normalization absorbed into the functions themselves. For instance, Lemma 1 establishes that $C_2(x) = \frac{1}{n}x'Mx = \frac{1}{n}\mathbb{E}[x'Mx] + o_p(1)$ and $C_1(x,y) = \frac{1}{n}\mathbb{E}[x'My] + o_p(1)$. Regarding the denominator, what matters is the size of its uncertainty relative to its signal—that is, the strength of identification. If $\frac{C_2(x)}{\mathbb{E}[C_2(x)]} \xrightarrow{p} 1$ and the above law of large numbers holds, then correct centering implies consistency (or asymptotic unbiasedness) of the estimator.

Lemma 3. *The OLS estimator is not correctly centered in the outcome model* (1) *with Assumption 1 or in the design-based model* (3) *with Assumption 2.*

To better understand the concept, it is helpful to recall that in a standard cross-sectional overidentified IV regression, the two-stage least squares (2SLS) estimator is not correctly centered. When the number of instruments is large, the resulting bias can be of first-order importance and is known as the many-instrument problem. In contrast, a one-step just-identified IV estimator—as well as jackknife or leave-one-out IV estimators in overidentified cross-sectional settings—remains correctly centered whenever the instruments are exogenous.

Lemma 4. *Consider an estimator $\hat{\beta}^A = \frac{x'Ay}{x'Ax}$, where A is an $n \times n$ fixed matrix.*[1] *It is correctly centered if and only if A satisfies the following conditions:*

$$\text{(POP) } AM = A \quad \textit{and} \quad \text{(CC) } A_{\tilde{\ell}\ell} = 0 \textit{ for all pairs } (\tilde{\ell}, \ell) \textit{ such that } \mathcal{E}_{\tilde{\ell}\ell} = 0.$$

Here, (POP) stands for the *partialling-out property*, and (CC) refers to *correct centering*. Under condition (POP), the estimator satisfies:

$$\hat{\beta}^A = \frac{x'Ay}{x'Ax} = \frac{x'A(\beta x + W\delta + e)}{x'Ax} = \beta + \frac{x'Ae}{x'Ax}.$$

Condition (CC) guarantees that $\mathbb{E}[x'Ae] = 0$, ensuring correct centering of the estimator.

The class of estimators described in Lemma 4 coincides with the class of just-identified linear IV estimators that use linear internal instruments. Specifically, define the instrument

[1] Not every correctly centered estimator has the form $\hat{\beta}^A$, or is even linear in y. The independence between clusters imposes strong restrictions on the joint distribution of the data and allows higher-order U-statistics to be added to the numerator. For instance, if $(x_i, y_i),\ i = 1, \ldots, 4$ are observations from distinct (and thus independent) clusters, and $y_i = \beta x_i + e_i$, then $\mathbb{E}\big[x_1x_2y_3y_4 - y_1y_2x_3x_4\big] = 0$. This observation connects to recent discussions in Hansen (2022) and Portnoy (2022); Lei and Wooldridge (2022).

$z = A'x$, a linear transformation of the regressors. Condition (CC) implies the exogeneity condition $\mathbb{E}[z_\ell e_\ell] = 0$ for all ℓ. If we estimate model (1) via a just-identified IV regression using z as the instrument for x, and including W as controls, then a straightforward generalization of the Frisch-Waugh-Lovell theorem for IV estimators yields:

$$\hat{\beta}^{\mathrm{IV}} = \frac{z'My}{z'Mx} = \frac{x'AMy}{x'AMx} = \frac{x'Ay}{x'Ax} = \hat{\beta}^{A},$$

where the final equality follows from the partialling-out property (POP).

Remark 1. If the data are generated from the design-based model (3) under Assumption 2, then the results of Lemma 4 continue to hold with a slightly modified (POP) condition. Specifically, we must replace (POP) with the following: (POP′) $MA = A$. In the case of the doubly robust model, the corresponding condition becomes: (POP″) $MA = AM = A$. □

Our approach is closely related to methods based on internal instruments for dynamic panel data, such as Anderson and Hsiao (1981), Arellano and Bond (1991), and Ahn and Schmidt (1995). Both Arellano and Bond (1991) and Ahn and Schmidt (1995) stack all available exclusion restrictions and estimate the model using overidentified 2SLS or GMM. In contrast, we construct a just-identified linear combination of internal instruments and implement a one-step IV procedure. This delivers a correctly centered estimator. Overidentified procedures generally do not enjoy this property, as the combination of many moment conditions together with an estimated weighting matrix introduces additional randomness. When the degree of overidentification is large, this can result in substantial many-instrument bias (Alvarez and Arellano, 2003). Different choices of the matrix A satisfying restrictions (CC) and (POP) span a broad class of linear internal instruments and, in particular, include the infeasible optimally weighted instrument of Arellano and Bond (1991) as a special case.

3.2 Asymptotic efficiency considerations

Let $\mathcal{A}$ denote the set of $n \times n$ matrices satisfying conditions (POP) and (CC). This set forms a linear subspace within the space of $n \times n$ matrices. The full space of $n \times n$ matrices has dimension n^2. Condition (POP), which is equivalent to $AW = 0_{n\times K}$, imposes nK linear constraints. Condition (CC) imposes an additional L zero restrictions, where L is the number of entries $A_{\tilde{\ell}\ell}$ required to be zero due to (CC). Therefore, the dimension of $\mathcal{A}$ is at least $n(n-K)-L$. If the dimension of $\mathcal{A}$ is positive, then our estimator class contains infinitely many candidates, raising the natural question: Which matrix A should we use? As a default, we propose the choice

$$A^* = \arg\min_{A\in\mathcal{A}} \|A - I_n\|_F = \arg\min_{A\in\mathcal{A}} \|A - M\|_F. \tag{4}$$

The equality holds due to condition (POP). The Frobenius norm defines a Euclidean geometry on the space of matrices, with the corresponding inner product given by $\langle A, B\rangle_F = \mathrm{tr}(A'B)$. The optimization problem (4) admits a unique solution: the orthogonal projection of M onto the subspace $\mathcal{A}$ with respect to the Frobenius inner product. In the lemma below, we show that this choice of A^* can be motivated by asymptotic efficiency considerations.

Lemma 5. *Suppose Assumption 1 holds and assume that $\mathbb{E}[ee'\,|\,x] = \sigma^2 I_n$ and $\mathbb{E}[xx'] = \lambda I_n$ for some constants $\sigma^2, \lambda > 0$. Consider the function*

$$V(A) = \frac{\mathbb{V}(x'Ae)}{\left(\mathbb{E}[x'Ax]\right)^2} = \frac{\mathbb{V}(z'e)}{\left(\mathbb{E}[z'x]\right)^2}$$

where $z = A'x$. Then the minimum of $V(A)$ over the set $\mathcal{A}$ is attained at A^*.

The function $V(A)$ represents the asymptotic variance of the IV estimator under strong identification, i.e., when $x'Ax/\mathbb{E}[x'Ax] \xrightarrow{p} 1$. This efficiency criterion follows the classical approach of Chamberlain (1987). Lemma 5 shows that, under ideal conditions, our proposed estimator using A^* from (4) achieves asymptotic efficiency within the class of correctly centered estimators. The two ideal conditions are: first, conditional homoskedasticity, $\mathbb{E}[ee' \mid x] = \sigma^2 I_n$, which implies that the optimal instrument z^* minimizes the first stage prediction error over all $z = A'x$ with $A \in \mathcal{A}$:

$$(A^*)'x = z^* = \arg\min_{z=A'x, A\in\mathcal{A}} \mathbb{E}\left[\|z - x\|^2\right]. \tag{5}$$

Second, the assumption $\mathbb{E}[xx'] = \lambda I_n$ imposes stochastic homogeneity of the regressors and ensures the equivalence between the optimization problems (4) and (5).

It is natural that the efficiency of an estimator depends on the distributional properties of the regressor x, especially in settings like ours where x must be treated as random. Any prior information—such as the scale of x or its dynamic dependence—can, in principle, be used to improve efficiency, for example, through cluster-specific weighting or time transformations (as in GLS). However, caution is needed: the matrix A must remain independent of the randomness in either x or e to avoid bias, such as the many-instrument bias observed in the two-step GMM estimator of Arellano and Bond (1991). If additional structure is available, such as $\mathbb{E}[xx'] = \Omega$, then asymptotic efficiency is attained by solving:

$$A^*_\Omega = \arg\min_{A\in\mathcal{A}} \operatorname{tr}\left((A - I_n)'\Omega(A - I_n)\right) = \arg\min_{A\in\mathcal{A}} \|\Omega^{1/2}(A - I_n)\|_F.$$

Most results in this paper extend naturally to this weighted setting, except for the leave-one-out projection characterization discussed below in Theorem 1(iii).

3.3 Solution to the optimization problem

This subsection derives and discusses the solution to the optimization problem (4).

We begin by introducing what we refer to as leave-out projections. For each index $\tilde{\ell}$, define $\mathcal{E}_{\tilde{\ell},\cdot}$ to be the $\tilde{\ell}$th row of the matrix $\mathcal{E}$, which contains indicators for those indices ℓ such that $\mathbb{E}[x_{\tilde{\ell}}e_\ell] = 0$. In other words, $\mathcal{E}_{\tilde{\ell},\cdot}$ identifies all observations for which $x_{\tilde{\ell}}$ is uncorrelated with their error term. Let $\mathbf{i}_{\tilde{\ell}} = \operatorname{diag}(\mathcal{E}_{\tilde{\ell},\cdot})$ denote the $n \times n$ diagonal selection matrix with ones on the diagonal corresponding to entries where $\mathcal{E}_{\tilde{\ell}\ell} = 1$, and zeros elsewhere. Then define $W_{\tilde{\ell}} = \mathbf{i}_{\tilde{\ell}}W$, an $n \times K$ matrix that retains the rows w_ℓ for which $\mathcal{E}_{\tilde{\ell}\ell} = 1$, and sets the remaining rows to zero. The corresponding leave-out projection matrix is given by $M^{(\tilde{\ell})} = I - W_{\tilde{\ell}}(W'_{\tilde{\ell}}W_{\tilde{\ell}})^+W'_{\tilde{\ell}}$, which partials out controls using only observations whose errors are uncorrelated with $x_{\tilde{\ell}}$. Unlike the full projection M, it excludes data points that violate exogeneity.

Theorem 1. *Assume W has full rank. Let A^* be the solution to the optimization problem (4). Then, A^* admits the following three equivalent characterizations:*

(i) Let the linear restrictions in (CC) be written as $\mathcal{L}vec(A) = 0$, then

$$vec(A^*) = (I_{n^2} - \mathcal{L}'(\mathcal{L}\mathcal{L}')^+\mathcal{L})vec(M).$$

(ii) There exists an $n \times n$ matrix B satisfying condition $B_{\tilde{\ell}\ell} = 0$ whenever $\mathcal{E}_{\tilde{\ell}\ell} = 1$ such that $A^* = (I_n - B)M$.

(iii) $A^*_{\tilde{\ell}\ell} = M^{(\tilde{\ell})}_{\tilde{\ell}\ell}$, *if* $\mathcal{E}_{\tilde{\ell}\ell} = 1$, *and* $A^*_{\tilde{\ell}\ell} = 0$, *otherwise.*

Characterization (i) in Theorem 1 highlights the analytical tractability of A^* by noting its connection with linear projection.

Characterization (ii) shows that A^* closely resembles the OLS projection matrix M, with the difference given by $A^* - M = -BM$, where B is a sparse matrix. Specifically, $B_{\tilde{\ell}\ell}$ is nonzero only if $\mathbb{E}[x_{\tilde{\ell}} e_\ell] \neq 0$. As a result, B is block-diagonal (with blocks corresponding to clusters) and has zeros on the main diagonal. Thus, A^* can be computed separately within each cluster, which is computationally appealing.

Finally, part (iii) of Theorem 1 offers an intuitive interpretation of the solution. For each observation $\tilde{\ell}$, we perform an observation-specific transformation of the original model (1) by partialling out the control $w_{\tilde{\ell}}$ using only observations uncorrelated with $x_{\tilde{\ell}}$. Define the residualized outcome as $y^*_{\tilde{\ell}} = y_{\tilde{\ell}} - w'_{\tilde{\ell}}\hat{\delta}^y_{\tilde{\ell}}$, where $\hat{\delta}^y_{\tilde{\ell}}$ is obtained from regressing y on w using only observations with errors uncorrelated with $x_{\tilde{\ell}}$. Similarly define $x^*_{\tilde{\ell}}$ and $e^*_{\tilde{\ell}}$. These definitions yield the transformed equation:

$$y^*_{\tilde{\ell}} = \beta x^*_{\tilde{\ell}} + e^*_{\tilde{\ell}}. \tag{6}$$

Since $e^*_{\tilde{\ell}}$ is constructed using only observations uncorrelated with $x_{\tilde{\ell}}$, we have $\mathbb{E}[x_{\tilde{\ell}} e^*_{\tilde{\ell}}] = 0$. Hence, $x_{\tilde{\ell}}$ is a valid instrument in (6), and the IV estimator is $\hat{\beta} = \frac{\sum_{\tilde{\ell}} x_{\tilde{\ell}} y^*_{\tilde{\ell}}}{\sum_{\tilde{\ell}} x_{\tilde{\ell}} x^*_{\tilde{\ell}}} = \frac{x'A^*y}{x'A^*x} = \hat{\beta}^{A^*}$.

