Using Multiple Outcomes to Adjust Standard Errors for Spatial Correlation

Stefano DellaVigna
UC Berkeley and NBER

Guido Imbens
Stanford University

Woojin Kim
NBER

David M. Ritzwoller
Stanford University

ABSTRACT. Empirical research in economics often examines the behavior of agents located in a geographic space. In such cases, statistical inference is complicated by the interdependence of economic outcomes across locations. A common approach to account for this dependence is to cluster standard errors based on a predefined geographic partition. A second strategy is to model dependence in terms of the distance between units. Dependence, however, does not necessarily stop at borders and is typically not determined by distance alone. This paper introduces a method that leverages observations of multiple outcomes to adjust standard errors for cross-sectional dependence. Specifically, a researcher, while interested in a particular outcome variable, often observes dozens of other variables for the same units. We show that these outcomes can be used to estimate dependence under the assumption that the cross-sectional correlation structure is shared across outcomes. We develop a procedure, which we call Thresholding Multiple Outcomes (TMO), that uses this estimate to adjust standard errors in a given regression setting. We show that adjustments of this form can lead to sizable reductions in the bias of standard errors in calibrated U.S. county-level regressions. Re-analyzing nine recent papers, we find that the proposed correction can make a substantial difference in practice.

Keywords: Spatial Correlation, Clustering, Large-Scale Inference
JEL: C10, C21, C31

Date: March 20, 2025
*Email: sdellavigna@econ.berkeley.edu, imbens@stanford.edu, woojin@berkeley.edu, ritzwoll@stanford.edu. We thank Ulrich Müller and Michael Pollmann for helpful comments and conversations. We are grateful to Malek Hassouneh, Rohan Jha, Chenxi Jiang, and Yi Yu for exceptional research assistance. We thank Desmond Ang for sharing data used in one of the applications. Imbens acknowledges support from the Office of Naval Research under grant numbers N00014-17-1-2131 and N00014-19-1-2468 and a gift from Amazon. Kim acknowledges support from the National Institute on Aging to the National Bureau of Economic Research through grant number T32-AG000186. Ritzwoller acknowledges support from the National Science Foundation through grant DGE-1656518.

1. Introduction

Empirical research in economics often considers data indexed by locations in space. In 2023 alone, nearly half—61 of 128—of all empirical, observational papers published in five leading, general interest journals primarily consider data with this characteristic.[1] The interpretation of inferences produced using these data is complicated by a ubiquitous form of dependence: economic outcomes are typically linked across locations in potentially complex ways, through direct causal relationships or a common underlying influence. Productivity shocks to one region may propagate to other areas through trade linkages or labor market flows (Greenstone et al., 2010; Bustos et al., 2016; Giroud et al., 2024), fiscal policies in one municipality may affect both the design of symmetric policies and economic outcomes in other locations (Case et al., 1993; DellaVigna and Kim, 2022), and political events in one jurisdiction may determine voter behavior in others (Besley and Case, 1995; Madestam et al., 2013).

Spatial dependence has the potential to induce substantial biases in statistical inferences that are made under the assumption that outcomes are independent across space. Existing methods for correcting standard errors fall into two broad categories. The first set of methods cluster standard errors based on a partition of the space. A prototypical instance of this approach allows for dependence between pairs of U.S. counties in the same state (for discussion, see Moulton 1986, 1990). The second set of methods parameterize dependence as a monotonic, decreasing function of geographic distance (Conley, 1999; Müller and Watson, 2022, 2023).

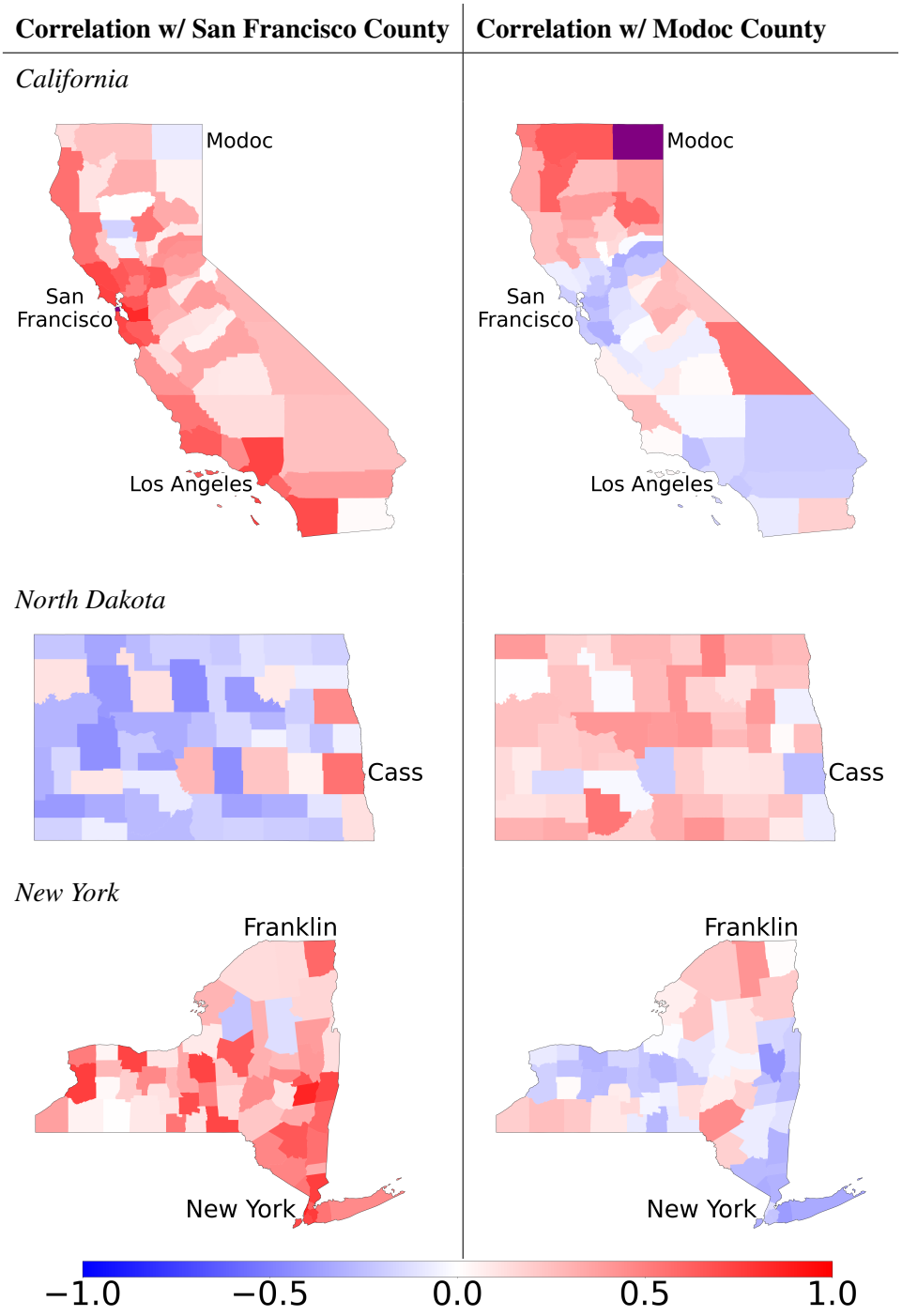
Both approaches impose strong, *ex ante* restrictions on how spatial dependence varies with geography, in one case with sharp boundaries, and in the other with smooth functions of distance. However, the correlation between economic outcomes at two locations need not be driven, in large or substantial part, by their geographic distance. Similarities in economic or political behavior, for example, are unlikely to be captured by methods based on modeling dependence with geographic proximity alone.

To illustrate this point, we collect a set of 91 U.S. county-level economic outcomes (including the unemployment rate, the average income, the poverty rate, etc.,) for each of the 3,144 U.S. counties.[2] We normalize each outcome to have mean zero and unit variance across the set of counties. The left column of Figure 1 displays the correlation between the economic outcomes of San Francisco County, California and the outcomes of every other county in the states of California, North Dakota, and New York. As one might expect, San Francisco is more correlated with other densely populated areas, such as Los Angeles County and New York County (which contains the borough of Manhattan), than with the less populated areas in the rural state of North Dakota. The same, intuitive, patterns hold among less populated areas. Take Modoc County, California, for instance. Modoc County is the third least populous county in the state of California, with an

[1] We hand-labeled the 370 papers published in the *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Review of Economic Studies*, and *Quarterly Journal of Economics* in 2023 according to whether they are, primarily, empirical, observational, and consider data indexed by locations in space. Further details concerning the criteria used to define these categories are given in Section 5. Examples of empirical analyses without a spatial aspect include those where units are products, or students in a single educational institution, or patents.

[2] We list these outcomes, and give further details concerning their construction, primarily from the U.S. Census, in Appendix B.1.

FIGURE 1. Correlation with Two California Counties



Notes: Figure 1 displays heat-maps giving the correlation of a collection of economic outcomes between pairs of U.S. counties. These outcomes are listed in Appendix B.1 and have been normalized to have mean zero and unit variance. The first column gives the correlation between outcomes in San Francisco County, California, with each county in the states of California, North Dakota, and New York. The second column is analogous, but gives correlations relative to Modoc County, California.

economy substantially driven by the agricultural sector (California Deprtment of Transportation, 2023). The right column of Figure 1 displays the correlation between the economic outcomes of Modoc County, and every other county in California, North Dakota, and New York. Modoc County is highly correlated with other rural, agricultural areas. In fact, Modoc County is more correlated with remote areas in North Dakota and New York than with more populated areas in California.[3]

Methods based on the assumption that cross-sectional dependence is determined by geographic proximity alone cannot capture the many of the main features of the estimates displayed in Figure 1, and so may not perform well in many regression problems considered in empirical economics. Indeed, in Section 2, we calibrate a simulation to the estimates of the correlation between pairs of U.S. counties across the outcomes introduced above. We find that, in this setting, many standard adjustments for spatial dependence yield standard errors that are badly biased.

In this paper, we introduce a method that uses observations of multiple outcomes to adjust standard errors for spatial dependence. We detail the proposal in Section 3. Often, in a single study, researchers observe many outcome variables for the same set of units. We give a condition under which collections of outcomes, of the type considered above, can be used to estimate dependence across units. In particular, rather than directly restricting cross-sectional dependence to be determined by, say, geographic proximity, we assume that the structure of cross-sectional correlation is shared across outcomes. This is a highly stylized restriction—so too are alternative assumptions. The main appeal is that, in settings where this assumption is reasonable, i.e., where the main features of cross-sectional correlation are captured by a collection of available outcomes, spatial dependence can be estimated directly from the data at hand.

We propose a procedure, which we call Thresholding Multiple Outcomes (TMO), for using estimates of spatial dependence to adjust standard errors in a given regression setting. The method consists of three steps. First, we estimate the correlation across locations using a collection of auxiliary outcomes. Second, we use the empirical distribution of these estimates to determine a threshold; pairs of locations whose correlation exceeds this threshold are treated as correlated, while pairs below the threshold are treated as uncorrelated. Third, we construct standard errors by allowing for dependence among location pairs treated as correlated. This approach is analogous to the construction of clustered standard errors. There, pairs of locations in the same cluster are assumed to correlate *a priori*. Here, we use the auxiliary outcomes to *determine* the pairs of locations that correlate. Revisiting the calibrated simulation from Section 2, we find that this procedure exhibits substantially smaller bias, and more accurate rejection rates, than existing methods.

Our procedure is premised on the hypothesis that, in many settings of practical interest in applied economics, pairs of locations can be effectively categorized as either "null" or "non-null." Outcomes at non-null location pairs exhibit meaningful correlation, whereas outcomes at null pairs do not. The main idea

[3]These results generalize beyond Modoc county. Figure A.1 displays the correlation in outcomes between each pair of counties in California, North Dakota, and New York. Counties are sorted by state and population. It is clear that both population and geography are important determinants of the correlation between two counties. Figure A.2 gives an analogous visualization for each pair of counties in the United States. The same pattern holds.

is that the empirical distribution of the estimates of the correlation between each pair of locations can be used to distinguish null pairs from non-null pairs. We appeal to the intuition that the center of this distribution should predominately consist of the correlations associated with null pairs. Thus, we estimate the null distribution using the center, and then extrapolate this estimate into the tails to infer the analogous distribution for the non-nulls. We rely on a Gaussian approximation to the null distribution. These inferred distributions can then be used to select a threshold that appropriately balances the contributions of null and non-null pairs to the error when constructing the resultant standard errors.

The main non-standard feature of our setting is that the data under consideration are correlated across both locations and outcomes. In Section 4, we consider how both sources of correlation impact the performance of the TMO estimator. We present two results. First, we give a quantitative description of the extent of the correlation across locations and outcomes under which spatial dependence is consistently estimable. Second, we characterize the quality of a Gaussian approximation to the empirical distribution of the null correlation estimates in terms of the extent of the correlation across outcomes. In both cases, the main takeaway is that our ability to recover spatial dependence is adversely impacted by the use of highly correlated outcomes. Thus, when applying the TMO estimator, researchers face a trade-off: pairs of outcomes with the most similar cross-sectional correlation are likely to be, themselves, more correlated. We provide recommendations on how to best balance these considerations in the remaining two sections.

In Section 5, we illustrate the application of TMO standard errors to nine recent papers. We begin by documenting that the potential for spatial correlation is widespread in empirical economics. In a set of 370 economics papers published in 2023 in five leading journals, we identify 126 observational papers, 48% of which have the potential for spatial correlation to affect a main estimate in the paper. The most common geographic unit in such papers is the U.S. county, employed in 15 papers. We re-analyze seven of these papers, which differ in their economic context and underlying methodology. For each paper, we identify a set of county-level outcomes, taken from each paper's replication packages as well as from external sources. We use these outcomes to construct the TMO estimator.

When applied to these papers, we find that the TMO estimator can substantively change the estimated standard errors, with a median increase of 37%. We then compare TMO to several leading alternatives based on modeling dependence in terms of geographic proximity: clustered standard errors, Conley (1999), and Müller and Watson (2022, 2023). In some cases, the TMO standard error leads to a similar adjustment. In others, the resultant standard errors are qualitatively different. To show that the method is applicable to other units of observation, we apply the TMO estimator to two additional papers—one that studies commuting zones and another that studies countries.

We conclude in Section 6 by detailing our recommendations for practice. We emphasize rules-of-thumb concerning what types of data our method is best suited for, diagnostics for determining that the method is performing as intended, and further illustrations of principals for choosing a relevant collection of outcomes. Auxiliary Figures and Tables are displayed in Appendix A. Further details and extensions are given in

Appendix B. Proofs for all results stated in the main text are given in Appendix C; Proofs for supporting Lemmas are given in Appendix D. Details concerning our empirical applications are given in Appendix E.

1.1 Related Literature

We contribute to an extensive econometric literature concerning quantification of uncertainty in data that exhibit cross-sectional dependence. Much of this research focuses on methods that partition units into clusters according to an observable characteristic, allowing for correlation among observations within the same cluster. Liang and Zeger (1986) introduce the first heteroskedasticity-consistent clustered standard error (see also MacKinnon and White 1985, Bell and McCaffrey 2002, and Imbens and Kolesar 2016 for alternatives that exhibit better finite-sample performance). Bertrand et al. (2004) provide a compelling demonstration of the practical importance of these methods in panel data. Cameron and Miller (2015) give a comprehensive review. Like this approach, we assume that only a small proportion of pairs of units correlate.[4] By contrast, we do not assume that these pairs are know *ex ante*, but rather, can be learned from observations of auxiliary outcomes.

A smaller, but widely applied, subset of this literature develops methods premised on the assumption that dependence can be modeled as a function of the distance between units. This line of work was initiated by Conley (1999), who emphasizes that dependence should be modeled by general measures of "economic distance" (see also Kelejian and Prucha 2007, Kim and Sun 2011, Bester et al. 2016, Pollmann (2020), Colella et al. (2023) and Müller and Watson 2022, 2023 for alternative estimators based on this program). Despite this, most applications of this framework parameterize dependence in terms of geographic proximity alone.

Separately, Barrios et al. (2012), Müller and Watson (2024), and Conley and Kelly (2025) demonstrate that applying standard variance estimators to spatially correlated data often yields spurious findings. The simulation experiments conducted in Section 2 generate results that are consistent with this view. Despite this, our outlook is pragmatic. The core of our proposal is that, due to the widespread availability of additional outcome data, at the very least, some of the cross-sectional correlation that is common across measured outcomes can be controlled.

Our approach for identifying pairs of locations that correlate draws on an extensive literature in multiple hypothesis testing, particularly as developed for DNA micro-array analyses. See Efron (2012) for a textbook treatment. As our primary aim is to construct a single standard error, the criteria that we use to distinguish nulls from non-nulls differs from the methods developed in that literature (see also Bailey et al. (2016) and Bailey et al. (2019) for connections between multiple hypothesis testing and the construction of spatial standard errors). Moreover, we emphasize simple methods that require minimal nonparametric modeling. More sophisticated methods for estimating null distributions, as well as optimality theory, are discussed in Langaas et al. (2005), Efron (2007), Jin and Cai (2007), and Cai and Jin (2010), among other places.

[4]Available methods for constructing clustered standard errors with large clusters are either conservative (Ibragimov and Müller, 2010, 2016) or rely an assumption that error covariance matrices are identical across clusters (Canay et al., 2017, 2021).

Finally, we contribute to a growing literature that considers how standard statistical procedures can be improved by incorporating measurements of multiple outcomes. For example, Ludwig et al. (2019), Sun et al. (2023), and Abadie et al. (2024) each consider how collections of outcomes can be used to improve estimates of causal effects. Our primary focus, by contrast, is on obtaining more accurate estimates of uncertainty in regression problems.

2. Standard Adjustments for Spatial Dependence are Biased

In this section, we use a simulation experiment to measure the performance of several methods for adjusting standard errors for spatial dependence. We consider settings with the following structure. A researcher observes the data (Y_i, W_i, X_i) for each unit i in $1, \ldots, n$, where Y_i is a scalar outcome, W_i is a scalar treatment, and X_i is a vector of covariates. For the sake of concreteness, Y_i could measure the change in the income per capita in U.S. county i over a given time period, W_i could denote the change in the percent of the population in county i that has completed a college degree, and X_i could collect a set of state fixed effects. The researcher assumes that the data arise from the linear model

$$Y_i = \alpha + \tau W_i + \theta^\top X_i + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | W, X] = 0, \quad i = 1, \ldots, n\,, \tag{2.1}$$

where $W = (W_i)_{i=1}^n$ and $X = (X_i)_{i=1}^n$ collect the treatment and covariate measurements over the full sample of units. Let $\hat{\tau}$ denote the ordinary least squares estimate of the parameter τ. Our interest is in estimating an appropriate standard error for $\hat{\tau}$.

Often, it is unreasonable to assume that the components of the residual vector $\varepsilon = (\varepsilon_i)_{i=1}^n$ are uncorrelated. For most settings of interest in applied economics, factors that determine outcomes in San Francisco, say, plausibly also determine outcomes in Berkeley. There are two common ways of adjusting standard errors for spatial dependence. One set of methods is based on clustering standard errors based on a partition of the space, thereby imposing geographic borders on the scope of spatial dependence (Moulton, 1986, 1990). A second set of methods parameterize spatial dependence smoothly in terms of geographic distance (Conley, 1999; Müller and Watson, 2022, 2023).

How well do these methods address the types of spatial dependence exhibited by the data considered in applied economics? An answer to this question requires the specification of a "typical" correlation structure. In Section 2.1, we measure the main features of the spatial dependence exhibited by the types of U.S. county-level outcomes considered in applied economics.[5] In Section 2.2, we use these measurements to calibrate a simulation and document the performance of several widely applied approaches to adjusting standard errors for spatial dependence.

[5]We focus our discussion on U.S. counties because this is the most common level unit of aggregation considered in leading general interest economics journals in 2023. In particular, 12% of all observational papers in this sample take U.S. counties as their unit of observation. This proportion increases to 25% if the sample is restricted to papers whose unit of observation has the potential for spatial correlation. Further details are given in Section 5.

2.1 Measuring The Structure of Spatial Dependence

What types of spatial dependence should we expect in the economic outcomes of U.S. counties? The basic premise of this paper is that non-trivial correlation structures can be inferred from measurements of multiple outcomes. To operationalize this idea, we collect 91 U.S. county-level outcomes. This is the same set of outcomes considered in Section 1. These outcomes, which include measurements of income, poverty, employment, crime, and health, and are intended to capture the main features of the types of outcomes considered in applied economics. We list these outcomes and give further details concerning their construction in Appendix B.1.

Given a "representative" outcome Y_i and treatment W_i, and a pair of counties i and i', should we expect the residuals ε_i and $\varepsilon_{i'}$ to correlate? We argue that an answer to this question can be recovered by averaging the product of the empirical versions of these residuals across the outcomes in our sample. Formally, let $Y_i^{(1)}, \ldots, Y_i^{(d)}$ denote the d outcomes measured for county i. As a working example, we take the treatment of interest W_i to be the change in the percentage of the population in county i that has completed a college degree from 1980 to 2009. For each outcome $Y_i^{(j)}$, let $\hat{\varepsilon}_i^{(j)}$ be the empirical residuals from the least squares regression of $Y_i^{(j)}$ on W_i and a vector of state fixed effects. Let $\tilde{\varepsilon}_i^{(j)}$ denote a version of $\hat{\varepsilon}_i^{(j)}$ that has been standardized to have variance one across units, given by

$$\tilde{\varepsilon}_i^{(j)} = \hat{\varepsilon}_i^{(j)} \Big/ \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\hat{\varepsilon}_i^{(j)})^2}\,. \tag{2.2}$$

The empirical covariance and correlation between the empirical residuals for counties i and i' *across outcomes* are given by

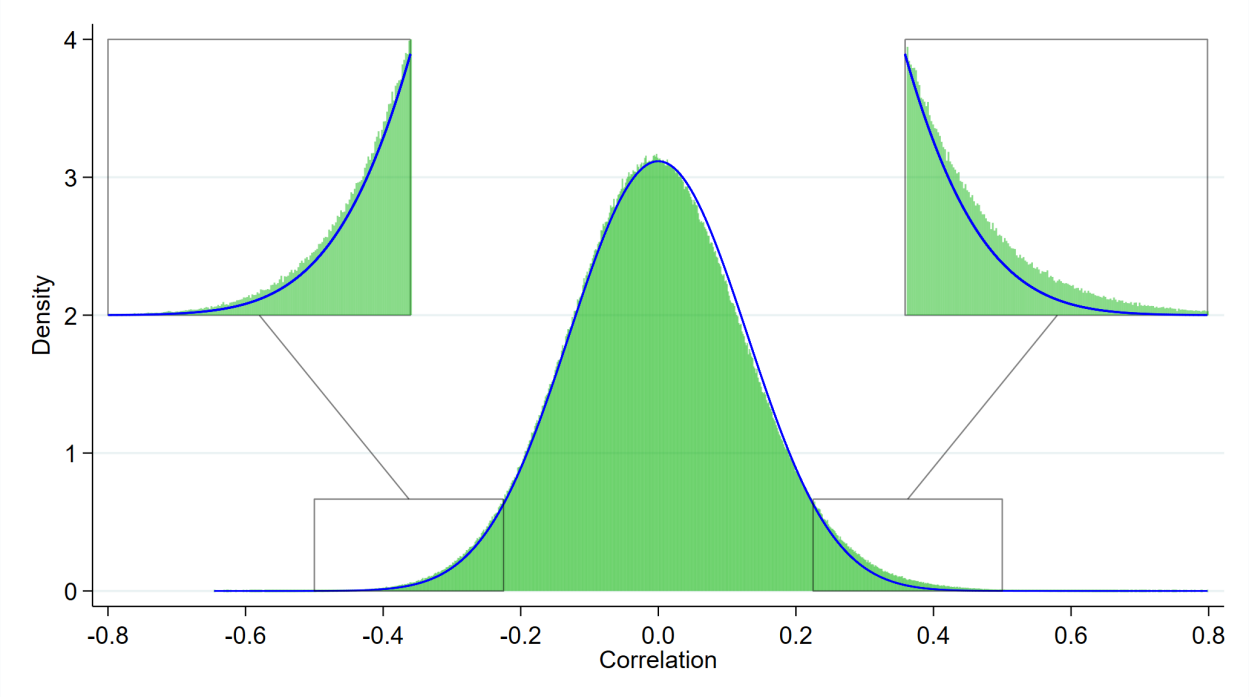
$$\hat{\lambda}_{i,i'} = \frac{1}{d}\sum_{j=1}^{d} (\tilde{\varepsilon}_i^{(j)} - \overline{\varepsilon}_i)(\tilde{\varepsilon}_{i'}^{(j)} - \overline{\varepsilon}_{i'}) \quad \text{and} \quad \hat{\rho}_{i,i'} = \frac{\hat{\lambda}_{i,i'}}{\sqrt{\hat{\lambda}_{i,i}\hat{\lambda}_{i',i'}}}\,, \tag{2.3}$$

respectively, where $\overline{\varepsilon}_i = \frac{1}{d}\sum_{j=1}^{d} \tilde{\varepsilon}_i^{(j)}$ is the average of the normalized residuals $\tilde{\varepsilon}_i^{(j)}$ across outcomes for county i. Intuitively, the statistic $\hat{\rho}_{i,i'}$ captures the tendency for residuals associated with counties i and i' to be correlated.

Figure 2 displays a histogram of the correlation estimates $\hat{\rho}_{i,i'}$ over each pair of distinct U.S. counties. A Gaussian density function, centered at zero, is overlaid in blue. The scaling of this density function has been chosen to best fit the quartiles of the distribution of the measurements $\hat{\rho}_{i,i'}$. The left and right tails of the distribution are displayed as enlarged insets. Two features stand out: The Gaussian distribution does an exceptional job at approximating the center, but under-estimates the right-tail, of the empirical distribution of the correlation estimates.

These observations can be rationalized as follows. Suppose that the correlation estimates $\hat{\rho}_{i,i'}$ are approximately Gaussian. This will hold if the number of outcomes d is sufficiently large and there is not too much correlation across residuals. (We will have more to say about this in Section 4.) The excellent approximation

FIGURE 2. The Distribution of Correlations Between Pairs of U.S. Counties



Notes: Figure 2 displays a histogram of the estimates $\hat{\rho}_{i,i'}$ over each pair of distinct U.S. counties. A mean-zero, Gaussian density function is overlaid in blue. The scaling of this density function has been chosen to best fit the corresponding quartiles of the distribution of the estimates $\hat{\rho}_{i,i'}$. The left and right tails of the distribution are displayed as enlarged insets.