Example 5 (Panel data with fixed effects)**.** Consider the model

$$y_{it} = \alpha_i + \beta x_{it} + e_{it}, \text{ with } \mathbb{E}[e_{it} | x_{it}, x_{i,t-1}, \ldots] = 0,$$

where the error is unpredictable given current and past values of the regressor, but may affect future values. In this case, the optimal matrix A^* has a block-diagonal structure, with each block corresponding to a cluster. Our estimator transforms the data by demeaning y_{it} and x_{it} using only current and future observations:

$$y^*_{it} = y_{it} - \frac{1}{T_i - t + 1}\sum_{s=t}^{T_i} y_{is}, \quad x^*_{it} = x_{it} - \frac{1}{T_i - t + 1}\sum_{s=t}^{T_i} x_{is}$$

and estimates the regression $y^*_{it} = \beta x^*_{it} + e^*_{it}$ using x_{it} as an instrument. This estimator lies within the class proposed by Arellano and Bond (1991), but applies more broadly. See also Hayashi and Sims (1983); Arellano and Bover (1995). □

Example 6 (Panel data with fixed effects with a limited time feedback)**.** Consider a setting

$$y_{it} = \alpha_i + \beta x_{it} + e_{it}, \text{ with } \mathbb{E}[e_{it} \,|\, \{x_{is}, s \neq t+1\}] = 0.$$

Unlike Example 5, the error e_{it} may affect the next period's regressor $x_{i,t+1}$ but not beyond; feedback is limited to one period. In this case, the optimal matrix A^* again has a block-diagonal structure across clusters. For each observation, we demean y_{it} and x_{it} using all time periods except $t-1$, which may be endogenous:

$$y^*_{it} = y_{it} - \frac{1}{T_i - 1} \sum_{s \neq t-1} y_{is}, \quad x^*_{it} = x_{it} - \frac{1}{T_i - 1} \sum_{s \neq t-1} x_{is}, \text{ for } t > 1,$$

and for $t = 1$, we use the full mean: $y^*_{i1} = y_{i1} - \frac{1}{T_i} \sum_{s=1}^{T_i} y_{is}$, $x^*_{i1} = x_{i1} - \frac{1}{T_i} \sum_{s=1}^{T_i} x_{is}$. We then estimate the regression $y^*_{it} = \beta x^*_{it} + e^*_{it}$ using x_{it} as the instrument. □

Price of robustness. Our estimator can be seen as a robust alternative to OLS when strict exogeneity, $\mathbb{E}[e_\ell \,|\, x] = 0$, is not credible. The exclusion matrix $\mathcal{E}$ determines the types of violations it is robust to: Example 6 allows $\mathbb{E}[x_{i,t+1} e_{it}] \neq 0$, while Example 5 allows broader future dependence. A natural question is whether this robustness reduces efficiency when strict exogeneity holds. Under the assumptions of Lemma 5, we have $V(A^*) = \frac{\sigma^2}{\lambda \operatorname{tr}(A^*)} = \frac{\sigma^2}{\lambda \|A^*\|_F^2}$, where $\operatorname{tr}(A^*)$ reflects the effective sample size. In the standard panel model with strict exogeneity and homoskedasticity, OLS achieves the efficiency bound with $\operatorname{tr}(M) = \sum_{i=1}^N (T_i - 1) = n - N$, as one degree of freedom is lost per cluster due to demeaning. Each block M_{ii} contributes $\operatorname{tr}(M_{ii}) = T_i - 1$.

In Example 6, we have $\operatorname{tr}(A_{ii}) = T_i - 1 - \frac{1}{T_i}$; the term $-\frac{1}{T_i}$ reflects the loss of information due to robustness against one-period feedback. In Example 5, $\operatorname{tr}(A_{ii}) = T_i - 1 - \sum_{t=2}^{T_i} \frac{1}{t}$. With $\sum_{t=2}^{T_i} \frac{1}{t} \geq \log(T_i + 1) - \log(2)$, this shows a substantially greater efficiency loss when guarding against dependence on any future regressor. Finally, in a panel model with fixed effects, assuming contemporaneous exogeneity only $\mathbb{E}[x_\ell e_\ell] = 0$ leads to $A^* = 0$. In this case, the model contains no usable identifying variation.

Remark 2. If one considers the design-based model (3) under Assumption 2, then a minor modification of Theorem 1, part (iii), is required. The procedure begins by constructing a proxy for v_ℓ, which is obtained as the ℓth residual from regression (3), estimated on a subsample consisting only of observations $\tilde{\ell}$ for which $\mathcal{E}_{\tilde{\ell}\ell} = 1$. This residual is then used as an instrument in a just-identified instrumental variables (IV) regression of y_ℓ on x_ℓ. □

4 Quantification of uncertainty

4.1 Importance of cross-dependence

Let x_i denote the $T_i \times 1$ vector of regressor values within cluster S_i—that is, all x_ℓ for which $i(\ell) = i$—and define e_i analogously. Due to the cluster structure, the pairs (x_i, e_i) are independent across i. Let A_{ij} denote the $T_i \times T_j$ blocks of the matrix $A \in \mathcal{A}$, corresponding to observations in clusters S_i and S_j, respectively. The numerators of the estimation errors take the form of quadratic expressions:

$$\hat{\beta}^A - \beta = \frac{x'Ae}{x'Ax} \quad \text{and} \quad x'Ae = \sum_{i,j} x_i' A_{ij} e_j.$$

As these are quadratic forms in random (x_i, e_i), the variance of the estimators has a non-standard structure. In particular, establishing asymptotic normality may require a CLT for quadratic forms.

Special case: block-diagonal structure. It is worth highlighting a special case —where the numerator reduces to linear forms. If the off-diagonal blocks A_{ij} are equal to zero for all $i \neq j$, then $x'Ae = \sum_i x_i' A_{ii} e_i$. Since (x_i, e_i) are independent across clusters, standard inference follows directly from the law of large numbers and central limit theorem, provided the number of clusters $N \to \infty$ and that the clusters satisfy a negligibility condition. This scenario arises in Examples 5 and 6 when only cluster fixed effects α_i are included in W. In that case, usual clustered standard errors can be used.

However, this structure breaks down when more complex controls are included.

Example 7 (Multi-way fixed effects)**.** Consider the model:

$$y_{it} = \alpha_i + \gamma_{g(i,t)} + \beta x_{it} + e_{it},$$

where α_i is an individual fixed effect, $g(i,t)$ denotes the group (e.g., teacher or classroom) assigned to individual i at time t, and γ_g is a group fixed effect. For instance, y_{it} may represent student i's academic achievement in year t, and x_{it} could be last year's score (to study persistence of achievement) or an intervention determined by past performance. The fixed effect α_i captures student ability, while γ_g represents teacher or classroom effects. The clustering is assumed at the individual level, and we assume weak exogeneity: $\mathbb{E}[e_{it} \,|\, \{x_{is}, s \geq t\}] = 0$.

Assume all students are observed in two periods. In year 1, students are split evenly between teachers $g = 1$ and $g = 2$; in year 2, they are randomly reassigned to $g = 3$ and $g = 4$. Each classroom has the same number of students. For a student i with teachers $(1, 3)$, the differenced equation is:

$$\Delta y_i = y_{i2} - y_{i1} = \gamma_3 - \gamma_1 + \beta \Delta x_i + \Delta e_i.$$

OLS partials out teachers' effects and transforms this to:

$$y_i^* = \Delta y_i - \left\{ \frac{3}{4}\overline{\Delta y}_{(1,3)} + \frac{1}{4}\overline{\Delta y}_{(1,4)} + \frac{1}{4}\overline{\Delta y}_{(2,3)} - \frac{1}{4}\overline{\Delta y}_{(2,4)} \right\},$$

where the term in brackets estimates $\gamma_3 - \gamma_1$, and $(2, 3)$ refers to the group of students who had teacher $g = 2$ in year 1 and $g = 3$ in year 2, and the upper bar denotes the average over the group. Similar transformations apply to x_i^* and e_i^*. Our estimator $\hat{\beta}^{A^*}$ is IV on the equation $y_i^* = \beta x_i^* + e_i^*$, using x_{i1} as an instrument: $\hat{\beta}^{A^*} - \beta = \frac{\sum_{i=1}^N x_{i1} e_i^*}{\sum_{i=1}^N x_{i1} x_i^*}$. But since e_i^* depends on all e_{jt}, the terms $x_{i1} e_i^*$ are dependent across i. Here, A^* is not block-diagonal, making standard inference approaches invalid and motivating the need for a different variance estimator. □

Example 7 highlights several key points. First, the difference between e_i^* and e_i reflects estimation error from recovering the teacher effects γ_g. Usage of the full sample to estimate these effects mixes observations and induces complex dependence, complicating inference. If γ_g estimates are very precise, then $\sum_i x_{i1} e_i^*$ will be close to $\sum_i x_{i1} e_i$, and ignoring the dependence may still yield valid inference.

Second, the OLS transformation for a student in group $(1, 3)$ involves not only students with shared teachers ($g = 1$ or $g = 3$), but also those indirectly connected, e.g., students in $(2, 4)$, who share a teacher with others in $(1, 4)$ or $(2, 3)$. More generally, we can view students as nodes in a network, where edges represent shared teachers. OLS demeaning propagates information across this network, creating strong dependencies within connected components and complicating standard inference methods.[2]

4.2 The correct formula for quantifying uncertainty

Denote $\lambda_i = \mathbb{E}[x_i]$ and $v_i = x_i - \lambda_i$. Then,

$$x'Ae = \sum_{j=1}^{N} \left(\sum_{i=1}^{N} \lambda_i' A_{ij} + v_j' A_{jj} \right) e_j + \sum_{j=1}^{N} \left(\sum_{i \neq j} v_i' A_{ij} \right) e_j = \sum_{j=1}^{N} \omega_j e_j + \sum_{j=1}^{N} Q_j e_j.$$

where $Q_j = \sum_{i \neq j} v_i' A_{ij}$ is mean-zero and independent of (e_j, v_j). The above is the so-called Hoeffding decomposition (Hoeffding, 1948); the two sums are the first two U-statistics, which are uncorrelated with each other by construction. Thus, the variance decomposes as:

$$\mathbb{V}(x'Ae) = \mathbb{V}\left(\sum_{j=1}^{n} \omega_j e_j \right) + \mathbb{V}\left(\sum_{j=1}^{n} Q_j e_j \right).$$

The first term is the first order U-statistic with $\mathbb{V}\left(\sum_{j=1}^{n} \omega_j e_j\right) = \sum_{j=1}^{n} \mathbb{V}\left(\omega_j e_j\right)$. The second term has correlated summands:

$$\mathbb{V}\left(\sum_{j=1}^{n} Q_j e_j \right) = \sum_{j=1}^{n} \mathbb{V}\left(Q_j e_j\right) + \sum_{j=1}^{N} \sum_{i \neq j} \mathbb{E}[v_i' A_{ij} e_j v_j' A_{ji} e_i]. \tag{7}$$

Standard cluster-robust variance formulas omits the final term $\sum_{j=1}^{N} \sum_{i \neq j} \mathbb{E}[v_i' A_{ij} e_j v_j' A_{ji} e_i]$, which captures cross-cluster dependence. This term vanishes in two important cases: (i) under strict exogeneity, $\mathbb{E}[e_j v_j'] = 0$ for all j; or (ii) when A is block-diagonal, i.e., $A_{ij} = 0$ for $i \neq j$. These are the settings where standard clustered standard errors yield valid inference.

[2] While OLS demeaning is efficient under the benchmark conditions of Lemma 5, one may prefer a less efficient estimator that induces less dependence. In Example 7, instead of using the full sample to estimate $\gamma_3 - \gamma_1$, one could rely only on students with the same teacher history, e.g., $\overline{\Delta y}_{(1,3)}$. This results in independent "super-clusters" defined by the teacher pairs: $(1,3)$, $(1,4)$, $(2,3)$, and $(2,4)$. One can go further by forming pairs of students with identical teacher histories and differencing them, creating many small, independent clusters, then using clustered standard errors on that level. However, this comes at a cost: the resulting estimator may be highly inefficient due to poor estimation of fixed effects. This strategy is explored in Verdier (2020).

5 Asymptotic Gaussianity

We establish a result on the asymptotic Gaussianity of the numerator of the estimation error, $x'Ae$. This result enables the construction of a hypothesis test for $H_0 : \beta = \beta_0$, as well as the development of a confidence set for β in the presence of clustered data. Implementation details are provided in Section 6. Subsection 5.1 presents a non-technical exposition and discusses the assumptions required to establish Gaussianity. Subsection 5.2 is intended for theoretical econometricians and contains the technical details. Readers whose focus is more applied may skip it.

5.1 Non-technical exposition

Two main challenges arise when establishing Gaussianity. First, the statistic includes not only a linear term but also a quadratic form in random shocks. Second, intra-cluster dependence introduces complexities, raising the question of how to restrict such dependence and what trade-offs arise between dependence and cluster size. Our asymptotic framework requires the number of clusters to grow, i.e., $N \to \infty$.