---

of a mean-zero, Gaussian distribution to the central quantiles of the correlation estimates suggests that most pairs of counties are not systematically correlated. By contrast, the poor approximation of the Gaussian distribution to the right tail of the correlation estimates suggests that there is a relatively small proportion of pairs of counties that are systematically positively correlated.

It is worth scrutinizing, however, whether the right-tail of the distribution of correlation estimates captures pairs of counties that are genuinely correlated. Consider the set of county pairs (i, i') in which the correlation estimate $\hat{\rho}_{i,i'}$ exceeds the positive threshold δ. Table 1 reports, for several values of δ, the proportion of county pairs (i, i') in this set in which county i is among the closest 10% of counties to i' on a collection of observable characteristics. Proximity on observable characteristics is very predictive of having highly correlated residuals. Geographic proximity does not tell the whole story, however. Demographic and socioeconomic proximity, e.g., income, population, and population density, are similarly predictive of correlations in residuals.

2.2 Calibrated Simulation

The estimates displayed in Figure 2 suggest that the outcomes of at least some pairs of U.S. counties are systematically correlated. The next question is whether this matters. We assess, through a calibrated simulation, whether ignoring these correlations leads to biased standard error estimates. Specifically, we use the estimates $\hat{\rho}_{i,i'}$ to parameterize a covariance matrix Σ for the residual vector ε in the specification (2.1).

TABLE 1. Predictors of Highly-Correlated County Pairs

Among county pairs (i, i') with $\hat{\rho}_{i,i'} > \delta$, percentage in which i ranks among the 10% closest to i' in:	δ		
	0	0.2	0.4
Population	13	21	41
Urban %	14	20	41
Geographic Distance	12	17	32
Median income	13	17	31
Non-white %	12	15	24
Vote-share	12	14	22
Any of the above	51	62	83

Notes: Table 1 shows the percent of county-pairs (i, i') in which county i is within the top 10% of counties closest to i' on a collection of observable characteristics, among the county-pairs that have a correlation greater than a threshold δ. The threshold δ increases across columns. The percentages of county-pairs with estimated correlations above 0, 0.2, and 0.4 are 49.3%, 6.4%, and 0.3%, respectively.

We assume that the residual vector ε has a Gaussian distribution with mean zero and variance Σ. Let $\sigma_{i,i'}$ denote the i, i'th component of Σ. Our goal is to assess the performance of standard methods that ignore spatial dependencies, as well as the performance of conventional methods for adjusting standard errors that are based on modeling spatial correlation in terms of geographic distance.

To do this, we must meet two objectives. First, the non-zero off-diagonal elements of Σ should primarily consist of pairs of counties whose correlation estimate $\hat{\rho}_{i,i'}$ is likely to reflect real dependence. Second, the covariance matrix Σ must be a proper, positive-definite, covariance matrix, so that it can be used to draw simulation replicates. We develop a simple approach that achieves these goals. We set $\sigma_{i,i'} = 0$ for all pairs of counties with $|\hat{\rho}_{i,i'}| < 0.45$. We then implement an algorithm, described in Appendix B.2, to produce a positive-definite matrix by forming "clusters" among the remaining pairs of counties that have $|\hat{\rho}_{i,i'}| \geq 0.45$. Within these clusters, which contain 39% of these highly correlated pairs, we set $\sigma_{i,i'} = \hat{\rho}_{i,i'}$. The simulation is structured as follows. For each of 1,000 simulation replicates, we draw a random vector ε from $\mathsf{N}(0, \Sigma)$, set $Y = \beta W + \varepsilon$, where $\beta = 0$ and W is a specified treatment vector, and apply various methods to (i) estimate the standard error associated with the least squares regression of Y on W and (ii) test the null hypothesis that $\beta = 0$ at level $\alpha = 0.05$.

Table 2 displays the results of this exercise. We consider two choices of the treatment vector W: The change in the percentage of the population in county i that has completed a college degree from 1980 to 2009 and the change in the per-acre value of farm-land. The first row displays the bias for heteroskedasticity-consistent standard errors (HC1). Across the 1,000 simulation draws, the median estimated standard error is only 0.44 of the true standard error, leading to a rejection rate of 0.39. Conventional methods to address

TABLE 2. Comparison of Methods for Adjusting for Spatial Dependence

	Δ % College Grad		Δ Per-acre Farm Value	
	Mean $\left(\frac{\text{Est. SE}}{\text{True SE}}\right)$	Rej. rate	Mean $\left(\frac{\text{Est. SE}}{\text{True SE}}\right)$	Rej. rate
Method	(1)	(2)	(3)	(4)
(1) HC1	0.44	0.39	0.54	0.29
(2) Cluster by state	0.47	0.36	0.57	0.25
(3) Conley (150mi)	0.47	0.37	0.57	0.26
(4) SCPC	0.64	0.24	0.71	0.17
(5) TMO	0.77	0.14	0.76	0.12

Notes: Table 2 reports the performance of various methods in the calibrated simulation experiment outlined in Section 2.2. Columns (1) and (3) report the mean standard error estimate, over simulation replications, as a proportion of the true standard error. Columns (2) and (4) reports the rejection rate at level 0.05 of the associated test that $\beta = 0$. Results are from 1,000 simulation draws. The treatments of interest are the change in the percent of college graduates from 1980-2009 (Columns 1-2) and the change in the per-acre value of farmland (Columns 3-4). "Robust" refers to HC1 heteroskedasticity consistent standard errors (Hinkley, 1977). "Conley (150mi)" refers to the standard error estimator proposed by Conley (1999), implemented with a bandwidth of 150 miles. "SCPC" refers to the Spatial Correlation Principal Components method proposed by Müller and Watson (2022, 2023). "TMO" refers to the Thresholding Multiple Outcomes estimator proposed in Section 3.

spatial correlation, for example clustering by state (row 2) or using Conley (1999) standard errors with a 150-mile radius (row 3), shows little improvement, with the median standard error at still less than half of the true standard error.[6] These adjustments have no impact on the bias in the standard error estimates (This can be attributed to the fact that Σ is based on the covariance of residuals that have already been been demeaned by state). The fourth row displays the performance of the Spatial Correlation Principal Components (SCPC) method of Müller and Watson (2022, 2023), which is based on parameterizing spatial correlation in terms of geographic distance. The results here exhibit a sizable improvement over the previous methods. The bias of the median estimated standard error is reduced to around 40% of the the true standard error and the rejection rate falls to 0.24.

3. USING MULTIPLE OUTCOMES TO ADJUST STANDARD ERRORS

The results reported in Table 2 are concerning. In a setting calibrated to approximate the spatial dependence exhibited by outcomes frequently studied in applied economics, we find that standard methods can be severely biased. In this section, we propose an alternative method. Our method takes as input a collection of outcomes, of the sort constructed in Section 2.1. This collection is used to recover an estimate of spatial dependence. This estimate is used to adjust standard errors.

[6] Conley (1999) standard errors are consistent only if the number of location pairs within the bandwidth is small relative to the sample size. In our setting, a bandwidth of 150 miles includes 3.1% of county pairs. Setting the bandwidth to 300 miles, say, includes 11.3% of county pairs, moving beyond the regime where we should expect consistency to hold. The methods proposed in Bester et al. (2016) and Müller and Watson (2022, 2023), by contrast, control size under various restrictions in strongly correlated data.

As before, we consider the linear model

$$Y_i^{(0)} = \alpha + \tau^{(0)} W_i + \varepsilon_i^{(0)}, \quad i = 1, \ldots, n \,, \tag{3.1}$$

where $Y_i^{(0)}$ measures some outcome and W_i denotes some treatment of interest. Extensions to settings with covariates, instruments, and panel data are treated in Appendix B.3. Again, the statistic $\hat{\tau}^{(0)}$ denotes the ordinary least squares estimate of $\tau^{(0)}$ in the specification (3.1) and we are interested in estimating an appropriate standard error for $\hat{\tau}^{(0)}$. For the purposes of this paper, we interpret this problem as desiring an estimate of the conditional variance

$$V(\Sigma^{(0)}) = \mathrm{Var}(\hat{\tau}^{(0)} \mid W) = S_n^{-2} \left(W^\top \Sigma^{(0)} W \right) \,, \tag{3.2}$$

where $W = (W_i)_{i=1}^n$ collects the observations W_i, $\Sigma^{(0)} = \mathrm{Var}(\varepsilon^{(0)} \mid W)$, and $S_n = W^\top W$. As a consequence, estimation of the conditional variance (3.2) reduces to estimation of the error covariance matrix

$$\Sigma^{(0)} = (\sigma_{i,i'}^{(0)})_{i,i'=1}^n, \quad \text{where} \quad \sigma_{i,i'}^{(0)} = \mathrm{Cov}(\varepsilon_i^{(0)}, \varepsilon_{i'}^{(0)} \mid W) \,. \tag{3.3}$$

This conditional perspective is shared by, e.g., Müller and Watson (2022, 2023). Design based settings, where interest is in estimating the randomness in the estimator generated by the treatment W, are worth further consideration (Abadie et al., 2020, 2023).

3.1 Multiple Outcomes and Proportionality

In absence of further assumptions on its structure, the residual covariance matrix $\Sigma^{(0)}$ is not recoverable. As a consequence, in practice, researchers impose a variety of restrictions on the structure of cross-sectional correlation. For example, in many settings, it is common to assume that all of the off-diagonal elements of the matrix $\Sigma^{(0)}$ are equal to zero (White, 1980). In cases where this is less credible, researchers often assume that units can be partitioned into clusters on the basis of some shared characteristic (Moulton, 1986, 1990; Liang and Zeger, 1986), or that cross-sectional correlation can be parameterized smoothly in terms of geographic proximity (Conley, 1999; Müller and Watson, 2022, 2023).

We take an alternative route. Our approach is premised on the observation that, often, in a single study, researchers observe many outcome variables for the same set of units. For example, a researcher, interested in the correlation across U.S. counties between average income and the proportion of the population with a college education, could also feasibly observe all of the outcomes considered in Section 2. In Section 5, we give several concrete examples from empirical economics with this characteristic. Formally, for each unit i, in addition to the outcome of interest $Y_i^{(0)}$, we assume that we also observe a collection $Y_i^{(1)}, \ldots, Y_i^{(d)}$ of d auxiliary, post-treatment outcomes. Let $\varepsilon_i^{(j)}$ and $\hat{\varepsilon}_i^{(j)}$ denote the population and empirical residuals associated with unit i when outcome $Y_i^{(j)}$ replaces $Y_i^{(0)}$ in the linear model (3.1), respectively.

The basic idea is to use the auxiliary outcomes to estimate the residual correlation matrix associated with the outcome of interest. To rationalize this approach, we must impose some conditions. First, we impose a

restriction that implies that the residual *correlation* matrix is shared across outcomes. We call a collection of outcomes with this property *proportional*.

Assumption 3.1 (Proportionality)**.** *There exist matrices* $\Lambda_n = (\lambda_{i,i'})_{i,i'=1}^n$ *and* $\Gamma_d = (\gamma_{j,j'})_{j,j'=0}^d$ *such that*

$$\sigma^{(0)}_{j,j'} = \mathrm{Cov}(\varepsilon_i^{(j)}, \varepsilon_{i'}^{(j')} \mid W) = \gamma_{i,i'}\lambda_{j,j'} \,, \tag{3.4}$$

for each pair of units i, i' *and outcomes* j, j'*, respectively. By convention, we set* $\gamma_{0,0} = 1$*.*

The restriction (3.4) means that the covariance between the residuals associated with two units and a pair of, potentially distinct, outcomes depends only on the outcomes through a constant factor.[7]

As a consequence, the empirical residuals for the auxiliary outcomes can be used to estimate the residual covariance matrix $\Sigma^{(0)}$. In particular, let $\tilde{\varepsilon}_i^{(j)}$ denote a version of the empirical residual $\hat{\varepsilon}_i^{(j)}$ that has been normalized to have variance one across units. That is, the normalized residual $\tilde{\varepsilon}_i^{(j)}$ is given by

$$\tilde{\varepsilon}_i^{(j)} = \hat{\varepsilon}_i^{(j)} \hat{\gamma}_j^{-1/2} \,, \quad \text{where} \quad \hat{\gamma}_j = \frac{1}{n}\sum_{i=1}^n (\hat{\varepsilon}_i^{(j)})^2 \tag{3.5}$$

denotes the sample variance of the residuals for the jth outcome. The covariance between these, normalized, residuals for units i and i', across outcomes, is given by

$$\hat{\lambda}_{i,i'} = \frac{1}{d}\sum_{j=1}^d (\tilde{\varepsilon}_i^{(j)} - \overline{\varepsilon}_i)(\tilde{\varepsilon}_{i'}^{(j)} - \overline{\varepsilon}_{i'}) \,, \quad \text{where} \quad \overline{\varepsilon}_i = \frac{1}{d}\sum_{j=1}^d \tilde{\varepsilon}_i^{(j)} \tag{3.6}$$

denotes the average normalized residual for ith unit. The associated correlation is given by

$$\hat{\rho}_{i,i'} = \frac{\hat{\lambda}_{i,i'}}{\sqrt{\hat{\lambda}_{i,i}\hat{\lambda}_{i',i'}}} \,. \tag{3.7}$$

The statistics $\hat{\rho}_{i,i'}$ are exactly the correlation estimates considered in Section 2. Under Assumption 3.1, the correlation estimates $\hat{\rho}_{i,i'}$ recover the residual covariances $\sigma^{(0)}_{j,j'}$ up to a scale factor.

Proportionality is a highly stylized condition and should be viewed as only ever holding approximately for any given collection of outcomes. The same holds, of course, for assumptions that decompose spatial correlation into clusters or parameterize spatial correlation in terms of geography. From a practical perspective, by imposing Assumption 3.1, we appeal to the intuition that the cross-sectional correlation exhibited by the auxiliary outcomes captures the main features of the cross-sectional correlation of the outcome of interest. On the other hand, we are ignoring aspects of spatial correlation that might be particular to any given outcome. Moreover, we are assuming that these particularities are idiosyncratic and, when considered in aggregate, negligible. Assumptions equivalent to proportionality have been considered in other fields of applied statistics. In particular, a literature in the analysis of micro-array experiments, initiated by Efron (2010), develops

[7] Assumption 3.1 is testable. Guggenberger et al. (2023) propose a computationally efficient test and illustrate its application to a setting related to identification robust inference for linear instrumental variables regression.

procedures for multiple hypothesis testing in proportional data. See, e.g., Muralidharan (2010), Allen and Tibshirani (2012), and Hoff (2016) for further discussion.[8]

Proportionality, however, is not enough. In fact, in the absence of further restrictions, any estimator of the conditional variance (3.2) can still be made arbitrarily inaccurate. We formalize this statement as follows.

Theorem 3.1. *Let $\breve{V}(Y, W)$ be any measurable, locally bounded function of the outcomes $Y = (Y^{(j)})_{j=0}^{d}$ and treatments W. For any constants $0 < \eta < 1$ and $0 < K$ and W with $W^\top W \neq 0$, there exists a conditional distribution for Y, that satisfies Assumption 3.1, such that either*

$$P\left\{\breve{V}(Y,W) - V(\Sigma^{(0)}) \geq K \mid W\right\} \geq 1-\eta \quad \text{or} \quad P\left\{V(\Sigma^{(0)}) - \breve{V}(Y,W) \geq K \mid W\right\} \geq 1-\eta \tag{3.8}$$

holds.

This result can be understood with a simple example. If we observe a single realization of a random variable X, then there is no way to estimate its variance. Translated to our context, without further conditions, there is no way to recover changes in residual covariance matrix $\Sigma^{(0)}$ that are proportional to $WW^\top$. Thus, some further restriction must be made for the standard error of the least squares estimator $\hat{\tau}^{(0)}$ to be estimable.

3.2 Sparsity and Thresholding

Our approach is based on the hypothesis that, in many settings of practical interest in applied economics, the residuals for *most* pairs of units are *not* meaningfully correlated. That is, pairs of units can be effectively categorized as "nulls" and "non-nulls." The components of the residual covariance matrix $\Sigma^{(0)}$ that correspond to the null pairs are equal to zero, the components that correspond to non-nulls can be large, and the nulls outnumber the non-nulls. In other words, the covariance matrix $\Sigma^{(0)}$ is sparse. This is the same idea that underlies clustering. There, pairs of units in the same cluster are known to be non-null *a priori*. Here, we use the auxiliary outcomes to *determine* the pairs of units that are non-nulls.

There are many ways that one might make such a determination. We propose to choose the pairs of units whose empirical residual correlation $\hat{\rho}_{i,i'}$ is large in absolute value.[9] There are two reasons for this. First, correlation coefficients are self-normalized. That is, the covariances $\hat{\lambda}_{i,i'}$ and $\hat{\lambda}_{i,i''}$ could be very different simply because the residuals for unit i are are much noisier than the residuals for unit i''. The correlations $\hat{\rho}_{i,i'}$ and $\hat{\rho}_{i,i''}$, by contrast, are comparable. Second, if the pair of units i, i' is a null, *i.e.,* the associated residuals are not correlated, then the distribution of the statistic $\hat{\rho}_{i,i'}$ can be characterized. Knowledge of the null distribution, then, allows us to discriminate between nulls and non-nulls. We treat this point in detail in Sections 3.3 and 3.4.

[8] More generally, Dawid (1981) gives a classical discussion of applications of models satisfying Assumption 3.1 to Bayesian inference. Gupta and Nagar (1999) gives a textbook treatment of related distributions. See also Zhou (2014) and Hornstein et al. (2018) for more recent analyses of related settings.

[9] An alternative approach, not pursued in this paper, is to use the estimates $\hat{\rho}_{i,i'}$ to learn clusters of units. This is worth further consideration, although it has the disadvantage that, necessarily, some geographically proximate units are determined to not correlate. Cao et al. (2024) pursue a related idea using a single outcome.

Before turning to this, however, we spell out how we can use the identified non-nulls to adjust standard errors. Suppose that we were willing to determine that all pairs of units whose absolute residual correlation is greater than some threshold δ are non-nulls. That is, we say that the pair of units i, i' is a non-null if $|\hat{\rho}_{i,i'}| \geq \delta$. How should this inform how we construct standard errors for the regression problem (3.1)? One sensible estimator of the covariance $\sigma^{(0)}_{i,i'}$ is given by

$$\hat{\sigma}^{(0)}_{i,i'}(\delta) = \begin{cases} \hat{\varepsilon}^{(0)}_i \hat{\varepsilon}^{(0)}_i \,, & i = i', \\ \hat{\varepsilon}^{(0)}_i \hat{\varepsilon}^{(0)}_{i'} \, \mathbb{I}\{|\hat{\rho}_{i,i'}| \geq \delta\} \,, & i \neq i' \,. \end{cases} \tag{3.9}$$

Recall that the outcomes $Y^{(0)}_i$ are not used to construct the correlations $\hat{\rho}_{i,i'}$. Collect these estimates into the matrix $\widehat{\Sigma}^{(0)}(\delta)$ and plug this matrix into the expression (3.2), giving the estimator

$$\widehat{V}(\delta) = S_n^{-2} \left(W^\top \widehat{\Sigma}^{(0)}(\delta) W \right) \,, \tag{3.10}$$

where we recall that $S_n = W^\top W$.

Written differently, the components of the covariance matrix $\widehat{\Sigma}^{(0)}$ for the null pairs are set to zero; the components for the non-null pairs are determined entirely by the empirical residuals associated with the outcome of interest. This is analogous to standard heteroskedasticity or cluster robust standard errors which use term like $\hat{\varepsilon}^{(0)}_i \hat{\varepsilon}^{(0)}_{i'}$ to estimate non-zero elements of residual covariance matrices (White, 1980; Liang and Zeger, 1986). The only difference is that, here, the auxiliary outcomes are used to determine which elements of the residual covariance matrix will be non-zero. As a consequence, the estimator (3.10) is minimally sensitive to Assumption 3.1—its value is affected by the auxiliary outcomes only through the choice of which covariances to zero out. We refer to the estimator (3.10) as the "TMO" variance estimator, for "Thresholding Multiple Outcomes."

In the formulation (3.9), the TMO estimator (3.10) does not threshold diagonal elements of the residual covariance matrix. That is, in effect, we augment the standard White (1980) heteroskedasticity consistent variance estimate. The same idea can be applied to wrap our approach around other existing variance estimators, particularly those that explicitly model cross-section dependence in terms of distance. For instance, an analogous estimator could be constructed, where elements that correspond to pairs of units in the same cluster, e.g., state, are never thresholded to zero. The spatial standard errors proposed by Conley (1999) and Müller and Watson (2022, 2023) can be augmented analogously. We give a detailed comparison of the practical performance of these alternatives in Section 5.

3.3 Threshold Choice

How should we choose the threshold δ? Many estimators of sparse covariance matrices considered in the literature involve thresholding each element of an initial estimate at critical values analogous to standard corrections for multiple hypothesis testing, such as the Bonferroni procedure (Cai et al., 2010; Cai and Liu, 2011; Bailey et al., 2019). That is, in these formulations, δ is chosen so that the probability of mistakenly

retaining even one null correlation is kept below a fixed level. If the objective is to consistently estimate a growing covariance matrix, then this is reasonable. If too many nulls are never thresholded, the aggregate error in the estimate will never converge.

In this section, we argue that, in fixed samples, thresholds constructed in this way can be too large to be practically useful. That is, rather than targeting consistent estimation of $\Sigma^{(0)}$, per se, we seek to choose a threshold δ that improves the downstream estimate of the regression variance $V(\Sigma^{(0)})$ for a fixed collection of units and outcomes. This task entails balancing the contributions of the null and non-null pairs of units to the error of the resultant estimator. To see this, consider the decomposition

$$\begin{aligned}
\widehat{V}(\delta) - V(\Sigma^{(0)}) &= S_n^{-2} \left(\sum_{i=1}^{n} W_i^2 (\hat{\varepsilon}_i^{(0)} \hat{\varepsilon}_i^{(0)} - \sigma_{i,i}^{(0)}) + \mathsf{T}(\delta) + \mathsf{F}(\delta) \right) , \quad \text{where} \\
\mathsf{T}(\delta) &= \sum_{i,i' \in \mathcal{T}} W_i W_{i'} \left((\hat{\varepsilon}_i^{(0)} \hat{\varepsilon}_{i'}^{(0)} - \sigma_{i,i'}^{(0)}) \mathbb{I}\{|\hat{\rho}_{i,i'}| \geq \delta\} - \sigma_{i,i'}^{(0)} \mathbb{I}\{|\hat{\rho}_{i,i'}| \leq \delta\} \right) \quad \text{and} \\
\mathsf{F}(\delta) &= \sum_{i,i' \in \mathcal{F}} W_i W_{i'} \hat{\varepsilon}_i^{(0)} \hat{\varepsilon}_{i'}^{(0)} \mathbb{I}\{|\hat{\rho}_{i,i'}| \geq \delta\} .
\end{aligned} \tag{3.11}$$

Here, the sets $\mathcal{T}$ and $\mathcal{F}$ collect non-null and null pairs of units, respectively.

Optimization of the expectation, or any higher order moment, of the error (3.11) requires knowledge of several unknown quantities, e.g., the distribution of the statistics $\hat{\rho}_{i,i'}$ and the values of the true residual covariances $\sigma_{i,i}^{(0)}$. To make this problem tractable, we apply several, heuristic, approximations. Taken in aggregate, these approximations will bias our analysis toward smaller values of δ, giving, more often than not, more conservative downstream standard errors. Let p_0 denote the proportion of pairs of units that are nulls. Let $f_0(\cdot)$ denote the density the absolute empirical correlation $|\hat{\rho}_{i,i'}|$ for a randomly selected null pair i, i'. Let $f_1(\cdot)$ denote the analogous density for the non-null pairs. Consider the loss function

$$\begin{aligned}
L(\delta) &= \widetilde{\mathsf{T}}(\delta) + \widetilde{\mathsf{F}}(\delta), \quad \text{where} \\
\widetilde{\mathsf{T}}(\delta) &= (1 - p_0) \int_0^{\delta} t \, \mathrm{d}f_1(t) \quad \text{and} \quad \widetilde{\mathsf{F}}(\delta) = p_0 \int_{\delta}^{\infty} t \, \mathrm{d}f_0(t) .
\end{aligned} \tag{3.12}$$

The function (3.12) abstracts away, or ignores, several features of the error (3.11). First, any heterogeneity across pairs i, i' generated by differences in the term $W_i W_{i'}$ is dropped. Moreover, for the non-nulls, when $|\hat{\rho}_{i,i'}| \geq \delta$, the difference $\hat{\varepsilon}_i^{(0)} \hat{\varepsilon}_{i'}^{(0)} - \sigma_{i,i'}^{(0)}$ is taken, in effect, to be zero.[10] In turn, when $|\hat{\rho}_{i,i'}| \leq \delta$, we ignore the fact that the distributions of $\hat{\rho}_{i,i'}$ and $\sigma_{i,i'}^{(0)}$ are not the same. For the nulls, likewise, we ignore the fact

[10] As the statistics $\hat{\varepsilon}_i^{(0)} \hat{\varepsilon}_{i'}^{(0)}$ and $\hat{\rho}_{i,i'}$ are likely to be positively correlated, on the event $|\hat{\rho}_{i,i'}| \geq \delta$, the statistic $|\hat{\varepsilon}_i^{(0)} \hat{\varepsilon}_{i'}^{(0)}|$ is likely to be be positively biased for $|\sigma_{i,i'}^{(0)}|$. We are ignoring this bias, appealing to the intuition that when $\hat{\rho}_{i,i'}$ is reasonably precisely estimated, its correlation with $\hat{\varepsilon}_i^{(0)} \hat{\varepsilon}_{i'}^{(0)}$ should not be very big, and so the resultant bias will not overwhelm the other contributions to the error.

that the distributions of $\hat{\rho}_{i,i'}$ and $\hat{\varepsilon}_i^{(0)}\hat{\varepsilon}_{i'}^{(0)}$ are not the same.[11] Finally, we drop the signs of the terms $\sigma_{i,i'}^{(0)}$ and $\hat{\varepsilon}_i^{(0)}\hat{\varepsilon}_{i'}^{(0)}$, letting each contribute to the loss only in absolute value.[12]

Nevertheless, (3.12) captures, in a parsimonious and tractable way, many features of the structure encoded in (3.11). Including a non-null reduces the error in proportion to the population covariance $\sigma_{i,i'}^{(0)}$. Including a null increases the error in proportion to the empirical covariance $\hat{\varepsilon}_i^{(0)}\hat{\varepsilon}_{i'}^{(0)}$. Under Assumption 3.1, these objects can be proxied by the correlation $\hat{\rho}_{i,i'}$. The following Theorem characterizes the threshold δ that minimizes the loss (3.12), under a collection of simple conditions.