The first challenge is addressed by developing a new Central Limit Theorem (Lemma 6) for quadratic forms of independent, normalized random vectors. These vectors represent observations within each cluster. The theorem establishes that, under certain negligibility conditions, a quadratic form involving matrix-weighted sums of such vectors converges in distribution to a normal distribution. The core requirement is that the contribution of any single cluster—as measured, for instance, by the Frobenius norm of the corresponding submatrix—becomes negligible relative to the total variance. A key strength of this result is that it avoids direct control over cluster sizes; the cluster size enters only implicitly through restrictions on the weight matrices.

Our main result—Gaussianity of $x'Ae$—is formalized in Theorem 2 for a general matrix A and later specialized to our proposed estimator matrix A^*. Theorem 2 combines a Lyapunov-type CLT for linear forms with the new quadratic CLT. Two sets of assumptions are required.

The first set (Assumption 3) imposes conditions on the distribution of errors, including moment bounds and tail behavior within clusters. Specifically, any normalized linear combination of errors within a cluster (i.e., one with unit variance) must have a uniformly bounded fourth moment. Likewise, normalized combinations of $e_i v_i'$ must have uniformly bounded $(2 + 2\delta)$ moments. These assumptions ensure that intra-cluster dependence is well-approximated by the covariance matrix and rule out extreme tail dependence, such as aligned outliers not reflected in the covariance. Further, we require that certain correlations are bounded away from one and that variances are bounded away from zero, thereby excluding cases of near-perfect predictability.

The second set (Assumption 4) concerns the interaction between the weight matrix A and the operator norms of the cluster-specific covariance matrices Σ_i. These assumptions ensure that each cluster's contribution to the total variance remains asymptotically negligible. By working with $\|\Sigma_i\|$, we allow for a trade-off between cluster size and intra-cluster dependence. Two stylized examples clarify the implications for the quadratic form (Assumption 4(iii)) that typically requires stronger restrictions :

- In the presence of a pervasive cluster-level shock (e.g., a factor structure), errors within a cluster may be strongly correlated. Then $\|\Sigma_i\|$ can grow proportionally with cluster size T_i, up to $\|\Sigma_i\| \leq CT_i$. To achieve Gaussianity of the quadratic term in this case, stronger conditions must be imposed on the weight matrix and cluster size, such as $\max_i T_i^4/n \to 0$.
- At the opposite extreme, if no pervasive shocks exist and dependence decays rapidly with distance (in some metric) within the cluster, then $\|\Sigma_i\|$ may remain uniformly bounded regardless of cluster size. In such cases, weaker conditions, such as $\max_i T_i^2/n \to 0$ suffice.

5.2 Technical details

Lemma 6. *Let ξ_i be independent $T_i \times 1$ random vectors with $\mathbb{E}[\xi_i] = 0$, $\mathbb{V}(\xi_i) = I_{T_i}$, and $\mathbb{E}|\tau'\xi_i|^4 \leq C\|\tau\|^4$ for all $T_i \times 1$ vectors τ. Let Γ be an $n \times n$ block matrix with blocks Γ_{ij} of size $T_i \times T_j$, symmetric across the diagonal ($\Gamma_{ij} = \Gamma'_{ji}$), and with $\Gamma_{ii} = 0$ for all i. Assume:*

$$\frac{\max_i \left(\sum_{j\neq i} \|\Gamma_{ij}\|_F^2\right)}{\sum_i \sum_{j\neq i} \|\Gamma_{ij}\|_F^2} \to 0 \textit{ and } \frac{\|\Gamma\|^2}{\|\Gamma\|_F^2} \to 0.$$

Then as $N \to \infty$

$$\frac{1}{\sqrt{V}} \sum_{i=1}^{N} \sum_{j \neq i} \xi_i' \Gamma_{ij} \xi_j \xrightarrow{d} N(0,1), \quad \textit{where } V = 2\|\Gamma\|_F^2 = 2 \sum_{i} \sum_{j \neq i} \|\Gamma_{ij}\|_F^2.$$

Let Σ_i denote the covariance matrix of $(e_i', v_i')'$, with submatrices $\Sigma_{e,i}$ and $\Sigma_{v,i}$. Define the normalized shock vector for cluster i as $\xi_i = \Sigma_i^{-1/2}(e_i', v_i')'$, which is a $2T_i \times 1$ vector. Also define the cross-covariance matrix $\Psi_i = \mathbb{E}[e_i(v_i \otimes e_i)']$ of dimension $T_i \times T_i^2$, and the variance matrix $\Phi_i = \mathbb{V}(v_i \otimes e_i)$ of dimension $T_i^2 \times T_i^2$.

Assumption 3. *The cluster-specific errors $(e_i', v_i')'$ satisfy the following conditions:*

(i) Each Σ_i is invertible, with $\|\Sigma_i^{-1}\| \leq C$, and ξ_i are independent, mean-zero vectors satisfying $\mathbb{E}[v_i' A_{ii} e_i] = 0$.

(ii) Lower bound on variance ratio: $\min_i \min_\tau \frac{\tau' \Phi_i \tau}{\tau'(\Sigma_{v,i} \otimes \Sigma_{e,i})\tau} > c$;

(iii) Bounded alignment: $\max_i \left\| \Sigma_{e,i}^{-1/2} \Psi_i \Phi_i^{-1/2} \right\| \leq 1 - c$;

(iv) Higher-order moment bounds: $\mathbb{E}|\tau' \xi_i|^4 \leq C\|\tau\|^4$, and $\mathbb{E}(\tau'(v_i \otimes e_i))^{2+2\delta} \leq C\left(\tau' \Phi_i \tau\right)^{1+\delta}$, for some $\delta > 0$.

The assumptions above ensure two linear CLTs: one for sums of e_i and one for sums of $v_i \otimes e_i$. Condition (ii) guarantees that $(\Sigma_{v,i}^{-1/2} v_i) \otimes (\Sigma_{e,i}^{-1/2} e_i)$ has eigenvalues bounded away from zero, preventing near-degeneracy. Condition (iii) ensures that correlations between e_i and $v_i \otimes e_i$ remain bounded away from one. Condition (iv) rules out heavy-tailed behavior and strong higher-order dependence that could dominate the covariance-based approximation.

Assumption 4. *Let A be an $n \times n$ matrix such that*

(i) $\sum_{i=1}^{N} (A'\lambda)_i' \Sigma_{e,i} (A'\lambda)_i + \|A\|_F^2 \geq cn$;

(ii) $\frac{\max_i \|\Sigma_i\|^2 \|A_{ii}\|_F^2}{n} \to 0$ *and* $\frac{\max_i (A'\lambda)_i' \Sigma_{e,i} (A'\lambda)_i}{n} \to 0$;

(iii) either one of the two conditions hold:

(a) $\frac{\max_i \|\Sigma_i\|^2 \cdot \sum_{i=1}^{N} \sum_{j \neq i} \|A_{ij}\|_F^2}{n} \to 0$

(b) $\frac{\max_i \|\Sigma_i\|^2 \cdot \max_i \sum_{j=1}^N (\|A_{ij}\|_F^2 + \|A_{ji}\|_F^2)}{n} \to 0$ *and* $\frac{\|A\|^2 \cdot \max_i \|\Sigma_i\|^2}{n} \to 0.$

Part (i) guarantees that the normalization in the CLT is no slower than $1/\sqrt{n}$. Part (ii) ensures cluster negligibility for the linear CLT, Part (iii) (b) does so for the quadratic CLT, while Part (iii) (a) assumes that the quadratic part is negligible in comparison to the linear one.

Theorem 2. *Let $x = \lambda + v$, where $\lambda = \mathbb{E}[x]$, and suppose Assumptions 3 and 4 hold. Then, as $N \to \infty$, $\frac{x'Ae}{\omega} \xrightarrow{d} \mathcal{N}(0,1)$, where $\omega^2 = \mathbb{V}(x'Ae) \geq cn$ for some $c > 0$.*

We now discuss how Assumption 4 applies to the matrix A^* defined in Theorem 1. According to Lemma 5, the Frobenius norm $\|A^*\|_F^2$ captures the effective sample size. Assumption 4(i) holds provided A^* retains sufficient variation relative to n. From Theorem 1(c), each row of A^* corresponds to a projection, and we have:

$$\sum_{\tilde{\ell}=1}^{n} (A^*_{\ell\tilde{\ell}})^2 = A^*_{\ell\ell} \text{ and } \|A^*\|_F^2 = \operatorname{tr}(A^*) = \sum_{\ell} A^*_{\ell\ell}.$$

Since A^* is observable, $\|A^*\|_F^2$ can be computed and directly compared to n. A simple sufficient condition for Assumption 4(i) is illustrated below:

Example 8. Suppose the data is indexed by $\ell = \ell(i,t)$ and that a standard weak exogeneity condition holds: $\mathbb{E}[e_{it} \mid x_{is} : s \leq t] = 0$. Let $\mathcal{X}_0 = \{\ell(i,1)\}$ denote the set of first-period observations for each cluster. For these ℓ, we have $\mathcal{E}_{\ell\tilde{\ell}} = 1$ for all $\tilde{\ell}$, so no observations are excluded from the projection, and thus $A^*_{\ell\ell} = M_{\ell\ell}$. If $\min_{\ell \in \mathcal{X}_0} M_{\ell\ell} > c > 0$ and average cluster size is bounded, then Assumption 4(i) holds. The condition $M_{\ell\ell} > c$ is standard in the many regressors/instruments literature (e.g., Cattaneo et al., 2018). □

Assumption 5 (Balanced design)**.** *Assume that the set of controls and the exclusion matrix $\mathcal{E}$ are such that for any non-trivial submatrix M_ℓ—formed from entries $M_{\ell_1\ell_2}$ with $\mathcal{E}_{\ell\ell_1} = 0$ and $\mathcal{E}_{\ell\ell_2} = 0$ —satisfies $\|M_\ell^{-1}\| \leq C$.*

Theorem 1 shows that $A^* = (I - B)M$, where B is a sparse matrix with relatively few nonzero entries. Assumption 5 ensures that A^* remains close to M under appropriate matrix norms.

Lemma 7. *The following primitive conditions are sufficient for parts of Assumption 4:*

- *If* $\frac{\max_i \|\Sigma_i\|^2 T_i}{n} \to 0$*, then the first part of Assumption 4(ii) holds;*
- *If Assumption 5 holds,* $\|\lambda\|_\infty < \infty$ *and* $\frac{(\max_i T_i)^3 \|\Sigma\| \cdot \|M\|_\infty^2}{n} \to 0$*, then the second part of Assumption 4(ii) holds;*
- *If Assumption 5 holds, and* $\frac{\max_i \|\Sigma_i\|^2 \cdot \max_i T_i^2}{n} \to 0$*, then Assumption 4(iii)(b) holds.*

For linear CLTs, the condition $\max_i T_i^2/n \to 0$ suffices (Hansen and Lee, 2019; Sasaki and Wang, 2022), while the quadratic CLT (Assumption 4(iii)(b)) requires the stricter condition $\max_i T_i^4/n \to 0$.

In high-dimensional settings with many regressors, an analog of the second part of Assumption 4(ii)—e.g., $\frac{\|M\lambda\|_\infty^2}{n} \to 0$—is often imposed as a high-level assumption because direct verification is difficult; see, for example, Assumption 3 in Cattaneo et al. (2018). Lemma 7 offers low-level primitive conditions that imply this assumption. The bound $\|\lambda\|_\infty < \infty$ is mild since $\lambda_\ell = \mathbb{E}[x_\ell]$, and for a fixed effects projection matrix, we have $\|M\|_\infty = 2$.

6 Inference with clustered data

6.1 Jackknife variance estimator

Suppose we test the null hypothesis $H_0 : \beta = \beta_0$, and define $U_i = y_i - x_i'\beta_0$. Even under H_0, these differ from the true errors e_i due to the inclusion of a predictable component: $U_i = e_i + w_i'\delta$. By the partialling-out property of A^*, we have $x'A^*U = x'A^*e$.

The jackknife variance estimator, introduced by Efron and Stein (1981) for symmetric statistics based on independent and identically distributed data, has also been advocated as a useful tool in settings involving non-symmetric statistics and non-identically distributed data (see, e.g., Bousquet et al., 2004; MacKinnon et al., 2023). In our setting, we define the statistic of interest as $\mathcal{Z} = x'A^*U = x'A^*e = f(D_1, \dots, D_N)$, where $D_i = (x_i, U_i)$ denotes the observed data for cluster i. Let $\mathcal{Z}_{(i)} = f(D_1, \dots, D_{i-1}, 0, D_{i+1}, \dots, D_N)$ denote the statistic computed when cluster i is removed (i.e., replaced by zero). The jackknife estimator of the variance is then given by

$$\hat{V}_{\text{JK}} = \sum_{i=1}^{N} (\mathcal{Z} - \mathcal{Z}_{(i)})^2 = \sum_{i=1}^{N} \left(\sum_{j \neq i} x_j' A_{ji}^* U_i + \sum_{j=1}^{N} x_i' A_{ij}^* U_j \right)^2 .$$

The following lemma establishes that $\hat{V}_{\text{JK}}$ is a conservative estimator, potentially overestimating the true variance.