Theorem 3.2. *Assume that the densities $f_0(\cdot)$ and $f_1(\cdot)$ are continuously differentiable. Moreover, assume that the distribution associated with $f_1(\cdot)$ is stochastically larger than the distribution associated with $f_0(\cdot)$. Define the local false discovery rate*

$$\mathsf{fdr}(\delta) = P\left\{(i,i') \in \mathcal{F} \mid |\hat{\rho}_{i,i'}| = \delta\right\}\ , \tag{3.13}$$

where $\mathcal{F}$ denotes the set of null pairs of units. The loss $L(\delta)$, defined in (3.12), has a unique minimum at the threshold $\delta^\star$ that solves the equality $\mathsf{fdr}(\delta) = 0.5$.

We refer to the point that $\delta^\star$ solves $\mathsf{fdr}(\delta) = 0.5$ as "the point of equalized classification." We aim to set the threshold δ at $\delta^\star$. The intuition underlying Theorem 3.2 is simple. Suppose that we choose a very large value of δ, e.g., 0.9. This far out in the tail, almost all of the pairs with $|\hat{\rho}_{i,i'}| \geq \delta$ will be non-nulls. Most of the non-null pairs, however, won't have correlations this large and so won't have been included in the variance estimate. Consider the decision to incrementally reduce the threshold. We are willing to do this as long as we are letting in more non-nulls than nulls. Iterating this process, we stop at the point where the proportion of new non-nulls and nulls is equalized.

It remains, then, to specify a procedure for estimating $\delta^\star$. For the time being, assume that we know of the distribution function $F_0(\cdot)$ of the empirical correlations $\hat{\rho}_{i,i'}$ for the null pairs of units. We treat the construction of an estimate of this quantity in the following subsection. Let $F_0^+(\cdot) = 2(1 - F_0(\cdot))$ denote the associated right distribution function for the absolute correlations $|\hat{\rho}_{i,i'}|$. Analogously, let $F^+(\delta) = P\{|\hat{\rho}_{i,i'}| \geq \delta\}$ denote the unconditional right distribution function of the absolute correlations. Observe that

$$\begin{aligned} &P\left\{(i,i') \in \mathcal{T} \mid |\hat{\rho}_{i,i'}| \geq \delta\right\} = (F^+(\delta) - p_0 F_0^+(\delta))/F^+(\delta) \quad \text{and} \\ &P\left\{(i,i') \in \mathcal{F} \mid |\hat{\rho}_{i,i'}| \geq \delta\right\} = p_0 F_0^+(\delta)/F^+(\delta) \end{aligned} \tag{3.14}$$

[11] The correlation estimates $\hat{\rho}_{i,i'}$ are more disperse than the true covariances $\sigma_{i,i'}^{(0)}$, but less disperse than the statistics $\hat{\varepsilon}_i^{(0)}\hat{\varepsilon}_{i'}^{(0)}$. Thus, in our approximation, we overestimate the contribution from the non-nulls and underestimate the contribution from the nulls. Both factors make the threshold smaller than would be optimal, and thereby yield, more often than not, a more conservative standard error.

[12] By letting the terms $\hat{\varepsilon}_i^{(0)}\hat{\varepsilon}_{i'}^{(0)}$ contribute only in absolute value, we rule out situations where positive and negative contributions from the nulls cancel each other out, as it would be undesirable to rely on this sort of cancellation in practice.

give the probabilities of being non-null and null, conditional on the absolute correlation being larger than δ, respectively. The difference between these probabilities is proportional to the function

$$Q(\delta) = (F^+(\delta) - p_0 F_0^+(\delta)) - p_0 F_0^+(\delta) = F^+(\delta) - 2p_0 F_0^+(\delta) \; . \tag{3.15}$$

The function $Q(\delta)$ is increasing as δ decreases if and only if more non-nulls are being let in than nulls. Thus, the point of equalized classification can be recovered by maximizing the function $Q(\delta)$.

An an unbiased, but infeasible, estimate of $Q(\delta)$ is given by

$$\tilde{Q}_{n,d}(\delta) = \widehat{F}_{n,d}(\delta) - 2p_0 F_0^+(\delta) \; , \tag{3.16}$$

where

$$\widehat{F}_{n,d}(\delta) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{i'<i} \mathbb{I}\{|\hat{\rho}_{i,i'}| \geq \delta\} \tag{3.17}$$

denotes the empirical right distribution function of the absolute correlations. Constructing a feasible version of the estimator (3.16) requires an estimate of the proportion of nulls p_0. There is a large literature that studies estimators of this quantity (see e.g., Langaas et al. 2005, Efron 2007, Jin and Cai 2007, and Cai and Jin 2010). For the most part, these estimators are quite complicated. Again, we take a simpler approach. In many applications of interest, the proportion of null pairs p_0 is close to one. In this case, the feasible estimator

$$\widehat{Q}_{n,d}(\delta) = \widehat{F}_{n,d}(\delta) - 2F_0^+(\delta) \; , \tag{3.18}$$

will be accurate. Our preferred approach chooses the threshold $\hat{\delta}^\star$ as the maximizer of (3.18).[13]

3.4 Estimating the Null Distribution

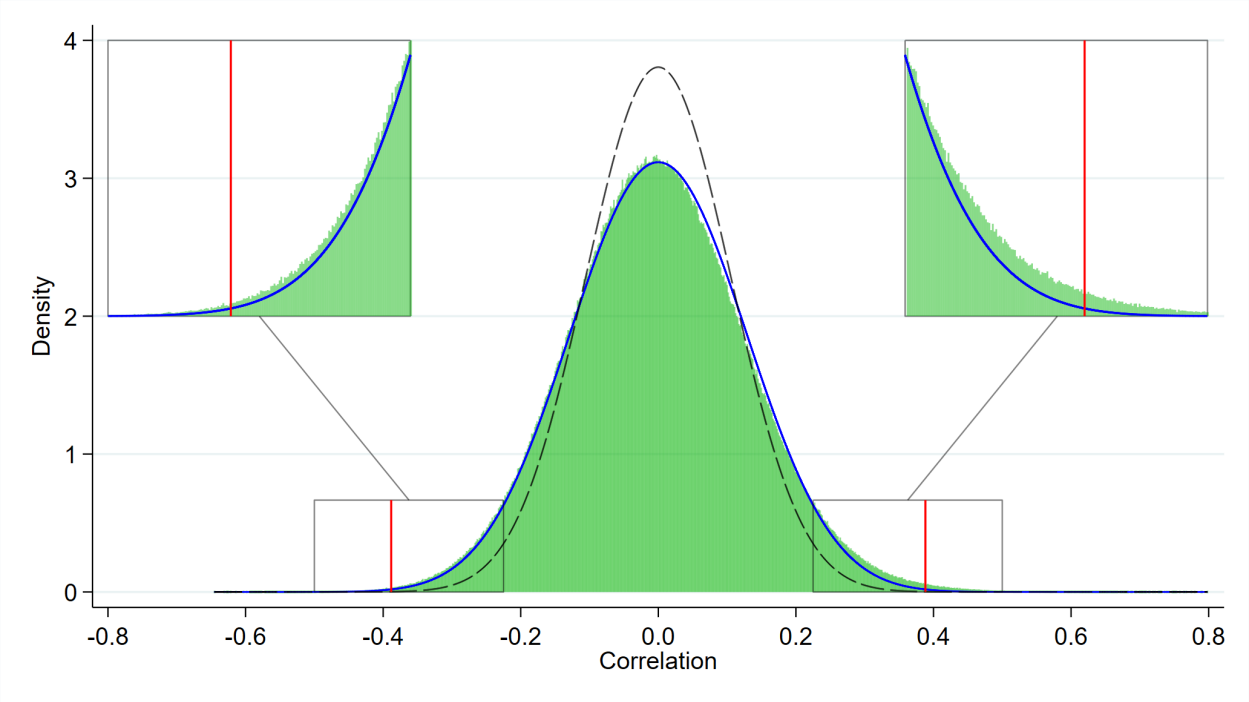
To close the procedure, it remains to specify an estimator of the null distribution $F_0(\cdot)$. If the outcomes $Y_i^{(1)}, ..., Y_i^{(d)}$ were independent, this would be straightforward. In particular, in this case, if the pair i, i' is a null, then the correlation coefficient $\hat{\rho}_{i,i'}$ will be approximately Gaussian with mean zero and variance $1/d$ (see e.g., Example 11.3.6, Lehmann and Romano, 2022). In practice, of course, outcomes are unlikely to be independently distributed. Correlation among the outcomes reduces the effective sample size associated with the estimator $\hat{\rho}_{i,i'}$, increasing its variance. To see this, we return to the collection of outcomes considered in Section 2. Like Figure 2, Panel A of Figure 3 displays a histogram of the correlation estimates $\hat{\rho}_{i,i'}$ across pairs of U.S. counties. A Gaussian density with mean zero and variance $1/d$ is overlaid in black; the dispersion in the empirical correlations is severely underestimated.

Consequently, to estimate a reasonable null distribution, we must account for the correlation across outcomes. One way to do this would involve estimating this correlation directly. We propose a simpler approach, motivated by methods developed in Efron (2007), that works by directly inspecting the empirical distribution of the correlations. In particular, we assume the center of the distribution largely consists of correlations associated with null pairs. Accordingly, we fit the variance of a mean-zero Gaussian distribution

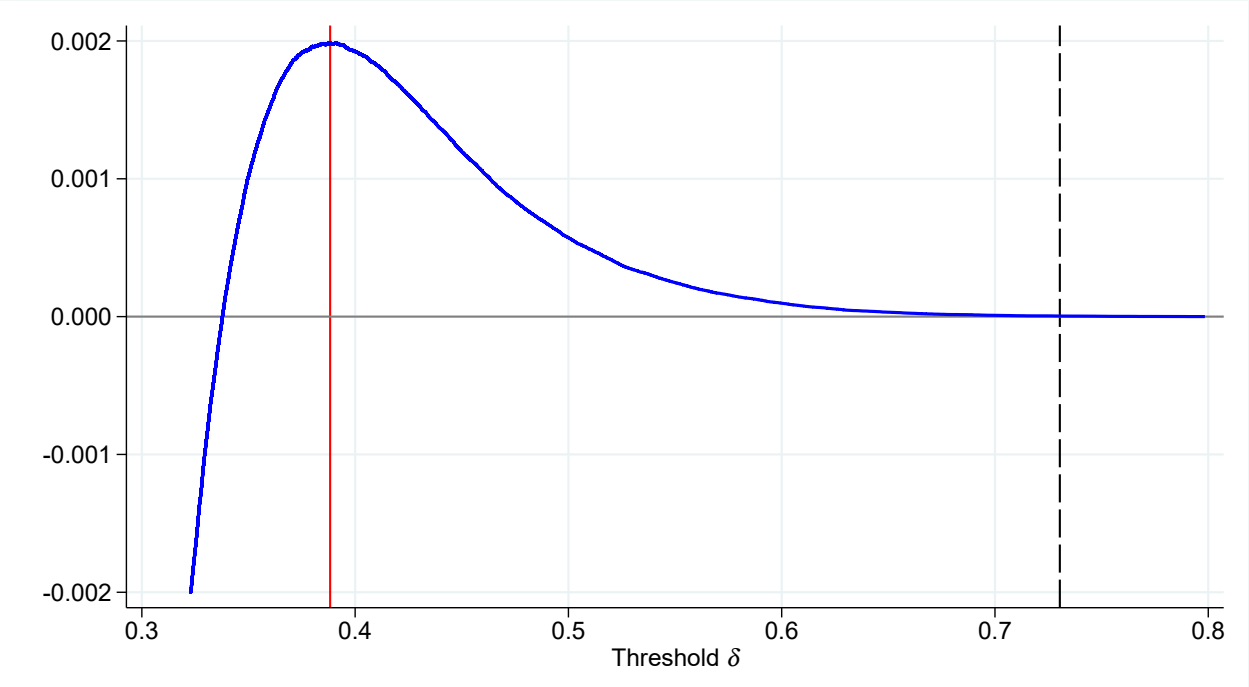
[13] An infeasible threshold chosen by maximizing the function (3.16) would be smaller than $\hat{\delta}^\star$, as $p_0 \leq 1$.

FIGURE 3. Estimating the Null Distribution and Choosing a Threshold

Panel A: Correlation Between Pairs of U.S. Counties, Revisited



Panel B: Approximation to Non-Nulls Minus Nulls



Notes: Panel A of Figure 3 reproduces Figure 2. A mean-zero, Gaussian density function with variance $1/d$ is overlaid in black. A mean-zero, Gaussian density function with variance $\hat{v}$ is overlaid in blue. Panel B displays the estimator estimator $\widehat{Q}_{n,d}(\delta)$, defined in (3.18), over a grid of values of δ, where we have replaced the the null distribution $F_0(\cdot)$ with its Gaussian approximation $\Phi_{\hat{v}}(\cdot)$. The solid vertical red line displays the threshold $\hat{\delta}^\star$, defined as the maximizer of $\widehat{Q}_{n,d}(\delta)$. The dashed vertical black line displays the Bonferroni threshold $\Phi_{\hat{v}}^{-1}(1-0.05/n^2)$.

to the center of the empirical distribution of the correlation estimates. Formally, let $\Phi_v(\cdot)$ denote the cumulative distribution function of a mean-zero Gaussian random variable with variance v and let $\Phi_v^{-1}(\cdot)$ and $\mathsf{IQR}_{v,0} = \Phi_v^{-1}(0.75) - \Phi_v^{-1}(0.25)$ denote the associated quantile function and interquartile range, respectively. In parallel, let

$$G_{n,d}(\delta) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{i'<i} \mathbb{I}\{\hat{\rho}_{i,i'} < \delta\} \tag{3.19}$$

denote the empirical distribution of the correlation estimates and let $G_{n,d}^{-1}(\cdot)$ and $\widehat{\mathsf{IQR}} = G_{n,d}^{-1}(0.75) - G_{n,d}^{-1}(0.25)$ denote the associated quantile function and interquartile range. We let $\hat{v}$ denote the value that equates the Gaussian and empirical interquartile ranges, *i.e.,* $\mathsf{IQR}_{\hat{v},0} = \widehat{\mathsf{IQR}}$. We refer to the estimate $\widehat{\mathsf{df}} = 1/\hat{v}$ as the effective sample size or degrees of freedom.

There is an extensive statistical literature that develops more sophisticated approaches for estimating null distributions in large-scale multiple testing problems, with an emphasis on applications to analysis of DNA microarray data (see e.g., Efron 2007, Jin and Cai 2007, and Cai and Jin 2010). Efron (2012) gives an influential textbook treatment. We adopt the approach outlined above, because it is computationally simple, *i.e.,* it does not require any nonparametric density estimation, and performs well in our applications. In Appendix B.4, we assess the sensitivity of our results to alternative null distribution estimators.

The density function associated with a mean-zero Gaussian with variance $\hat{v}$ is overlaid on Panel A of Figure 3 in blue. The fit at the center of the distribution is remarkably accurate. The right-tail of the Gaussian density underestimates the empirical distribution of the correlation estimates. That is, if our estimate of the null distribution is accurate, then there is evidence of a substantial number of non-null pairs of units whose residuals are significantly positively correlated.

Panel B of Figure 3 displays the values of the estimator $\widehat{Q}_{n,d}(\delta)$, defined in (3.18), over a grid of values of δ, where we have replaced the null distribution $F_0(\cdot)$ with its Gaussian approximation $\Phi_{\hat{v}}(\cdot)$. A vertical red line denotes the maximizer of this curve, i.e., $\hat{\delta}^\star$, at roughly 0.4. It is worth highlighting that the threshold $\hat{\delta}^\star$ is much smaller than the threshold that would be chosen with a standard correction for multiple hypothesis testing. The threshold that would be used in an application of a Bonferroni correction, i.e., $\Phi_{\hat{v}}^{-1}(1 - 0.05/n^2) \approx 0.73$, is displayed with a vertical black line.

We conclude by noting that, in small samples, correlation coefficients have a highly skewed distribution away from the null. In other words, correlation coefficients are not pivotal—even in Gaussian data their variances depend on the true underlying correlation. Fisher (1915) shows that correlation coefficients can be made approximately Gaussian through the transformation

$$\tilde{\rho}_{i,i'} = \frac{1}{2} \log\left(\frac{1 + \hat{\rho}_{i,i'}}{1 - \hat{\rho}_{i,i'}}\right) . \tag{3.20}$$

See Hotelling (1953) and Efron (1982) for further discussion.[14] In our applications, we find that estimating a null distribution by using a Gaussian approximation to the Fisher transformed correlations $\tilde{\rho}_{i,i'}$, rather than

[14] Muralidharan (2010) shows that the Fisher transformation maintains it normalizing property in data satisfying Assumption 3.1.

the empirical correlations $\hat{\rho}_{i,i'}$, is more robust. In practice, we recommend using the Fisher transformed correlation in place of the empirical correlation when implementing the construction proposed in this section. We have centered our discussion around correlation coefficients to ease exposition.

3.5 Performance

We now return to the calibrated simulation considered in Section 2. We show that the TMO estimator, implemented with the threshold $\hat{\delta}^{\star}$, exhibits smaller bias and has more accurate rejection rates than standard methods for constructing spatial standard errors. In particular, the fifth row of Table 2 displays estimates of the bias and rejection rate of the TMO estimator when used to construct standard errors for the two regressions considered in Section 2. The biases are substantially smaller, and the associated rejection rates are closer to the nominal Type I error rate, *i.e.,* $\alpha = 0.05$.

It is worth acknowledging, however, that this simulation environment is particularly suited to the TMO estimator, as the underlying spatial correlation structure is determined by the same set of outcomes used by TMO. Moreover, the spatial correlation structure used in the simulation is quite sparse. More diffuse, geographically structured correlation patterns may be better suited for the alternative methods. Despite this, the good performance of the TMO estimator in this environment can be interpreted as demonstrating that it is able to adjust for forms of spatial correlation that are not captured by existing methods.

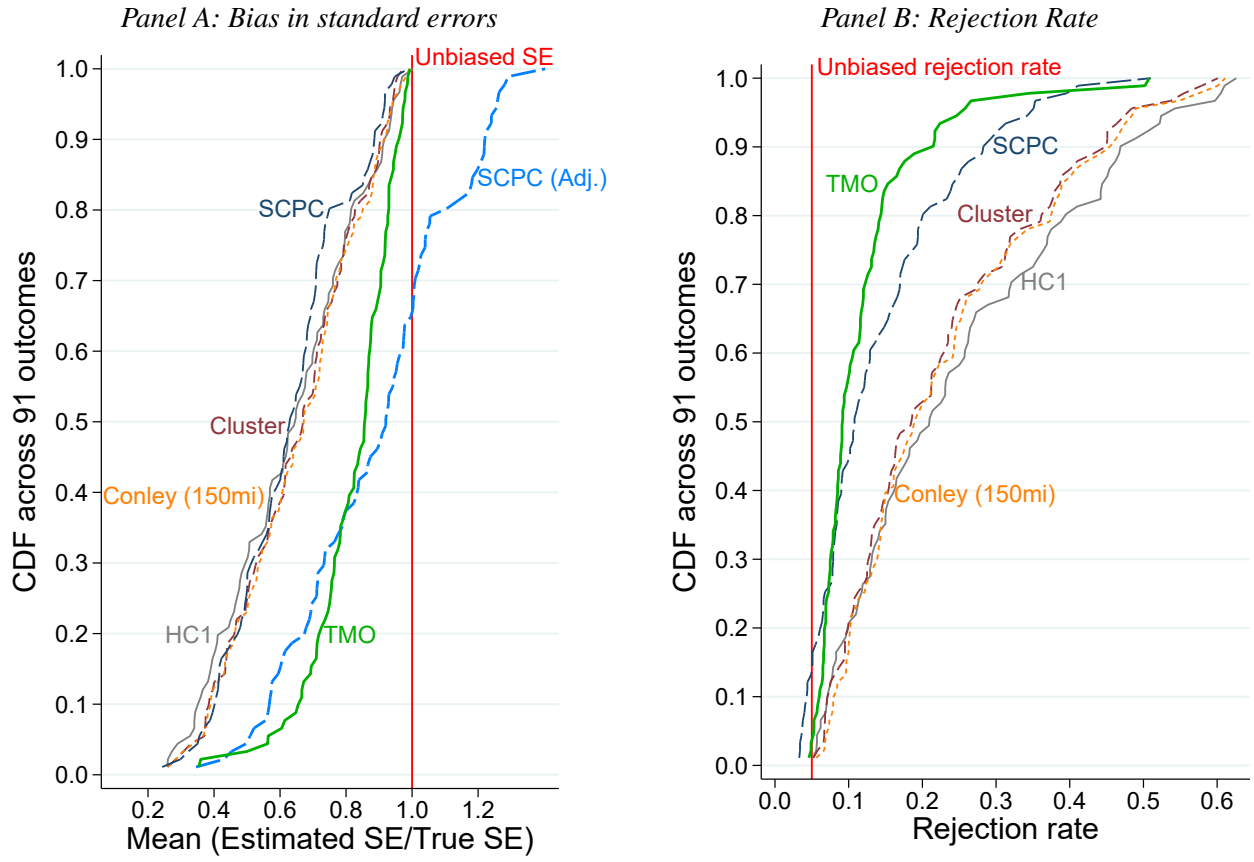
There may be some concern that these results are driven by the two choices of treatments considered in Section 2, *i.e.,* the change in the percentage of the population in county i that has completed a college degree and the change in the per-acre value of farm-land. To address this concern, we replicate the Monte Carlo experiment, setting each of the 91 outcome variables discussed in Section 2 as the treatment variable. Figure 4 displays the distribution functions of the bias and rejection rate associated with each method across these specifications. The TMO standard errors have uniformly better bias, and more accurate rejection rates, than the HC1 standard errors, robust standard errors clustered at the state level, and Conley (1999) standard errors. In Panel A, we display two forms of the SCPC standard errors. These differ according to whether or not the alternative critical value associated with this procedure is incorporated into the standard error.[15] The SCPC standard errors, once adjusted in this way, more closely approximate the TMO correction, though, unlike the TMO correction, they under-reject in about 20% of the specifications.

4. Consistency and Gaussian Approximation across Correlated Outcomes

The main non-standard feature of our setting is that the data under consideration are correlated across both locations and outcomes. In this section, we consider how both sources of correlation affect the methods developed in Section 3. We have two objectives. First, we give sufficient conditions under which the spatial correlation across units is consistently estimable. In particular, we give a quantitative description of the extent of the correlation across outcomes and units allowed by our estimator. Second, we give sufficient conditions

[15] Unless otherwise specified, when reporting results associated with SCPC, we incorporate the critical value into the estimate of the standard error.

FIGURE 4. Distributional Comparison of Performance Across Treatments



Notes: Figure 4 displays CDFs of the bias and rejection rates associated with various methods across several specifications of the calibrated simulation experiment outlined in Section 2.2. In particular, the distributions are taken across the specifications where each of the 91 outcome variables discussed in Section 2 is set as the treatment variable. For each specification, we take 1,000 simulation draws. In Panels A and B, the vertical red lines display the hypothetical performance of an unbiased estimator and an unbiased test, respectively. In Panel A, we display two forms of the SCPC standard errors. The label "SCPC (Adj.)" denotes that the alternative critical value associated with the SCPC procedure has been incorporated into the standard error.

---

for the validity of a Gaussian approximation to the empirical distribution of residual correlations of null pairs of units. Here, we highlight how the extent of the correlation across outcomes impacts the dispersion of the null distribution and the quality of the Gaussian approximation.

4.1 Sparsity and Consistency

We begin by imposing a series of simplifications that will greatly ease the notational burden, while preserving the key features of our setting. First, we focus our attention on inference for the mean $\tau^{(0)} = \mathbb{E}[Y_i^{(0)}]$. That is, we set $W_i = 1$ for all units. Second, we impose the normalization $\mathbb{E}[Y_i^{(j)}] = 0$ for all units i and outcomes j and consider the simplified estimators

$$\hat{\lambda}_{i,i'} = \frac{1}{d}\sum_{j=1}^{d} \tilde{Y}_i^{(j)} \tilde{Y}_{i'}^{(j)}, \quad \text{where} \quad \tilde{Y}_i^{(j)} = \hat{\gamma}_j^{-1/2} Y_i^{(j)} \quad \text{and} \quad \hat{\gamma}_j = \frac{1}{n}\sum_{i=1}^{n} (Y_i^{(j)})^2 . \tag{4.1}$$

That is, relative to the feasible estimators (3.5) and (3.6), the infeasible estimators (4.1) have been centered with the true values of the unknown parameters $\mathbb{E}[Y_i^{(j)}]$.

The variance estimator proposed in Section 3 is premised on the assumption that there is not too much dependence across either locations or outcomes. Here, we describe this assumption formally. That is, we give sufficient conditions on the scope of the dependence across units and outcomes under which we can recover the spatial correlation across units.

Our results are expressed in terms of the matrices Λ_n and Γ_d introduced in Assumption 3.1. Recall that the matrix Λ_n parametrizes the structure of the dependence across units. We assume that the rows of this matrix are sparse, in the following sense.

Assumption 4.1 (Sparsity)**.** *There exists a constant $0 \leq q < 1$ and a sequence κ_n such that the components of Λ_n satisfy*

$$\max_{i \in \{1,\ldots,n\}} \sum_{i'=1}^{n} |\lambda_{i,i'}|^q \leq \kappa_n \tag{4.2}$$

for each integer n, where we adopt the convention that $0^0 = 0$.

In the extreme case that $q = 0$, Assumption 4.1 means that at most κ_n elements of any row of Λ_n are non-zero. Non-zero values of q accommodate less stringent forms of sparsity.[16] In turn, the matrix Γ_d parametrizes the structure of the dependence across outcomes. We place no *a priori* restrictions on this quantity. In the statement of our results we will reference the matrix

$$\Omega_d = \Gamma_d^{1/2} \Xi_d \Gamma_d^{1/2}, \quad \text{where} \quad \Xi_d = \mathsf{diag}(1/\gamma_{1,1}, \ldots, 1/\gamma_{d,d}) \tag{4.3}$$

and $A^{1/2}$ denotes the square root of the matrix A. Observe that Ω_d can be interpreted as the correlation matrix across outcomes. Let $\omega_{j,j'}$ denote the j, j'th element of Ω_d.