Lemma 8. *Under Assumptions 1 and 3, the jackknife variance estimator satisfies* $\mathbb{E}[\hat{V}_{JK}] \geq \mathbb{V}(\mathcal{Z})$*. Moreover, if* $A^*_{ij} = 0$ *for all* $i \neq j$ *(i.e.,* A^* *is block-diagonal), then the estimator is unbiased under the null:* $\mathbb{E}[\hat{V}_{JK}] = \mathbb{V}(\mathcal{Z}) = \omega^2$ *when* $\beta = \beta_0$*.*

Efron and Stein (1981) showed that the jackknife variance estimator tends to be conservative for symmetric statistics derived from independent and identically distributed data. While the jackknife correctly captures the variance of first-order U-statistics, it double-counts second-order terms. In our setting, $\mathcal{Z}$ is non-symmetric and based on independent but non-identically distributed data. The jackknife is conservative for two main reasons: it gives double weight to the quadratic term (as noted in Efron and Stein (1981)), and $\mathcal{Z}_{(i)}$ is improperly centered since $\mathbb{E}[U_i - e_i] \neq 0$. The jackknife variance estimator is unbiased when $A^*_{ij} = 0$ for all $i \neq j$, in which case the removal of a single cluster does not affect the residualization of others. Consequently, $\mathcal{Z}_{(i)}$ is properly centered, and no cross-cluster covariances arise. In general, the degree of conservativeness is small when $\sum_{j=1}^N \sum_{i \neq j} \|A^*_{ij}\|_F^2$ is much smaller than $\sum_{j=1}^N \|A^*_{jj}\|_F^2$.

6.2 Tests and confidence sets

Our estimator, $\hat{\beta}^{A^*} = \frac{x'A^*y}{x'A^*x} = \frac{z'y}{z'x}$, belongs to the class of just-identified internal IV estimators. Similar to the IV estimator for dynamic panels proposed by Arellano and Bond (1991), our approach may be subject to weak identification concerns (see Bun and Windmeijer (2010)). To construct inference procedures that are robust to weak identification, we propose using the Anderson–Rubin (AR) test. The AR test delivers valid inference under both strong and weak identification.

To test the null hypothesis $H_0 : \beta = \beta_0$, we employ the test statistic

$$AR(\beta_0) = \frac{(x'A^*(y - x\beta_0))^2}{\hat{V}(\beta_0)},$$

and reject the null if this statistic exceeds the $1 - \alpha$ quantile of the χ^2_1 distribution. The variance estimator $\hat{V}(\beta_0)$ is given by $\hat{V}_{\text{JK}}$, which is based on the implied errors $U = U(\beta_0) = y - x\beta_0$. By Theorem 2, the AR test has asymptotic size no greater than the nominal level.

A confidence set is obtained by inverting the AR test. Since both the numerator and the denominator are quadratic in β_0, the inversion reduces to solving a quadratic inequality. The resulting confidence set is always non-empty and always includes the estimator $\hat{\beta}^{A^*}$.

When identification is strong, the AR-based test and confidence set are asymptotically equivalent to those based on the t-statistic, using $\hat{\beta}^{A^*}$ as the point estimator and $\frac{\hat{V}(\hat{\beta}^{A^*})}{(x'A^*x)^2}$ as the squared standard error. Nevertheless, we advocate using the AR test, as the validity of t-statistic-based inference is not generally guaranteed.

7 Empirical Application

Policy interventions with spatial effects present a compelling setting for applying our method. In such contexts, researchers often distinguish between the *direct* effect of treatment—on the treated units themselves—and the *total* effect, which includes indirect impacts arising from spillovers. A prominent example is provided by Egger et al. (2022) (hereafter EHMNW), who study a large-scale fiscal intervention in rural Kenya. In their experiment, one-time cash transfers in total equivalent to approximately 15 percent of local GDP were randomly assigned to 653 villages. We apply our method to this setting to estimate the direct effect of the intervention on treated villages.

The villages (our units of observation) were divided into 68 sub-locations (clusters), the median sub-location contains 7 villages. Treated villages were randomly selected through a two-stage randomization procedure. Within each treated village, all eligible households received a one-time cash transfer of USD 1,871 (PPP-adjusted). Eligibility was determined through a poverty-based means test. The median village includes 98 households, approximately one-third of whom met the eligibility criteria. EHMNW examine a range of outcomes, including household consumption, firm behavior, and local prices. Here, we focus exclusively on village-aggregated outcomes for total household consumption. Results for other outcomes are similar and available upon request.

The parameter of interest in our application is the village-level *direct* treatment effect on average household consumption. Although treatment was randomly assigned, a simple OLS regression is unreliable due to anticipated interference from neighboring villages. EHMNW assume that spatial spillovers do not extend beyond a radius of R kilometers, with their preferred specification (selected through BIC criteria) being $R = 2$.

This setting fits naturally within our design-based framework, which we illustrate through estimation under two specifications. In both cases, we define the exogeneity matrix $\mathcal{E}$ using cutoffs based on pairwise distances between villages. We assume the exclusion restriction $E[v_{\tilde{\ell}}(y_\ell - \beta x_\ell)] = 0$ holds if and only if $\ell = \tilde{\ell}$ or the distance between villages $\tilde{\ell}$ and ℓ is at least R kilometers. In practice, researchers should choose a distance cutoff below which exogeneity may be questionable. Here, we illustrate the behavior of the $\hat{\beta}^{A^*}$ estimator using a range of cutoffs from 0 to 3 kilometers. In the data, the average pairwise distance between villages within a sub-location is 2.2 kilometers. On average, each village has 1.5, 4.6, and 6.7 neighbors within the sub-location and within 1, 2, and 3 kilometers[3], respectively. Our estimates under different choices of R are shown in Figure 2.

(a) Baseline (b) Continuous treatment

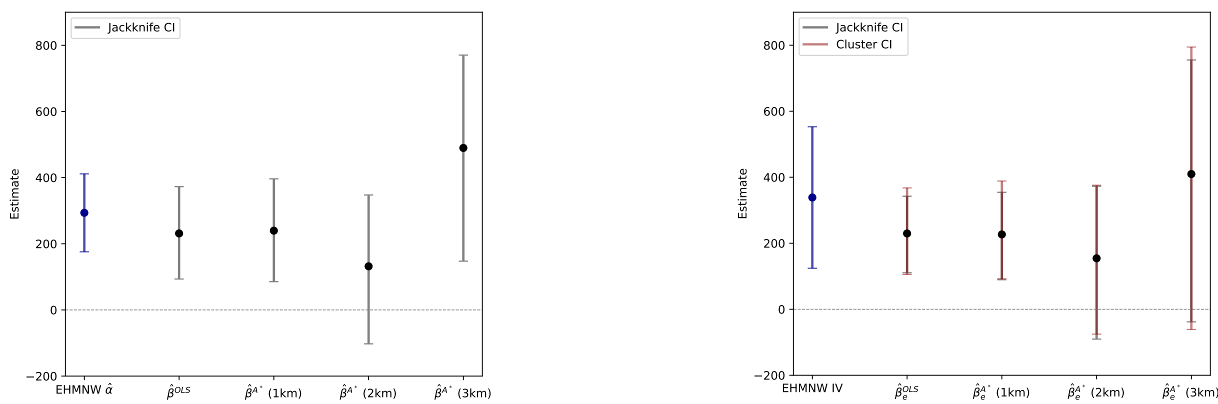

Figure 2: Estimated effects on household consumption

Notes. The outcome variable in Panel (a) is average household consumption; in Panel (b), it is average household consumption among eligible households. Effect sizes in Panel (b) are scaled by the average transfer amount. Jackknife and cluster-robust confidence intervals are computed by inverting the respective AR test statistics. The estimates labeled "EHMNW" are taken directly from the original paper: the $\hat{\alpha}$ estimate in Panel (a) is based on Specification (1) from EHMNW and reported in their Table 1 (row 1, column 1); the IV estimate in Panel (b) is based on Specification (2) from the paper and reported in Table 1 (row 1, column 2). These estimates use spatial standard errors based on household GPS coordinates, following Conley (2008).

For specification (a), let y_ℓ denote the average consumption of households in village ℓ, and let x_ℓ denote the binary treatment status of the village. We assume the correctly specified

[3] Based on GPS coordinates. We dropped nine observations for which we do not have GPS coordinates.

treatment equation: $x_\ell = p_{i(\ell)} + v_\ell$, where p_i is a cluster (sub-location) fixed effect capturing treatment propensity. We include sub-location fixed effects to flexibly absorb the realized treatment share, which may differ substantially from the ex ante assignment probability. In practice, the realized propensity can depend on cluster size or arise from re-randomization or re-balancing procedures implemented during the experiment.

The results are reported in Panel (a), alongside the baseline estimator from the specification used in EHMNW. The baseline estimator is reported in Table 1 (row 1, column 1) of EHMNW. We find that the estimates remain relatively stable for distance cutoffs below 2 km, suggesting that the bias in estimating the direct effect due to inter-village spillovers is limited. As larger values of R correspond to more relaxed exogeneity assumptions, the effective sample size decreases, leading to larger standard errors. Figure 2, Panel (a), also displays confidence intervals based on the jackknife variance estimator described in Section 6. In this specification, since the only controls are cluster fixed effects, the A^* matrix is block-diagonal, and the jackknife estimator coincides with the standard cluster-robust variance estimator.

Specification (b), reported in Figure 2, Panel (b), makes use of the full variation in treatment intensity. Here, we define y_ℓ as the average consumption of eligible households in village ℓ, and let x_ℓ denote the total transfer allocated to the village. The assumed treatment assignment equation in this case is $x_\ell = W_\ell' \delta + v_\ell$, where W_ℓ includes the number of eligible households in village ℓ along with sub-location fixed effects. The parameter of interest, β, again represents the village-level direct treatment effect. This parameter is somewhat comparable to the total effect on eligible households reported in EHMNW, which was estimated using an instrumental variables strategy and reported in Table 1 (row 1, column 2) of the paper. Estimates under different exogeneity assumptions are presented along with both jackknife and cluster-robust standard errors. In this specification, the controls W_ℓ include a covariate that varies within clusters and is not absorbed by the sub-location fixed effects. As a result, matrix M is no longer block-diagonal, and neither is A^*, so the two variance estimators do not coincide.

Discussion. Both panels illustrate our main takeaway: the point estimates—and especially their precision—are highly sensitive to the researcher's maintained exogeneity assumptions. Relaxing these assumptions, for example by allowing spillovers to extend within 3 km rather than 2 km, leads to substantially wider confidence intervals. This reflects a reduction in the effective sample size captured by the trace of the A^* matrix under less restrictive exogeneity assumptions.

(a) A^* with 1 km cutoff   (b) A^* with 3 km cutoff

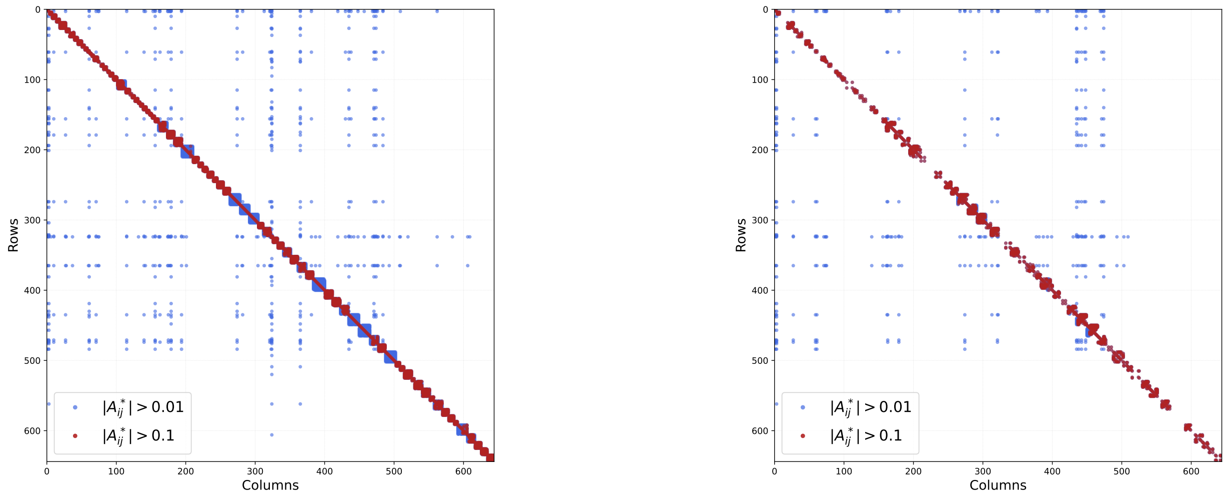

Figure 3: Non-zero structure of A^*

Notes. This figure visualizes entries of the matrix A^* with absolute value greater than 0.01 (blue) or 0.1 (red). Panel (a) shows A^* computed under 1 km cutoff and Panel (b) shows A^* under 3 km cutoff.