The following Theorem characterizes the rate of convergence of the estimator $\hat{\gamma}_0 \hat{\lambda}_{i,i'}$, where we normalize the covariance estimator $\hat{\lambda}_{i,i'}$ by the variance estimate $\hat{\gamma}_0$ to fix the scale. To ease exposition, throughout the main text, we allow the constants c and C to depend on the largest and smallest diagonal terms of the matrices Λ_n and Γ_d as well as the maximum sub-Gaussian norm of the data $Y_i^{(j)}$. We track the dependence of the rate of convergence on these quantities in the proof.

Theorem 4.1 (Consistency)**.** *Suppose that a strengthened version of Assumption 3.1, stated in Appendix C.3, and Assumption 4.1 hold. Fix a constant $0 < \phi < 1$ and define the sequence*

$$\varphi_{n,d} = \left(\frac{1}{d} \sqrt{\sum_{j=1}^{d} \sum_{j'=1}^{d} \omega_{j,j'}^2} + \sqrt{\frac{\kappa_n}{n}} \right) \log^{3/2}(nd/\phi) \,. \tag{4.4}$$

[16] A version of Assumption 4.1 was first considered in the normal-means problem by Abramovich et al. (2006). It has been applied to estimation of covariance matrices in independent data by Bickel and Levina (2008) and Cai and Liu (2011), among others.

There exist constants $c < 1$ and C such that if $\varphi_{n,d} < c$, then

$$\max_{i,i' \in \{1,\ldots,n\}} |\hat{\gamma}_0 \hat{\lambda}_{i,i'} - \lambda_{i,i'}| \leq C\varphi_{n,d} \tag{4.5}$$

with probability greater than $1 - C\phi$.

Theorem 4.1 bounds the error in our estimate of cross-sectional dependence in terms of two quantities:

$$\frac{1}{d}\sqrt{\sum_{j=1}^{d}\sum_{j'=1}^{d}\omega_{j,j'}^2} \quad \text{and} \quad \sqrt{\frac{\kappa_n}{n}}\,. \tag{4.6}$$

The former parametrizes the extent of the correlation across outcomes. Observe that if the matrix Ω_d is diagonal, then this term is equal to $d^{-1/2}$, *i.e.,* the parametric rate. As the correlation across outcomes increases, this term becomes larger, and the precision of our estimate decreases.

Consequently, when constructing collections of outcomes, researchers face an essential trade-off. Outcomes that more credibly reflect the same underlying correlation structure might tend to be highly correlated. If outcomes are too similar, the resultant estimate of cross-sectional dependence will be imprecise. If outcomes are too different, they might not capture the same cross-sectional correlation, and estimates of cross-sectional dependence will be biased. We test approaches for navigating this trade-off in the applications considered in Section 5. We summarize our recommendations in Section 6.

The second term in (4.6) describes the sparsity of the matrix Λ_n, and arises due to the error in variance estimator $\hat{\gamma}_j$. The interpretation of this quantity is cleanest when Assumption 4.1 holds with $q = 0$. In this case, our estimate of cross-sectional dependence is accurate so long as the maximal number of non-zero elements of any row of Γ_n is small relative to the total number of units n. As a result, there is less scope to apply our procedure in settings where the number of units is small, e.g., in cases where the units of observation are U.S. states. This situation echos results in the literature on estimation of clustered standard errors. For example, Hansen and Lee (2019) show that standard estimators of clustered standard errors are consistent only if the number of units in each cluster is small relative to the sample size.

4.2 Gaussian Approximation

In Section 3.4, we apply a Gaussian approximation to the empirical distribution of the correlations of the residuals of the null pairs of units. Approximations of this sort are widely applied in independent data (see e.g., Efron 2007 for discussion). However, in our context, there are two forms of dependence that complicate this program. First, outcomes are likely to be highly correlated with each other. Second, the correlations $\hat{\rho}_{i,i'}$ and $\hat{\rho}_{i,i''}$ will be dependent. In this section, we characterize how dependence across outcomes, and across correlations taken with the same unit, impact a Gaussian approximation to the null distribution.[17]

[17] Efron (2010) and Azriel and Schwartzman (2015) give less specialized methods for approximating the empirical distribution of correlated observations.

In particular, continuing with the simplifications introduced in the previous subsection, we give a bound on the difference between the cumulative distribution

$$G_{n,d}(\delta) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{i'<i} \mathbb{I}\{\hat{\gamma}_0 \hat{\lambda}_{i,i'} \leq \delta\} \, . \tag{4.7}$$

and an appropriately scaled Gaussian distribution function. For the sake of simplicity, we assume that all pairs of units are nulls. Further, as the distribution function (4.7) is taken over covariances, rather than correlations, we assume that the data are homoskedastic. That is, we assume that the matrix Λ_n is given by λI_n for some constant λ, where I_n is the $n \times n$ identity matrix. Moreover, we assume that the underlying data $Y_i^{(j)}$ are Gaussian. Even in this highly stylized case, the calculations involved are non-standard.

Theorem 4.2 (Gaussian Approximation)**.** *Suppose that Assumption 3.1 holds, $\Lambda_n = \lambda I_n$ for some constant λ, and that the data $Y_i^{(j)}$ are normally distributed. Let $\eta_1, \ldots, \eta_d$ denote the eigenvalues of Ω_d, defined in (4.3). Fix the parameter*

$$v^\star = \frac{\lambda^2}{d^2} \sum_{j=1}^{d} \sum_{j'=1}^{d} \omega_{j,j'}^2 \tag{4.8}$$

and recall that $\Phi_v(\cdot)$ denotes the cumulative distribution function of a mean-zero Gaussian random variable with variance v. Fix a constant $0 < \phi < 1$. For each threshold δ in $\mathbb{R}$, it holds that

$$|G_{n,d}(\delta) - \Phi_{v^\star}(\delta)| \leq C \left(\frac{\sum_{j=1}^{d} \eta_j^3}{(\sum_{j=1}^{d} \eta_j^2)^{3/2}} + \sqrt{\frac{\log^3(dn/\phi)}{n}} \right) \tag{4.9}$$

with probability greater than $1 - \phi$.

Theorem 4.2 establishes that the empirical distribution of the null covariances fluctuates around a Gaussian distribution whose variance $v^\star$ is given by (4.8). Observe that the square-root of $v^\star$ appears in the bound (4.4). Written differently, given a collection of outcomes with correlation matrix Ω_d, the true "effective sample-size" or "degrees of freedom" is given by

$$\mathsf{df}^\star = \frac{1}{v^\star} = \frac{1}{\lambda^2} \frac{d^2}{\sum_{j=1}^{d} \sum_{j'=1}^{d} \omega_{j,j'}^2} \, . \tag{4.10}$$

If the outcomes are mutually independent, then $\mathsf{df}^\star = d/\lambda^2$. As the correlation across outcomes increases, the effective sample size decreases, and the dispersion of the null distribution increases.

Correlation across outcomes also reduces the quality of the Gaussian approximation. Consider the term

$$\frac{\sum_{j=1}^{d} \eta_j^3}{(\sum_{j=1}^{d} \eta_j^2)^{3/2}} \tag{4.11}$$

appearing in the bound (4.9). If the outcomes are mutually independent, then all of the eigenvalues of the correlation matrix Ω_d are equal to one, and the error (4.11) is equal to $d^{-1/2}$. Indeed, this is the order of the error that we should expect in independent data. As the correlation across outcomes increases, this

term becomes larger. In the extreme, when Ω_d is rank one, having a single eigenvalue equal to d, the error (4.11) is equal to one. In sum, correlation across outcomes impacts both our ability to discriminate nulls from non-nulls, through the dispersion of the null distribution, and our ability to accurately recover the null distribution through a Gaussian approximation.

5. APPLICATIONS

We now examine the impact of applying the TMO variance estimator, as well as alternative spatial standard error estimators, to a sample of recently published papers in applied economics. We find that the proposed estimator can make a substantial difference in practice.

5.1 Spatial Correlation in Applied Economics

We start off by documenting the prevalence of the potential for spatial correlation in a set of recently published papers. We review all 370 papers published in 2023 in the *American Economic Review*, *Econometrica*, the *Journal of Political Economy*, the *Quarterly Journal of Economics*, and the *Review of Economic Studies*. We hand-code each paper as belonging to four (overlapping) categories: Econometrics (26 papers), Economic Theory (108 papers), Macroeconomics (66 papers), and Empirical work (197 papers). We take an expansive view of the latter category, including macroeconomic, econometric, or theoretical papers with empirical applications. Within the 197 papers with empirical content, there are 17 laboratory experiments, 28 field experiments (RCTs), and 23 descriptive papers (e.g., measures of inequality or inter-generational mobility). Empirical studies in these three categories do not typically give rise to the issue of spatial correlation in the errors, due to the experimental design or to the nature of the issue being studied.

We classify the remaining 128 papers, a third of the original sample, as belonging to an "observational" category, which includes all (non-randomized) studies of how variation in a variable W_i affects an outcome variable Y_i. This category includes difference-in-difference designs, structural papers, and shift-share instruments, among other designs. Within these 128 papers, we code whether a main specification of the paper can be affected by spatial correlation of the observations. If so, we record the geographical level of such correlation. In our assessment, spatial correlation of the errors plays a role in a striking number of cases: 61 cases out of 128, or 48 percent, of observational papers. These spatial designs are present in all five journals and especially common in the *Review of Economic Studies* (21 papers), the *American Economic Review* (18 papers), and the *Quarterly journal of Economics* (12 papers). Our categorization is likely a lower bound of the importance for spatial correlation because in some papers the analysis has a spatial aspect, but it is not emphasized, e.g., the study of establishment-level productivity in cases in which the location of the establishment is not recorded.

The most common geographic unit at which the correlation is likely to occur is the U.S. county, with 15 papers in this category. For papers examining outcomes in the United States (31 papers overall), the next most common geographic units are states (6 papers), census blocks (2 papers), zip codes (2 papers),

TABLE 3. Survey of Papers Published in Leading Economics Journals

	All	*AER*	*ECTA*	*JPE*	*QJE*	*RES*
All papers (N)	370	93	66	71	48	92
Fields (not mutually exclusive)						
Economic Theory	108	18	23	34	6	27
Econometrics	26	4	12	2	2	6
Macroeconomics	66	15	12	15	7	17
Empirical	197	67	20	30	33	47
Method						
Descriptive	23	5	1	3	4	10
Randomized control trial	28	11	4	5	4	4
Lab experiment	17	8	2	2	3	2
Observational	128	41	13	19	22	33
Potential for spatial correlation						
No	67	23	10	12	10	12
Yes	61	18	3	7	12	21
U.S. unit of observation	31	9	2	2	7	11
Most common: County	15	4	1	1	6	3
Second most common: State	6	1	1	1	1	2
Non-U.S. unit of observation	30	9	1	5	5	10
Most common: Country	6	2	1	2	1	0
Second most common: Chinese county	2	1	0	0	1	0

Notes: Table 3 gives the results of a survey of papers published in the *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Review of Economic Studies*, and *Quarterly Journal of Economics* in 2023. Each paper is categorized by field. Empirical papers are categorized according to their primary methodology. Observational papers are categorized according to whether their primary analyses have the potential to be affected by spatial correlation. For these papers, the geographic level of such correlation is coded.

and commuting zones (2 papers). For papers with coverage outside the United States (30 papers), the most common levels of geography are the country (6 papers), Chinese counties (2 papers), followed by a variety of other local data units (e.g., French municipalities, Austrian regions, and Ukrainian districts). Given that the most common spatial-based setting in this sample is the county-level spatial design, accounting for 8 percent of all empirical papers, in the next subsection we focus on this category of papers, systematically revisiting the estimates in the 15 county-level papers.

5.2 Results

For the 15 papers that use U.S. county-level observations, we query replication packages and, where necessary, authors to access underlying data. For eight out of the 15 papers, we are not able to reproduce a main specification, due to a key variable being confidential. In the remaining seven cases, we identify and reproduce a main empirical specification. Panel A of Table 4 displays the point estimate, original standard

error, and level of clustering used for each of these papers. For the majority, there is no clustering at a higher level of geographic aggregation. (Online appendices often include additional corrections, typically a Conley (1999) standard error.) The papers differ in a number of ways, ranging from cross-sections to panel data (as indicated by the number of periods in Column 7), from modern to historical outcome data, and span the most common observational designs. For each of the papers, to apply the TMO correction we construct a collection of auxiliary outcome variables, typically from replication packages associated with the papers themselves. In several cases, these outcomes are supplemented with additional variables obtained from external sources. Column 8 reports the number of outcomes obtained for each paper. Further details on how we handle the data associated with each paper, and construct a relevant set of auxiliary outcomes, are given in Appendix E.

Column 1 presents the key finding: the ratio of the TMO standard error (Column 2) to the original standard error (Column 3). For each paper, we use the TMO procedure to augment the original level of clustering (Column 4). That is, a county pair (i, i') is never thresholded if counties i and i' are elements of the same cluster. Written differently, the TMO procedure augments the original standard error by allowing for dependence between pairs of counties, from different clusters, whose absolute correlation $|\hat{\rho}_{i,i'}|$ is above the threshold $\hat{\delta}^{\star}$ (Column 10). Column 11 reports the proportion of location pairs that meet this criteria and contribute to the TMO standard error. The proportion varies from 0.18% to 4.06%. TMO leaves the standard error unaltered in one case, and increases the remaining six. The median increase is 37% percent.