The structure of the matrix A^* in the continuous treatment specification (same as in Figure 2, Panel (b)) is shown in Figure 3, Panel (a) for $R = 1$ and Panel (b) for $R = 3$. A blue dot indicates that the absolute value of the corresponding element of A^* exceeds 0.01, while a red dot indicates that it exceeds 0.1. Villages are sorted by cluster (sub-location), so that block structure, if present, is visible. The figure reveals that the matrix A^* is far from block-diagonal: some villages contribute substantially to the residualization of controls across many clusters. As the cutoff radius R increases, the trace of the matrix A^* decreases. This pattern reflects a reduction in the effective sample size. The ratio of Frobenius norms for off-diagonal blocks to diagonal blocks increases with distance, but remains small. When $R = 1$, we have $\mathrm{tr}(A^*) = 539$, while the relative Frobenius norm of off-diagonal blocks is 0.043. When $R = 3$, $\mathrm{tr}(A^*) = 273$ while the Frobenius norm of off-diagonal blocks is 0.048. Due to the relative smallness of off-diagonal blocks, the confidence sets using clustered and jackknife variance estimators are very close to each other.

References

Ahn, S. C. and P. Schmidt (1995). Efficient estimation of models for dynamic panel data. *Journal of Econometrics 68*(1), 5–27.

Alvarez, J. and M. Arellano (2003). The time series and cross-section asymptotics of dynamic panel data estimators. *Econometrica 71*(4), 1121–1159.

Anderson, T. W. and C. Hsiao (1981). Estimation of dynamic models with error components. *Journal of the American Statistical Association 76*(375), 598–606.

Arellano, M. and S. Bond (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The Review of Economic Studies 58*(2), 277–297.

Arellano, M. and O. Bover (1995). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics 68*(1), 29–51.

Banerjee, A., M. Alsan, E. Breza, A. G. Chandrasekhar, A. Chowdury, E. Duflo, P. Goldsmith-Pinkham, and B. A. Olken (2024). Can a trusted messenger change behavior when information is plentiful? evidence from the first months of the covid-19 pandemic in west bengal. *Review of Economics and Statistics*, 1–33.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2013). The diffusion of microfinance. *Science 341*(6144), 1236498.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2019). Using gossips to spread information: Theory and evidence from two randomized controlled trials. *The Review of Economic Studies 86*(6), 2453–2490.

Blattman, C., D. P. Green, D. Ortega, and S. Tobón (2021). Place-based interventions at scale: The direct and spillover effects of policing and city services on crime. *Journal of the European Economic Association 19*(4), 2022–2051.

Blundell, R. and S. Bond (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics 87*(1), 115–143.

Bousquet, O., S. Boucheron, and G. Lugosi (2004). Concentration inequalities. In *Advanced Lectures on Machine Learning*, pp. 208–240. Springer.

Bun, M. J. G. and V. Sarafidis (2015). Dynamic panel data models. In *The Oxford Handbook of Panel Data*. Oxford University Press.

Bun, M. J. G. and F. Windmeijer (2010). The weak instrument problem of the system gmm estimator in dynamic panel data models. *The Econometrics Journal 13*(1), 95–126.

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

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics 34*(3), 305–334.

Djogbenou, A. A., J. G. MacKinnon, and M. Ø. Nielsen (2019). Asymptotic theory and wild bootstrap inference with clustered errors. *Journal of Econometrics 212*(2), 393–412.

Efron, B. and C. Stein (1981). The jackknife estimate of variance. *The Annals of Statistics*, 586–596.

Egger, D., J. Haushofer, E. Miguel, P. Niehaus, and M. Walker (2022). General equilibrium effects of cash transfers: experimental evidence from kenya. *Econometrica 90*(6), 2603–2643.

Hansen, B. E. (2022). A modern gauss–markov theorem. *Econometrica 90*(3), 1283–1294.

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

Hayashi, F. and C. Sims (1983). Nearly efficient estimation of time series models with predetermined, but not exogenous, instruments. *Econometrica*, 783–798.

Henderson, H. V. and S. R. Searle (1981). On deriving the inverse of a sum of matrices. *SIAM review 23*(1), 53–60.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 293–325.

Jayachandran, S., J. De Laat, E. F. Lambin, C. Y. Stanton, R. Audy, and N. E. Thomas (2017). Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation. *Science 357*(6348), 267–273.

Lehmann, E. L. and G. Casella (1998). *Theory of point estimation.* Springer.

Lei, L. and J. Wooldridge (2022). What estimators are unbiased for linear models? *arXiv preprint arXiv:2212.14185*.

Ligtenberg, J. W. (2023). Inference in iv models with clustered dependence, many instruments and weak identification. *arXiv preprint arXiv:2306.08559*.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2023). Fast and reliable jackknife and bootstrap methods for cluster-robust inference. *Journal of Applied Econometrics 38*(5), 671–694.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica*, 1417–1426.

Okui, R. (2021). 1. linear dynamic panel data models. *Handbook of Research Methods and Applications in Empirical Microeconomics*, 1.

Paluck, E. L., H. Shepherd, and P. M. Aronow (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences 113*(3), 566–571.

Portnoy, S. (2022). Linearity of unbiased linear model estimators. *The American Statistician 76*(4), 372–375.

Sasaki, Y. and Y. Wang (2022). Non-robustness of the cluster-robust inference: with a proposal of a new robust method. *arXiv preprint arXiv:2210.16991*.

Sølvsten, M. (2020). Robust estimation with many instruments. *Journal of Econometrics 214*(2), 495–512.

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

A Appendix

Assumption 6. *(i) Observations are independent across clusters: the set* $\{e_\ell, x_\ell\}_{\ell \in S_i}$ *is independent across* i *with* $\max_\ell \mathbb{E}[x_\ell^4 + e_\ell^4] < C$.

(ii) The expected variation of the residualized regressor $\tilde{x}_\ell$ *diverges with the sample size* n*;* $\frac{1}{n}\sum_{\ell=1}^n \mathbb{E}[\tilde{x}_\ell^2] \to Q > 0$.

(iii) The number of controls K *may grow with* n*, subject to* $\frac{K}{n} < 1 - c$*. The largest cluster size satisfies* $\max_i T_i/n = o(1)$.

Discussion of assumptions. Assumption 6(i) formalizes the earlier description: observations are independent across clusters but may be dependent within clusters. The finite fourth moment condition is standard in linear econometrics. Assumption 6(ii) ensures sufficient variation in the residualized regressor $\tilde{x}$ for consistency of OLS, when asymptotic bias is not an issue. Assumption 6(iii) accommodates high-dimensional controls W, including multiple sets of fixed effects. While clusters may be large and grow with n, each cluster must remain asymptotically negligible, as also required in Hansen and Lee (2019); Djogbenou et al. (2019). Thus, the number of clusters N diverges with the sample size n.

Proof of Lemma 1. By definition, $\tilde{x}_\ell = \sum_{\ell^*=1}^n M_{\ell\ell^*} x_{\ell^*}$. Assumption 1 implies: $\mathbb{E}[\tilde{x}_\ell e_\ell] = \sum_{\ell^*=1}^n M_{\ell\ell^*}\mathbb{E}[x_{\ell^*} e_\ell]$. Therefore,

$$\frac{1}{n}\sum_{\ell=1}^n \mathbb{E}[\tilde{x}_\ell e_\ell] = \frac{1}{n}\sum_{\ell=1}^n \sum_{\tilde{\ell}=1}^n M_{\ell\tilde{\ell}}\mathbb{E}[x_{\tilde{\ell}} e_\ell]. \tag{8}$$

To establish convergence in probability, we use an Efron–Stein-type argument. For generic clustered variables (v_ℓ, u_ℓ), with $\max_\ell \mathbb{E}[v_\ell^4 + u_\ell^4] < C$, define:

$$\tilde{v}_\ell = \sum_{\ell^*=1}^n M_{\ell\ell^*} v_{\ell^*}, \quad \mathcal{V}_i = \sum_{\ell \in S_i} \tilde{v}_\ell u_\ell, \quad \mathcal{U}_i = \sum_{\ell,\ell^* \in S_i} M_{\ell\ell^*} v_\ell u_{\ell^*}.$$

Note: $\sum_{\ell=1}^n \tilde{v}_\ell^2 \leq \sum_{\ell=1}^n v_\ell^2, \sum_{\ell\in S_i}\left(\sum_{\ell^*\in S_i} M_{\ell\ell^*} v_{\ell^*}\right)^2 \leq \sum_{\ell \in S_i} v_\ell^2$. Define $\mathcal{M} = \max_i \sum_{\ell \in S_i} u_\ell^2$.

Then:

$$\frac{1}{n^2}\sum_{i=1}^{N}\mathbb{E}[\mathcal{V}_i^2] \leq \frac{1}{n^2}\sum_{i=1}^{N}\mathbb{E}\left[\sum_{\ell\in S_i}\tilde{v}_\ell^2\mathcal{M}\right] \leq \mathbb{E}\left[\frac{\mathcal{M}}{n^2}\sum_{\ell=1}^{n}v_\ell^2\right] \leq C\sqrt{\mathbb{E}\left(\frac{\mathcal{M}^2}{n^2}\right)}$$
$$\leq C\sqrt{\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{N}\left(\sum_{\ell\in S_i}u_\ell^2\right)^2\right]} \leq C\sqrt{\frac{\max_i T_i}{n}} = o(1). \tag{9}$$

Similarly,

$$\frac{1}{n^2}\sum_{i=1}^{N}\mathbb{E}[\mathcal{U}_i^2] \leq \frac{1}{n^2}\sum_{i=1}^{N}\mathbb{E}\left[\sum_{\ell\in S_i}v_\ell^2\sum_{\ell\in S_i}u_\ell^2\right] \leq C\frac{1}{n^2}\sum_{i=1}^{N}T_i^2 \leq C\frac{\max_i T_i}{n} = o(1). \tag{10}$$

Now consider: $\mathcal{S}_n = \frac{1}{n}\sum_{\ell=1}^{n}\tilde{x}_\ell e_\ell = \frac{1}{n}\sum_{\ell=1}^{n}x_\ell\tilde{e}_\ell = \frac{1}{n}\sum_{\ell=1}^{n}\sum_{\ell^*=1}^{n}M_{\ell\ell^*}x_\ell e_{\ell^*}$. Define:

$$\Delta_i\mathcal{S}_n := \mathcal{S}_n - \frac{1}{n}\sum_{\ell\notin S_i}\sum_{\ell^*\notin S_i}M_{\ell\ell^*}x_\ell e_{\ell^*} = \frac{1}{n}\sum_{\ell\in S_i}(\tilde{x}_\ell e_\ell + x_\ell\tilde{e}_\ell) - \frac{1}{n}\sum_{\ell,\ell^*\in S_i}M_{\ell\ell^*}x_\ell e_{\ell^*}.$$

Applying (9) and (10) with $(v_\ell, u_\ell) \in \{(x_\ell, e_\ell), (e_\ell, x_\ell)\}$, we conclude:$\sum_{i=1}^{N}\mathbb{E}[(\Delta_i\mathcal{S}_n)^2] = o(1)$. By the Efron-Stein inequality, $\mathbb{V}(\mathcal{S}_n) \leq \sum_{i=1}^{N}\mathbb{E}[(\Delta_i\mathcal{S}_n)^2]$, this implies: $\frac{1}{n}\sum_{\ell=1}^{n}\tilde{x}_\ell e_\ell - \mathbb{E}\left[\frac{1}{n}\sum_{\ell=1}^{n}\tilde{x}_\ell e_\ell\right] = o_p(1)$. A similar argument for $\mathcal{S}_n = \frac{1}{n}\sum_{\ell=1}^{n}\tilde{x}_\ell^2 = \frac{1}{n}\sum_{\ell=1}^{n}\tilde{x}_\ell x_\ell$, where:

$$\Delta_i\mathcal{S}_n = \frac{2}{n}\sum_{\ell\in S_i}\tilde{x}_\ell x_\ell - \frac{1}{n}\sum_{\ell,\ell^*\in S_i}M_{\ell\ell^*}x_\ell x_{\ell^*},$$

implies: $\frac{1}{n}\sum_{\ell=1}^{n}\tilde{x}_\ell^2 - \mathbb{E}\left[\frac{1}{n}\sum_{\ell=1}^{n}\tilde{x}_\ell^2\right] = o_p(1)$. By Assumption 6(ii), we have: $\frac{1}{n}\sum_{\ell=1}^{n}\tilde{x}_\ell^2 = Q + o_p(1)$, with $Q > 0$. Therefore, combining with (8) and applying the continuous mapping theorem, the result follows. □

Proof of Lemma 2. The following example borrows from Chapter 2.1 of Lehmann and Casella (1998). Consider i.i.d. Bernoulli variables $\{z_i\}_{i=1}^{n}$ with $p \in (0,1)$ and i.i.d. Rademacher variables $\{q_i\}_{i=1}^{n}$, independent of $\{z_i\}$. Define $x_i = q_i(\sqrt{2}\,z_i + 1 - z_i)$, $\quad y_i = q_i(z_i/\sqrt{2} + 1 - z_i)$, $\quad i = 1, \ldots, n$. Then $x_i y_i = 1$ a.s., $x_i \in \{\sqrt{2}, -\sqrt{2}, 1, -1\}$ with probabilities $p/2, p/2, (1-$

$p)/2, (1-p)/2$, and $\mathbb{E}[x_i] = \mathbb{E}[y_i] = 0$. The data satisfy the correctly specified regression

$$y_i = \beta x_i + e_i, \quad \beta = \frac{\mathbb{E}[y_i x_i]}{\mathbb{E}[x_i^2]} = \frac{1}{1+p}, \quad \mathbb{E}[e_i] = \mathbb{E}[x_i e_i] = 0.$$

Thus, this is one of the distributions from class $\mathcal{F}$. Any estimator $u(x, y)$ has expectation

$$\mathbb{E}[u(x,y)] = \sum_{(x_1,...,x_n,y_1,...,y_n)} u(x_1, ..., x_n, y_1, ..., y_n)\, p^k (1-p)^{n-k} 2^{-n},$$

where k is the number of indices with $z_i = 1$. Hence $\mathbb{E}[u]$ is a polynomial in p of degree at most n and cannot equal $\beta = (1+p)^{-1}$ for all $p \in (0, 1)$. Thus, no unbiased estimator exists for this much simpler model. □

Proof of Lemma 3. A possible data generating process has $\mathbb{E}[x_{\tilde{\ell}} e_\ell] = M_{\tilde{\ell}\ell}(1 - \mathcal{E}_{\tilde{\ell}\ell})$. For the outcome model, the lemma then follows from (8). The derivation for the design-based model is similar. □

Proof of Lemma 4. Fix a column coordinate j. Take $F \in \mathcal{F}$ such that $y = x\beta + W\delta_j + e$ where δ_j is the unit vector with 1 in the j-th position. In addition, let the marginal distribution of x be a point mass at x_0, and let e be independent of x. In this case, for $u(x, y)$ to be correctly centered on F, we have

$$\mathbb{E}_F[x'Ay] - \beta \mathbb{E}_F[x'Ax] = \mathbb{E}_F[x'AW\delta_j] = x_0' AW_j = 0.$$

This holds for all values of x_0. This implies $AW_j = 0$ for all j, hence $AW = 0$, i.e., $AM = A$.

Finally, take a pair of indexes $(\tilde{\ell}_0, \ell_0)$ such that $\mathcal{E}_{\tilde{\ell}_0, \ell_0} = 0$. Take $F \in \mathcal{F}$ such that $\delta = 0, \beta = \beta_0$ and $\mathbb{E}_F[x_{\tilde{\ell}} e_\ell] = 0$ for all $(\tilde{\ell}, \ell) \neq (\tilde{\ell}_0, \ell_0)$ and $\mathbb{E}_F[x_{\tilde{\ell}_0} e_{\ell_0}] = 1$. Let $u(x, y) = \frac{x'Ay}{x'Ax}$ be correctly centered, then

$$0 = \mathbb{E}_F[x'Ay] - \beta_0 \mathbb{E}_F[x'Ax] = \mathbb{E}_F[x'Ae] = \sum_{\ell, \tilde{\ell}} A_{\tilde{\ell}\ell} E_F[x_{\tilde{\ell}} e_\ell] = A_{\tilde{\ell}_0 \ell_0} \mathbb{E}_F[x_{\tilde{\ell}_0} e_{\ell_0}].$$

Hence we must have $A_{\tilde{\ell}\ell} = 0$ whenever $\mathcal{E}_{\tilde{\ell}\ell} = 0$. □

Proof of Lemma 5. Consider the optimization problem (5), and suppose the minimum is attained at $z^* = \tilde{A}'x$. For any other $z = A'x$, with $A \in \mathcal{A}$, define a linear combination

$z_a = z^* + az$, which also lies in the same class. Minimizing $\mathbb{E}\|z_a - x\|^2$ over a, the first-order condition at $a = 0$ yields: $\mathbb{E}[z'z^*] = \mathbb{E}[z'x]$. Now evaluate the objective function:

$$V(A) = \frac{\sigma^2 \mathbb{E}[z'z]}{[\mathbb{E}(z'x)]^2} = \frac{\sigma^2 \mathbb{E}[z'z]}{[\mathbb{E}(z'z^*)]^2} \geq \frac{\sigma^2}{\mathbb{E}[z^{*\prime}z^*]} = V(\tilde{A}),$$

where the inequality follows from the Cauchy–Schwarz inequality and the first-order condition $\mathbb{E}[z^{*\prime}z^*] = \mathbb{E}[z^{*\prime}x]$. Finally, observe that the objective in (5) simplifies as:

$$\mathbb{E}\|z - x\|^2 = \operatorname{tr}\left[(A' - I_n)\mathbb{E}[xx'](A - I_n)\right] = \lambda\|A - I_n\|_F^2,$$

which matches the objective in (4). Hence, $\tilde{A} = A^*$ minimizes both objectives. □

Proof of Theorem 1. **Part (i).** Consider the optimization problem:

$$\|A - M\|_F^2 = (\operatorname{vec}(A) - \operatorname{vec}(M))'(\operatorname{vec}(A) - \operatorname{vec}(M)) \to \min \quad \text{s.t. } \mathcal{L}\operatorname{vec}(A) = 0.$$

The solution is given by: $\operatorname{vec}(A^*) = \left(I_{n^2} - \mathcal{L}'(\mathcal{L}\mathcal{L}')^+\mathcal{L}\right)\operatorname{vec}(M)$.

Part (ii). Since $\mathcal{A}$ is a linear subspace and the Frobenius norm arises from an inner product, the solution to the projection problem

$$A^* = \arg\min_{A \in \mathcal{A}} \|A - M\|_F^2 = \operatorname{proj}_F(M, \mathcal{A})$$

is the orthogonal projection of M onto $\mathcal{A}$. Hence, $M - A^*$ is orthogonal to any $A \in \mathcal{A}$ in Frobenius inner product: $\langle M - A^*, A\rangle_F = 0$.

Consider a matrix B with property $B_{\tilde{\ell}\ell} = 0$ whenever $\mathcal{E}_{\tilde{\ell}\ell} = 1$, then

$$\langle A, BM\rangle_F = \operatorname{tr}(A'BM) = \operatorname{tr}(MA'B) = \operatorname{tr}((AM)'B) = \operatorname{tr}(A'B) = \sum_{\ell_1,\ell_2} A_{\ell_1,\ell_2}B_{\ell_1,\ell_2} = 0,$$

since for any pair (ℓ_1, ℓ_2), either $B_{\ell_1,\ell_2} = 0$ or $A_{\ell_1,\ell_2} = 0$ for all $A \in \mathcal{A}$. then $\langle BM, A\rangle_F = 0$ for all $A \in \mathcal{A}$. If there exists a matrix B with that property and such that $A^* = M - BM \in \mathcal{A}$, then A^* must be the projection of M onto $\mathcal{A}$. To find such B, we need $(M - BM)_{\tilde{\ell}\ell} = 0$

whenever $\mathcal{E}_{\tilde{\ell}\ell} = 0$. This condition leads to the linear system:

$$M_{\tilde{\ell}\ell} - \sum_{\ell_1 : \mathcal{E}_{\tilde{\ell}\ell_1} = 0} B_{\tilde{\ell}\ell_1} M_{\ell_1 \ell} = 0, \qquad \text{whenever } \mathcal{E}_{\tilde{\ell}\ell} = 0.$$

This system has as many equations as unknowns. Define $M_{\tilde{\ell}}$ to be sub-matrix of M containing only elements $M_{\ell_1\ell_2}$ for indexes such that $\mathcal{E}_{\tilde{\ell}\ell_1} = 0$ and $\mathcal{E}_{\tilde{\ell}\ell_2} = 0$. If the matrix $M_{\tilde{\ell}}$ is full rank, the solution is:

$$B_{\tilde{\ell}\ell_1} = \sum_{\ell : \mathcal{E}_{\tilde{\ell}\ell} = 0} M_{\tilde{\ell}\ell} \left(M_{\tilde{\ell}}^{-1} \right)_{\ell\ell_1}, \quad \text{for } \ell_1 \text{ s.t. } \mathcal{E}_{\tilde{\ell}\ell_1} = 0, \text{ and } 0 \text{ otherwise.}$$

If $M_{\tilde{\ell}}$ is rank-deficient, the system still has a solution if having $\sum_{\ell : \mathcal{E}_{\tilde{\ell}\ell} = 0} c_\ell M_{\ell_1,\ell} = 0$ for all ℓ_1 such that $\mathcal{E}_{\tilde{\ell}\ell_1} = 0$ implies that $\sum_{\ell : \mathcal{E}_{\tilde{\ell}\ell} = 0} c_\ell M_{\tilde{\ell}\ell} = 0$. This holds since $M_{\ell_1,\ell} = \tilde{w}_{\ell_1}' \tilde{w}_\ell$, where $\tilde{w}_\ell = (W'W)^{-1/2} w_\ell$, so:

$$\sum_{\ell : \mathcal{E}_{\tilde{\ell}\ell} = 0} c_\ell M_{\ell_1,\ell} = 0 \Rightarrow \sum_{\ell : \mathcal{E}_{\tilde{\ell}\ell} = 0} c_\ell \tilde{w}_\ell = 0 \Rightarrow \sum_{\ell : \mathcal{E}_{\tilde{\ell}\ell} = 0} c_\ell M_{\tilde{\ell}\ell} = 0.$$

Thus, the system is solvable, and we can choose $B_{\tilde{\ell}\ell_1} = \sum_{\ell : \mathcal{E}_{\tilde{\ell}\ell} = 0} M_{\tilde{\ell}\ell} \left(M_{\tilde{\ell}}^{+} \right)_{\ell\ell_1}$ as a concrete solution using a generalized inverse.

Part (iii). Let $U = W'W$, and define selection matrices $\mathbf{i}_{\tilde{\ell}} = \text{diag}(\mathcal{E}_{\tilde{\ell},\cdot})$ and $\mathbf{i}_{\tilde{\ell}^c} = I_n - \mathbf{i}_{\tilde{\ell}}$ as diagonal matrices selecting observations whose errors are uncorrelated and correlated with $x_{\tilde{\ell}}$, respectively. Let $W_{\tilde{\ell}} = \mathbf{i}_{\tilde{\ell}} W, \quad W_{\tilde{\ell}^c} = \mathbf{i}_{\tilde{\ell}^c} W$, so that:

$$U = W'W = W_{\tilde{\ell}}' W_{\tilde{\ell}} + W_{\tilde{\ell}^c}' W_{\tilde{\ell}^c}.$$

By the generalized Woodbury identity (Henderson and Searle, 1981):

$$(W_{\tilde{\ell}}' W_{\tilde{\ell}})^+ = U^{-1} + U^{-1} W_{\tilde{\ell}^c}' (I_n - W_{\tilde{\ell}^c} U^{-1} W_{\tilde{\ell}^c}')^+ W_{\tilde{\ell}^c} U^{-1}.$$

Using notation: $M = I_n - WU^{-1}W', \quad M^{(\tilde{\ell})} = I_n - W_{\tilde{\ell}}(W_{\tilde{\ell}}' W_{\tilde{\ell}})^+ W_{\tilde{\ell}}'$. The term

$$\mathbf{i}_{\tilde{\ell}^c}' (I_n - W_{\tilde{\ell}^c} U^{-1} W_{\tilde{\ell}^c}')^+ \mathbf{i}_{\tilde{\ell}^c} = \mathbf{i}_{\tilde{\ell}^c} (\mathbf{i}_{\tilde{\ell}^c} M \mathbf{i}_{\tilde{\ell}^c})^+ \mathbf{i}_{\tilde{\ell}^c}$$

is a submatrix of M. Pre- and post-multiplying the generalized Woodbury formula by the $\tilde{\ell}$-th and $\tilde{\ell}_1$-th rows of W, we obtain:

$$M^{(\tilde{\ell})}_{\tilde{\ell}\tilde{\ell}_1} = M_{\tilde{\ell}\tilde{\ell}_1} - \sum_{\ell_1,\ell_2:\mathcal{E}_{\tilde{\ell}\ell_1}=0,\mathcal{E}_{\tilde{\ell}\ell_2}=0} M_{\tilde{\ell}\ell_1}(M^{+}_{\tilde{\ell}})_{\ell_1,\ell_2}M_{\ell_2,\tilde{\ell}_1} = M_{\tilde{\ell}\tilde{\ell}_1} - \sum_{\ell_2:\mathcal{E}_{\tilde{\ell}\ell_2}=0} B_{\tilde{\ell}\ell_2}M_{\ell_2,\tilde{\ell}_1} = A^{*}_{\tilde{\ell}\tilde{\ell}_1},$$

where the last equality follows from part (ii). □

Lemma 9. *Assume a random vector $\xi \in \mathbb{R}^T$ satisfies $\mathbb{E}[\xi] = 0$, $\mathbb{V}(\xi) = I_T$, and for all non-random vectors τ, $\mathbb{E}|\tau'\xi|^4 \leq C\|\tau\|^4$. Then for any positive definite matrix $A \in \mathbb{R}^{T\times T}$,*

$$\mathbb{E}[(\xi' A\xi)^2] \leq C(\mathrm{tr}(A))^2.$$

Proof of Lemma 9. As A is positive definite, it has spectral decomposition $A = \sum_{j=1}^{T} \mu_j\psi_j\psi_j'$, where $\mu_j > 0$ and $\{\psi_j\}$ is an orthonormal set of eigenvectors. Then:

$$\begin{aligned}\mathbb{E}[(\xi' A\xi)^2] &= \mathbb{E}\left(\sum_{j=1}^{T} \mu_j(\psi_j'\xi)^2\right)^2 = \sum_{j,k=1}^{T} \mu_j\mu_k\mathbb{E}[(\psi_j'\xi)^2(\psi_k'\xi)^2] \\ &\leq \sum_{j,k=1}^{T} \mu_j\mu_k\sqrt{\mathbb{E}(\psi_j'\xi)^4}\sqrt{\mathbb{E}(\psi_k'\xi)^4} \leq C\sum_{j,k=1}^{T} \mu_j\mu_k = C(\mathrm{tr}(A))^2.\end{aligned}$$ □

Proof of Lemma 6. We use Lemma A2.1 from Sølvsten (2020). Define:

$$\Delta_i W = \frac{1}{\sqrt{V}}(\xi_i - \tilde{\xi}_i)'\sum_{j\neq i}\Gamma_{ij}\xi_j,$$

where $\tilde{\xi}_i$ is an independent copy of ξ_i. To apply Lemma A2.1, we must verify:

$$\sum_{i=1}^{N}\mathbb{E}[(\Delta_i W)^2] = O(1) \quad \text{and} \quad \sum_{i=1}^{N}\mathbb{E}[(\Delta_i W)^4] \to 0.$$

For the second moment: $\sum_i \mathbb{E}[(\Delta_i W)^2] = \frac{2}{V}\sum_{i=1}^{N}\sum_{j\neq i}\mathrm{tr}(\Gamma_{ij}\Gamma_{ij}') = 1$. For the fourth moment, decompose the expression into two terms.

$$\sum_i \mathbb{E}[|\Delta_i W|^4] = \frac{4}{V^2}\sum_{i=1}^N \mathbb{E}[\sum_{j\neq i}(\xi_i'\Gamma_{ij}\xi_j)^4 + \sum_{j\neq i}\sum_{k\notin\{i,j\}}(\xi_i'\Gamma_{ij}\xi_j)^2(\xi_i'\Gamma_{ik}\xi_k)^2].$$

The first term is:

$$\begin{aligned}
&\frac{4}{V^2}\mathbb{E}\left\{\sum_{i=1}\sum_{j\neq i}\mathbb{E}\left[(\xi_i'\Gamma_{ij}\xi_j)^4|\xi_j\right]\right\} \leq \frac{C}{V^2}\mathbb{E}\sum_i\sum_{j\neq i}\|\Gamma_{ij}\xi_j\|^4 \\
&= \frac{C}{V^2}\mathbb{E}\left\{\sum_i\sum_{j\neq i}\left(\xi_j'\Gamma_{ij}'\Gamma_{ij}\xi_j\right)^2\right\} \leq \frac{C}{V^2}\sum_i\sum_{j\neq i}\left(\mathrm{tr}(\Gamma_{ij}'\Gamma_{ij})\right)^2 = \frac{C}{V^2}\sum_i\sum_{j\neq i}\|\Gamma_{ij}\|_F^4,
\end{aligned}$$

where we used Lemma 9. For the second term, we notice that

$$\begin{aligned}
&\mathbb{E}[(\xi_i'\Gamma_{ij}\xi_j)^2(\xi_i'\Gamma_{ik}\xi_k)^2] = \mathbb{E}(\xi_i'\Gamma_{ij}\Gamma_{ij}'\xi_i\xi_i'\Gamma_{ik}\Gamma_{ik}'\xi_i) \\
&\leq \sqrt{\mathbb{E}[(\xi_i'\Gamma_{ij}\Gamma_{ij}'\xi_i)^2]}\sqrt{\mathbb{E}[(\xi_i'\Gamma_{ik}\Gamma_{ik}'\xi_i)^2]} \leq C\|\Gamma_{ij}\|_F^2\|\Gamma_{ik}\|_F^2.
\end{aligned}$$

Combining both yields

$$\sum_i \mathbb{E}|\Delta_i W|^4 \leq \frac{C}{V^2}\sum_i\sum_{j\neq i}\left(\|\Gamma_{ij}\|_F^4 + \sum_{k\notin\{i,j\}}\|\Gamma_{ij}\|_F^2\|\Gamma_{ik}\|_F^2\right) \leq \frac{C\max_i\left(\sum_{j\neq i}\|\Gamma_{ij}\|_F^2\right)}{V} \to 0.$$

Condition (i) of Lemma A2.1 of Sølvsten (2020): We need to verify:

$$\frac{2}{V}\sum_{i=1}^N\sum_{k>i}\sum_{j\neq i}\xi_j'\Gamma_{ji}(\xi_i\xi_i' + I_{T_i})\Gamma_{ik}\xi_k \xrightarrow{p} 1.$$

This expression decomposes into five terms: $2I_1 + 4I_2 + I_3 + 6I_4 + \frac{4}{V}\sum_{i,k>i}\|\Gamma_{ik}\|_F^2$. The final term converges to 1 as $V = 2\|\Gamma\|_F^2$, so it suffices to show that $I_1, I_2, I_3, I_4 \xrightarrow{p} 0$.

$$I_1 = \frac{1}{V}\sum_{i=1}^{N}\sum_{k>i}\sum_{j\notin\{i,k\}} \xi_j'\Gamma_{ji}(\xi_i\xi_i' - I_{T_i})\Gamma_{ik}\xi_k \xrightarrow{p} 0,$$

$$I_2 = \frac{1}{V}\sum_{i=1}^{N}\sum_{k>i}\sum_{j\notin\{i,k\}} \xi_j'\Gamma_{ji}\Gamma_{ik}\xi_k \xrightarrow{p} 0,$$

$$I_3 = \frac{2}{V}\sum_{i=1}^{N}\sum_{k>i} \operatorname{tr}\left(\Gamma_{ik}(\xi_k\xi_k' - I)\Gamma_{ki}(\xi_i\xi_i' - I)\right) \xrightarrow{p} 0,$$

$$I_4 = \frac{2}{V}\sum_{i=1}^{N}\sum_{k>i} \operatorname{tr}\left((\xi_k\xi_k' - I)\Gamma_{ki}\Gamma_{ik}\right) \xrightarrow{p} 0.$$

In all cases, the expectation of the left-hand side is zero, so we only need to check that the variance of each expression converges to zero.

$$\mathbb{V}(I_1) \leq \frac{C}{V^2}\sum_{i=1}^{N}\sum_{k\neq i}\sum_{j\notin\{i,k\}} \mathbb{E}\left(\operatorname{tr}(\Gamma_{ji}(\xi_i\xi_i' - I_{T_i})\Gamma_{ik}\Gamma_{ki}(\xi_i\xi_i' - I_{T_i})\Gamma_{ij}) + (\xi_j'\Gamma_{ji}\xi_i)^2(\xi_i'\Gamma_{ik}\Gamma_{kj}\xi_j)\right)$$

$$= \frac{C}{V^2}\sum_{i,k}^{N}\sum_{j\neq k}\left[\mathbb{E}\left(\xi_i'\Gamma_{ij}\Gamma_{ji}\xi_i\xi_i'\Gamma_{ik}\Gamma_{ki}\xi_i\right) - \operatorname{tr}(\Gamma_{ji}\Gamma_{ik}\Gamma_{ki}\Gamma_{ij})\right] + \frac{C}{V^2}\sum_{i,k,j}^{N}\mathbb{E}[(\xi_j'\Gamma_{ji}\xi_i)^2(\xi_i'\Gamma_{ik}\Gamma_{kj}\xi_j)].$$

Here we used the circular property of traces, the fact that ξ's with different indexes are independent with identity covariance, and $\Gamma_{jj} = 0$. The first summand of the last sum is bounded by

$$\frac{C}{V^2}\left(\sum_{i=1}^{N}\mathbb{E}\left(\xi_i'\bar{\Gamma}_{ii}\xi_i\right)^2 - \operatorname{tr}(\Gamma^4) + \sum_{i,j}\operatorname{tr}(\Gamma_{ji}\Gamma_{ij}\Gamma_{ji}\Gamma_{ij})\right) \leq \frac{C}{V^2}\left(\sum_i(\operatorname{tr}(\bar{\Gamma}_{ii}))^2 + \sum_{i,j}\|\Gamma_{ij}\|_F^4\right)$$

$$\leq \frac{C}{V^2}\left(\sum_i(\sum_j\|\Gamma_{ij}\|_F^2)^2 + \sum_{i,j}\|\Gamma_{ij}\|_F^4\right) \leq \frac{C}{V}\max_i\sum_j\|\Gamma_{ij}\|_F^2 \to 0.$$

Here we used $\mathrm{tr}(\Gamma_{ji}\Gamma_{ij}\Gamma_{ji}\Gamma_{ij}) = \|\Gamma_{ji}\Gamma_{ij}\|_F^2 \le \|\Gamma_{ji}\|^2\|\Gamma_{ij}\|_F^2 \le \|\Gamma_{ij}\|_F^4$. Denote $\bar{\Gamma} = \Gamma^2$. Then

$$\frac{C}{V^2}\sum_{i,k,j=1}^N \left|\mathbb{E}[(\xi_j'\Gamma_{ji}\xi_i)^2(\xi_i'\Gamma_{ik}\Gamma_{kj}\xi_j)]\right| = \frac{C}{V^2}\sum_{i,j=1}^N \left|\mathbb{E}[(\xi_j'\Gamma_{ji}\xi_i)^2(\xi_i'\bar{\Gamma}_{ij}\xi_j)]\right|$$
$$\le \frac{C}{V^2}\sum_{i,j=1}^N \sqrt{\mathbb{E}(\xi_j'\Gamma_{ji}\xi_i)^4}\sqrt{\mathbb{E}(\xi_i'\bar{\Gamma}_{ij}\xi_j)^2} \le \frac{C}{V^2}\sum_{i,j=1}^N \|\Gamma_{ji}\|_F^2\cdot\|\bar{\Gamma}_{ij}\|_F \le \frac{C\max_{i,j}\|\bar{\Gamma}_{ij}\|_F}{V}.$$

We notice that $\|\bar{\Gamma}_{ij}\|_F = \|\sum_k \Gamma_{ik}\Gamma_{kj}\|_F \le \sqrt{\sum_k\|\Gamma_{ik}\|_F^2}\sqrt{\sum_k\|\Gamma_{kj}\|_F^2} \le \max_i\sum_j\|\Gamma_{ij}\|_F^2$. Thus, $I_1 \xrightarrow{p} 0$. All summands in I_2 are uncorrelated unless the set of indexes j,k coincides.

$$\mathbb{V}(I_2) \le \frac{C}{V^2}\sum_{i,l=1^N}\sum_{k>i,k>l}\sum_{j\ne i}\mathrm{tr}(\Gamma_{ji}\Gamma_{ik}\Gamma_{kl}\Gamma_{lj}) = \frac{C}{V^2}\mathrm{tr}(\Gamma^2\tilde{\Gamma}) \le \frac{C}{V^2}\|\Gamma^2\|\|\tilde{\Gamma}\|_F^2 \le \frac{C\|\Gamma^2\|}{\|\Gamma\|_F^2} \to 0.$$

In this argument we defined $\tilde{\Gamma}_{il} = \sum_{k>i,k>l}\Gamma_{ik}\Gamma_{kl}$ and $\|\tilde{\Gamma}\|_F^2 \le \|\Gamma\|_F^2$. This gives $I_2 \xrightarrow{p} 0$.

Introduce $\eta_{ik} = \mathrm{tr}\left(\Gamma_{ik}(\xi_k\xi_k' - I)\Gamma_{ki}(\xi_i\xi_i' - I)\right)$.

$$\mathbb{E}[\eta_{ik}^2] = \mathbb{E}\left(\xi_k'\Gamma_{ki}(\xi_i\xi_i' - I)\Gamma_{ik}\xi_k - \mathrm{tr}(\Gamma_{ik}\Gamma_{ki}(\xi_i\xi_i' - I))\right)^2$$
$$= \mathbb{E}[\mathbb{V}(\xi_k'\Gamma_{ki}(\xi_i\xi_i' - I)\Gamma_{ik}\xi_k|\xi_i)] \le \mathbb{E}[\mathbb{E}[(\xi_k'\Gamma_{ki}(\xi_i\xi_i' - I)\Gamma_{ik}\xi_k)^2|\xi_i]]$$
$$\le C\mathbb{E}\left(\mathrm{tr}(\Gamma_{ki}(\xi_i\xi_i' - I)\Gamma_{ik})\right)^2 = C\mathbb{E}\left(\xi_i'\Gamma_{ik}\Gamma_{ki}\xi_i - \|\Gamma_{ik}\|_F^2\right)^2$$
$$\le C\mathbb{E}\left(\xi_i'\Gamma_{ik}\Gamma_{ki}\xi_i\right)^2 \le C\|\Gamma_{ik}\|_F^4.$$

Notice that η_{ik} are uncorrelated with any other η's except η_{ki}. Thus

$$\mathbb{V}(I_3) = \mathbb{V}\left(\frac{1}{V}\sum_{i=1}^N\sum_{k\ne i}\eta_{ik}\right) = \frac{2}{V^2}\sum_{i=1}^N\sum_{k\ne i}\mathbb{E}[\eta_{ik}^2] \le \frac{C\max_{ik}\|\Gamma_{ik}\|_F^2}{V} \to 0.$$

Finally, to prove $I_4 \xrightarrow{p} 0$ we notice that all summands are independent, thus

$$\mathbb{V}\left(\frac{1}{V}\sum_{k=1}^N\xi_k'\bar{\Gamma}_{kk}\xi_k\right) = \frac{1}{V^2}\sum_{k=1}^N\mathbb{V}\left(\xi_k'\bar{\Gamma}_{kk}\xi_k\right) \le \frac{1}{V^2}\sum_{k=1}^N\mathbb{E}\left(\xi_k'\bar{\Gamma}_{kk}\xi_k\right)^2$$
$$\le C\frac{1}{V^2}\sum_{k=1}^N(\mathrm{tr}\,\bar{\Gamma}_{kk})^2 = C\frac{1}{V^2}\sum_{k=1}^N(\mathrm{tr}\sum_j\Gamma_{kj}\Gamma_{jk})^2 = C\frac{1}{V^2}\sum_{k=1}^N(\sum_j\|\Gamma_{kj}\|_F^2)^2 \to 0. \qquad \square$$

Proof of Theorem 2. Define $\tilde{A}_{ij} = \Sigma_i^{1/2} \begin{pmatrix} 0 & 0 \\ A_{ij} & 0 \end{pmatrix} \Sigma_j^{1/2}$ to be a $(2T_i) \times (2T_j)$ matrix. Construct the full $(2n) \times (2n)$ matrix $\tilde{A}$ with $\tilde{A}_{ij}$ in block (i, j) and $\tilde{A}_{ii} = 0$ for all i. Note that $\tilde{A}$ is generally not symmetric. The variance of the quadratic component can be written as:

$$\mathbb{V}\left(\sum_j \sum_{i \neq j} v_i' A_{ij} e_j\right) = \mathbb{V}\left(\sum_j \sum_{i \neq j} \xi_i' \tilde{A}_{ij} \xi_j\right) = \sum_j \sum_{i \neq j} \mathrm{tr}(\tilde{A}_{ij} \tilde{A}_{ij}' + \tilde{A}_{ij} \tilde{A}_{ji}) = \mathrm{tr}(\tilde{A}\tilde{A}') + \mathrm{tr}(\tilde{A}^2).$$

The first term, $\mathrm{tr}(\tilde{A}\tilde{A}')$, corresponds to $\sum_j \mathbb{V}(Q_j e_j)$ in the decomposition of Equation (7). The second term, $\mathrm{tr}(\tilde{A}^2)$, captures the correction due to cross-cluster dependence.

If two random variables η_1 and η_2 have correlation bounded by $1 - c$, then $\mathbb{V}(\eta_1 + \eta_2) \geq c(\mathbb{V}(\eta_1) + \mathbb{V}(\eta_2))$. Assumption 4(i) on bounded alignment and Assumption 3(i) imply

$$\max_{\tau_1, \tau_2} \left|\mathrm{corr}(\tau_1' e_i, \tau_2'(v_i \otimes e_i))\right| \leq 1 - c,$$

$$\omega^2 \geq \sum_{j=1}^N c\left(\mathbb{V}((\lambda' A)_j e_j) + \mathbb{V}(\mathrm{vec}(A_{jj}')'(v_j \otimes e_j))\right) + \mathrm{tr}(\tilde{A}\tilde{A}').$$

Notice that $\mathrm{tr}(\tilde{A}\tilde{A}') = \|\tilde{A}\|_F^2 \geq C^{-2} \sum_{i \neq j} \|A_{ij}\|_F^2$ due to Assumption 3(i).

$$\mathbb{V}(v_j' A_{jj} e_j) = \mathbb{V}(\mathrm{vec}(A_{jj}')'(v_j \otimes e_j)) \geq c \cdot \mathrm{vec}(A_{jj}')'(\Sigma_{v,j} \otimes \Sigma_{e,j}) \mathrm{vec}(A_{jj}') = c\|A_{jj}\|_F^2.$$

Combining the above with Assumption 4(i), we conclude $\omega^2 \geq cn$.

Next, recall that: $x' A e = \sum_{j=1}^N \omega_j e_j + \sum_{j=1}^N \sum_{i \neq j} v_i' A_{ij} e_j$. We first establish asymptotic normality for the linear component, $\sum_{j=1}^N \omega_j e_j$, using Lyapunov's condition. For $\sum_j (\lambda' A)_j e_j$, Lyapunov's condition holds due to bounded fourth moments from Assumption 3(iv). For $\sum_j v_j' A_{jj} e_j$, we apply the higher-moment bound in Assumption 3(iv):

$$\begin{aligned} \mathbb{E}|v_j' A_{jj} e_j|^{2+2\delta} &\leq \mathbb{E}|\mathrm{vec}(A_{jj}')'(v_j \otimes e_j)|^{2+2\delta} \leq C\left(\mathrm{vec}(A_{jj}')' \Phi_j \mathrm{vec}(A_{jj}')\right)^{1+\delta} \\ &\leq C\left(\mathrm{vec}(A_{jj}')'(\Sigma_{v,j} \otimes \Sigma_{e,j}) \mathrm{vec}(A_{jj}')\right)^{1+\delta} = C\|\tilde{A}_{jj}\|_F^{2+2\delta}. \end{aligned}$$

Hence: $\sum_{j=1}^N \mathbb{E}|v_j' A_{jj} e_j|^{2+2\delta} \leq C \max_j \|\tilde{A}_{jj}\|_F^{2\delta} \sum_{j=1}^N \|\tilde{A}_{jj}\|_F^2$. Using Assumption 4(ii), which implies $\frac{\max_j \|\tilde{A}_{jj}\|_F^2}{n} \to 0$, we have:

$$\frac{\sum_{j=1}^N \mathbb{E}\left|v_j' A_{jj} e_j\right|^{2+2\delta}}{\omega^{1+\delta}} \le C \frac{\max_j \|\tilde{A}_{jj}\|_F^{2\delta} \sum_{j=1}^N \|\tilde{A}_{jj}\|_F^2}{\left(\mathbb{V}(v_j' A_{jj} e_j)\right) \cdot \omega^\delta} \le C \frac{\max_j \|\tilde{A}_{jj}\|_F^{2\delta}}{n^\delta} \to 0.$$

Thus, the linear part satisfies a CLT: $\frac{1}{\omega} \sum_{j=1}^N \omega_j e_j \xrightarrow{d} \mathcal{N}(0,1)$.

We now apply Lemma 6 to the quadratic part: $\sum_{i=1}^N \sum_{j \neq i} v_i' A_{ij} e_j = \sum_{i,j} \xi_i' \tilde{A}_{ij} \xi_j$, with $\Gamma = \frac{1}{2}(\tilde{A} + \tilde{A}')$. Assumption 4(iii) ensures that either the quadratic term is asymptotically negligible compared to ω, or the conditions of Lemma 6 are satisfied. Because the linear and the quadratic terms are uncorrelated in finite samples, their asymptotic normality implies that they are asymptotically independent and hence jointly normal. □

Proof of Lemma 7. For the first statement notice that $\sum_{\tilde{\ell}} (A^*_{\ell\tilde{\ell}})^2 = A^*_{\ell\ell} \le 1$, so:

$$\sum_{j=1}^N \|A^*_{ij}\|_F^2 \le \sum_{\ell \in S_i} A^*_{\ell\ell} \le C \max_i T_i.$$

According to Theorem 1, $A^* = (I - B)M$, where B is sparse with non-zero elements $B_{\ell\tilde{\ell}}$ only if $\mathcal{E}_{\ell\tilde{\ell}} = 0$. If $\|M_\ell^{-1}\| \le C$, then $\|B_\ell\|^2 \le C$ and all elements of B are bounded. Thus, $\|I - B\|_\infty \le C \max_i T_i$,

$$\begin{aligned}
&\max_i (A'\lambda)_i' \Sigma_{e,i} (A'\lambda)_i \le C \max_i \|\Sigma_i\| \cdot \|(A'\lambda)_i\|_2^2 \le C \max_i \|\Sigma_i\| \cdot \max_i T_i \|(A'\lambda)\|_\infty^2 \\
&\le C \max_i \|\Sigma_i\| \cdot \max_i T_i \cdot \|M\|_\infty^2 \|1 - B\|_\infty^2 \|\lambda\|_\infty^2 \le C \max_i \|\Sigma_i\| \cdot \max_i T_i^3 \|M\|_\infty^2.
\end{aligned}$$

For the third statement:

$$\begin{aligned}
&\sum_\ell (A^*_{\ell\ell_1})^2 \le \sum_\ell M^2_{\ell\ell_1} + \sum_\ell \Big(\sum_{\tilde{\ell}:\mathcal{E}_{\ell\tilde{\ell}}=0} B_{\ell\tilde{\ell}} M_{\tilde{\ell}\ell_1}\Big)^2 \le M_{\ell_1,\ell_1} + \sum_\ell \|B_\ell\|^2 \sum_{\tilde{\ell}:\mathcal{E}_{\ell\tilde{\ell}}=0} M^2_{\tilde{\ell}\ell_1} \\
&\le 1 + C \sum_\ell \sum_{\tilde{\ell}:\mathcal{E}_{\ell\tilde{\ell}}=0} M^2_{\tilde{\ell}\ell_1} \le 1 + C(\max_i T_i) M_{\ell_1,\ell_1} \le C \max_i T_i.
\end{aligned}$$

Here we used the fact that $\mathcal{E}_{\ell\tilde{\ell}} = 0$ only within a cluster and thus has less than T_i elements.

$$\sum_{i=1}^{N} \|A^*_{ij}\|_F^2 \leq \sum_{\ell_1 \in S_j} \sum_{\ell=1}^{n} A^*_{\ell\ell_1} \leq C \max_k T_k^2.$$

From Theorem 1, B is block-diagonal with row/column sparsity bounded by $\max_i T_i$. Thus: $\|B\|^2 \leq \|B\|_1 \|B\|_\infty \leq C(\max_i T_i)^2$. Hence, $\|A^*\|^2 \leq \|(I - B)M\|^2 \leq C(\max_i T_i)^2$. □

Proof of Lemma 8. We start by expanding the jackknife difference:

$$\mathcal{Z} - \mathcal{Z}_{(j)} = -\mu_j + \omega_j e_j + \sum_{i \neq j} (\lambda_j' A^*_{ji} e_i + v_i' A^*_{ij} w_j' \delta) + \sum_{i \neq j} (v_i' A^*_{ij} e_j + v_j' A^*_{ji} e_i),$$

where $\mu_j = \sum_{i \neq j} \lambda_i' A^*_{ij} w_j' \delta$, $\omega_j = (A^* \lambda)_j + v_j' A^*_{jj}$. We now compute the squared expectation:

$$\mathbb{E}[(\mathcal{Z} - \mathcal{Z}_{(j)})^2] = \mu_j^2 + \mathbb{V}(\omega_j e_j) + \sum_{i \neq j} \mathbb{V}\left(\lambda_j' A^*_{ji} e_i + v_i' A^*_{ij} w_j' \delta\right) + \sum_{i \neq j} \mathbb{V}\left(v_i' A^*_{ij} e_j + v_j' A^*_{ji} e_i\right),$$

using the fact that all components are uncorrelated due to independence across clusters. Summing over j and comparing to the true variance $\mathbb{V}(\mathcal{Z})$, we obtain:

$$\mathbb{E}[\hat{V}_{\mathrm{JK}}] = \mathbb{V}(\mathcal{Z}) + \sum_j \mu_j^2 + \sum_{i=1}^{N} \sum_{j<i} \mathbb{V}\left(v_i' A^*_{ij} e_j + v_j' A^*_{ji} e_i\right) + \sum_j \sum_{i \neq j} \mathbb{V}\left(\lambda_j' A^*_{ji} e_i + v_i' A^*_{ij} w_j' \delta\right).$$

Therefore, $\mathbb{E}[\hat{V}_{\mathrm{JK}}] \geq \mathbb{V}(\mathcal{Z})$, and the jackknife variance estimator is conservative. In the special case $A^*_{ij} = 0$ for all $i \neq j$, all additional terms vanish, yielding $\mathbb{E}[\hat{V}_{\mathrm{JK}}] = \mathbb{V}(\mathcal{Z})$. □

B Simulation in Section 2.4

The DGP is a stochastic block model with $N = 50$ clusters of size $T = 10$, for a total of $n = 500$ units. Within each cluster, each pair of units forms an undirected edge independently with probability $p_{\mathrm{in}} = 0.3$; there are no cross-cluster edges. The weighted adjacency matrix G is defined as $G_{\ell k} = F_{\ell k} \cdot \omega_{\ell k}$, where $F_{\ell k} \in \{0, 1\}$ indicates whether ℓ and k are connected and $\omega_{\ell k} = \omega_{k\ell} \overset{\mathrm{iid}}{\sim} \mathrm{Exp}(1)$ for each edge. Crucially, G is *not* row-normalized: the row sums $s_\ell = \sum_k G_{\ell k}$ vary across units. Cluster-level treatment saturation is drawn as $\mu_i \overset{\mathrm{iid}}{\sim} U[0.1, 0.9]$

for $i = 1, \dots, N$. The network structure, edge weights, and cluster-level saturations are sampled once and held fixed across all simulation replications. Unit-level treatment drawn as $x_\ell \sim \text{Bernoulli}(\mu_{i(\ell)})$ independently across units and the outcome is $y_\ell = \beta x_\ell + \alpha \sum_k G_{\ell k} x_k + e_\ell$ where $e_\ell \overset{\text{iid}}{\sim} N(0,1)$. x_ℓ, e_ℓ are redrawn in each of $S = 2{,}000$ simulation replications and $\beta = 1$.