To illustrate that TMO is applicable to geographic units beyond U.S. counties, we consider two additional, prominent, non-county papers: Chetty et al. (2014), which consider U.S. commuting zone level data, and Acemoglu et al. (2019), which studies country level data.[18] The findings are broadly consistent, although the magnitude of the corrections are both on the lower end of the U.S. county-level corrections.

As a concrete example, Bernini et al. (2023) studies the impact of the Voting Rights Act on the racial makeup of local governments in the Southern U.S. We obtain 60 outcome variables from the replication package, which we standardize and use to compute the correlation across counties. Given the correlation across outcomes, find that the null distribution has approximately 26 degrees of freedom (as reported in Column 9), stressing the importance of starting with a large number of outcomes. For this paper, the optimal threshold is set at 0.54, keeping only 0.7% of county-pairs (as reported in Columns 10 and 11). The TMO method increases the standard error by 37% relative to the estimate reported in the paper.

Figure 5 provides further details concerning the application to Bernini et al. (2023). Panel A shows the distribution of the correlation estimates $\hat{\rho}_{i,i'}$. The solid curve corresponds to the estimate of the null density, proposed in Section 3.4. Panel B plots the ratio of the TMO standard error and the original standard error, as the threshold δ varies. At low thresholds (< 0.3), the standard error ratio hovers around 1, indicating the high prevalence of noise, relative to signal, added to the standard error. In the neighborhood around the optimal threshold $\hat{\delta}^{\star} = 0.54$, the TMO estimate is relatively stable. As the threshold increases, the standard error reverts to the the value reported in the paper.

[18] Müller and Watson (2024) also consider the data from Chetty et al. (2014).

TABLE 4. Ratio of TMO to Original Standard Errors in Nine Recent Papers

SE Ratio (TMO/Orig.)	TMO	Original SE	Clustering	Coefficient	Units n	Periods	Outcomes d	$\widehat{\text{df}}$	$\hat{\delta}^\star$	$\% \geqslant \hat{\delta}^\star$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Panel A. County-level papers from 2023 survey										
Cook et al. (2023): Effect of White casualties in WWII on Green Book establishments (Table 3 Column 4)										
3.56	0.03	0.01	County	0.06	3104	12	76	37.8	0.41	3.64
Caprettini and Voth (2023): Effect of New Deal grants on purchase of WWII bonds (Table 2 Column 1)										
2.10	0.05	0.02	County	0.19	3022	1	44	26.1	0.51	1.25
Esposito et al. (2023): Effect of The Birth of a Nation screening on patriotic vs. divisive language in newspapers (Table 3 Column 3)										
1.76	0.12	0.07	County	1.09	786	132	172	192.4	0.19	4.06
Bernini et al. (2023): Effect of Black % × Voting Rights Act on share of black elected officials (Table 2 Column 4)										
1.37	0.06	0.04	Judicial div.	0.10	971	1	60	25.8	0.54	0.70
Bazzi et al. (2023): Effect of Southern Whites % on 2016 Trump vote-share (Table 2 Column 4)										
1.20	0.20	0.16	60mi^2	1.03	1886	1	60	21.6	0.54	1.99
Calderon et al. (2023): Effect of Black population on Democratic vote-share (Table 2 Column 6)										
1.10	0.49	0.44	County	1.88	1263	3	78	21.4	0.55	1.62
Moscona and Sastry (2023): Effect of extreme temperature exposure × innovation on price per agricultural land acre (Table 3 Column 1)										
1.00	0.09	0.09	State	0.25	3000	2	39	17.1	0.67	0.18
Panel B. Prominent non-county examples										
Chetty et al. (2014): Correlation between share of African American and upward mobility (Figure 8 Row 1)										
1.08	0.05	0.05	State	-0.36	693	1	158	14.4	0.70	0.30
Acemoglu et al. (2019): Effect of democracy on log GDP per capita (Table 2 Column 3)										
1.04	0.24	0.23	Country	0.78	174	47	79	78.7	0.28	5.73

Notes: Panel A of Table 4 shows the results from applying TMO to seven recent county-level papers surveyed in Table 3. Panel B shows the results for two prominent non-county examples, Chetty et al. (2014) at the commuting-zone level and Acemoglu et al. (2019) at the country level. Column 1 reports the ratio of the TMO standard error estimate (Column 2) relative to the original standard error in each paper (Column 3). The TMO standard error in Column 2 is combined with the original level of clustering in the paper (Column 4). The table also shows the coefficient on the main regressor of interest (Column 5), the number of locations in each sample (Column 6), the number of time periods (Column 7), the number of outcomes used in the TMO procedure (Column 8), the estimated degrees of freedom for the null distribution (Column 9), the optimal threshold $\hat{\delta}^\star$ that maximizes (3.18) (Column 10), and the percent of location pairs (i, i') that have a correlation $\hat{\rho}_{i,i'}$ greater than $\hat{\delta}^\star$ in absolute value, out of all location pairs (i, i') where i and i' are in different clusters (Column 11).

5.3 Comparison

How does TMO compare, in practice, to other popular corrections for spatial correlation? Column 1 of Table 5 displays the ratio of the TMO standard error to the original standard error. Here, unlike Column 1 of Table 4, TMO does not augment the original level of clustering. That is, TMO is permitted to threshold pairs of counties in the same cluster. Column 2 gives the analogous ratio, where now TMO augments state level clusters, i.e., pairs of counties in the same state are never thresholded. In most cases, the standard error increases moderately, although the increase tends not to be dramatic. Combining TMO with state level clustering, in this way, is well suited for settings where policies are adopted at the state level.[19]

Columns 3 and 5 display analogous results for two leading alternatives: Conley (1999) standard errors, constructed with a 150 mile bandwidth, and Müller and Watson (2022, 2023) SCPC standard errors. For each procedure, we report the ratio of the adjusted standard error to the original standard error. In several cases, the Conley (1999) and SCPC are similar to the TMO. In others, they are qualitatively different. Columns 4 and 6 display analogous ratios, where now TMO has been applied to augment the Conley (1999) and SCPC, respectively. Combining the two forms of correction tends to lead to further increases in the resultant standard error. This is intuitive—distance-based methods are better suited to capture more diffuse, geographically driven, dependencies. The TMO standard errors, by contrast, will better capture strong dependencies that are not driven by geographic proximity.

The applications that we have considered thus far exhibit spatial correlation predicted well by both geographic and non-geographic factors. In an exercise similar to Table 1, Table A.1 displays, for each of the U.S. county or commuting zone level papers, the proportion of highly correlated pairs of locations that are similar along observable characteristics. Pairs of locations with strong positive correlations tend to be close (within the top 10%) in at least one of these observable characteristics. While geographic proximity is typically the first or second most predictive dimension, there are still a significant proportions of location pairs that are highly correlated, but not necessarily close in geographic distance.

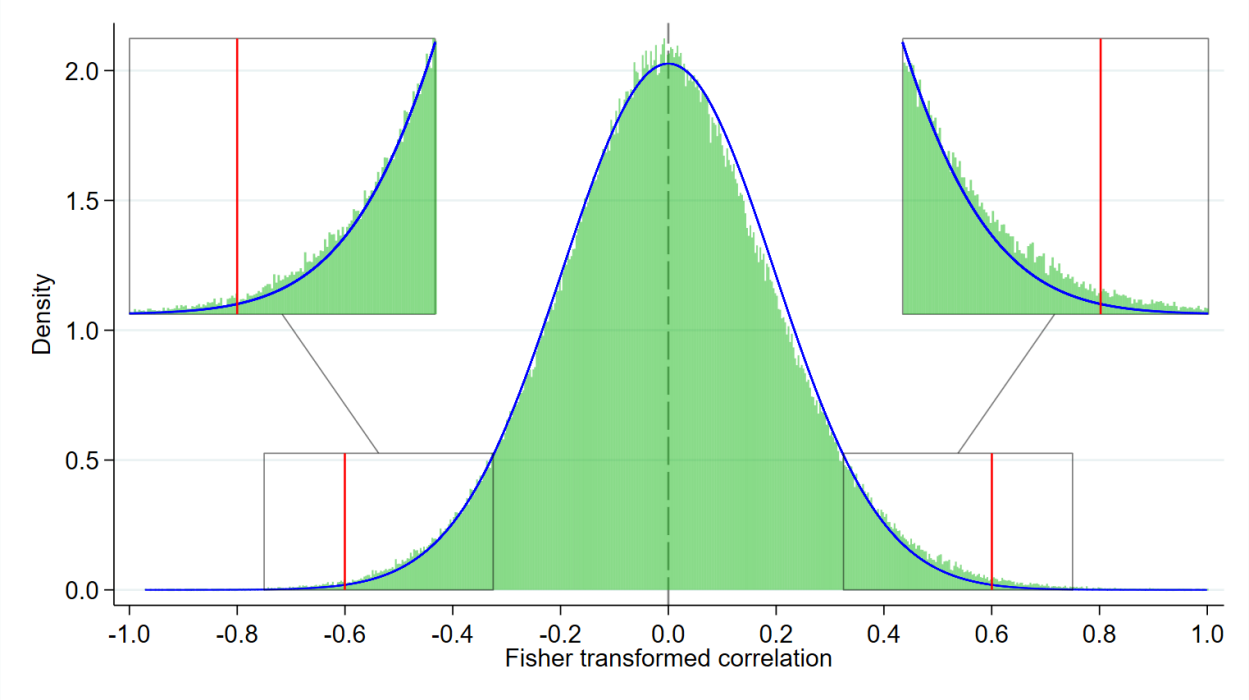
6. Recommendations for Practice

Economic outcomes are often linked across locations. Statistical inferences that do not account for these linkages can exhibit substantial biases. This paper introduces a method for adjusting standard errors for spatial correlation. We call this method "Thresholding Multiple Outcomes" (TMO). The essential input is a collection of auxiliary outcome variables. The main assumption is that the spatial correlation in the residuals for a regression problem of interest is shared by the analogous residuals constructed using these auxiliary outcomes. Under this assumption, the auxiliary outcomes can be used to estimate spatial correlation. We propose to use the empirical distribution of these estimates to determine pairs of locations that are very

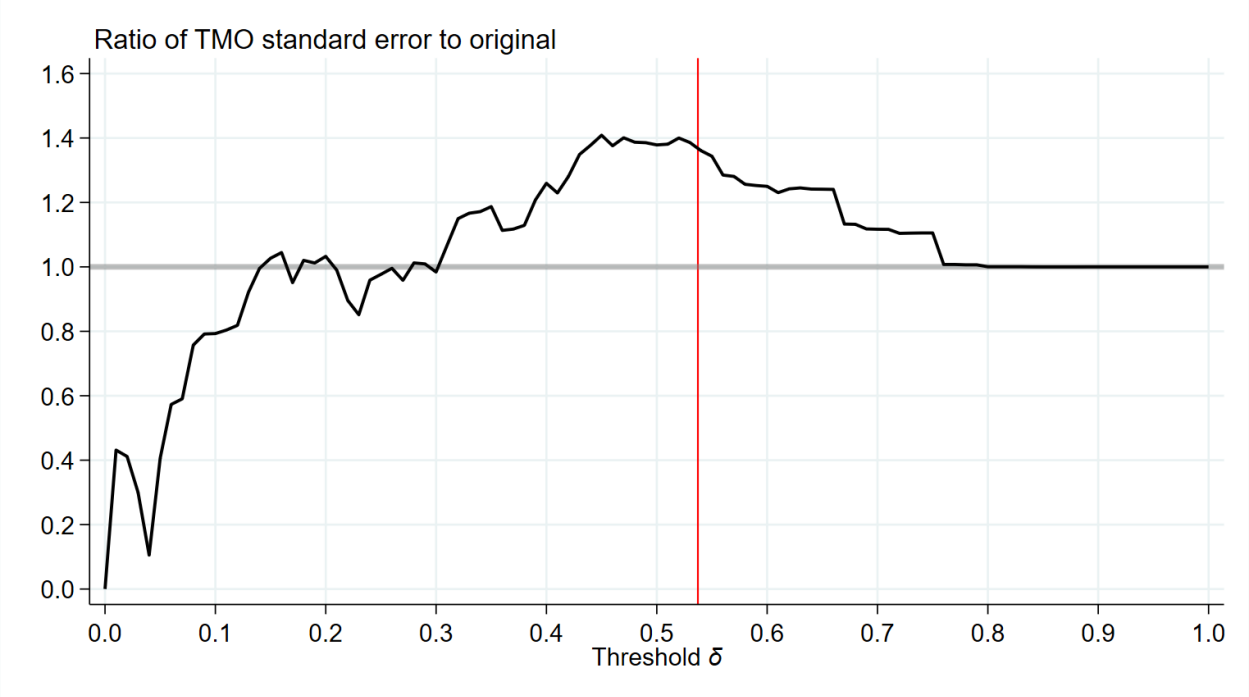
[19]Clustering by state can be problematic in samples with a small number of states. For example, the data used in Bernini et al. (2023) cover only 12 states. There, 12% of county pairs are in the same state.

FIGURE 5. Details for Application to Bernini et al. (2023)

Panel A: Correlation Between Pairs of Counties



Panel B: TMO across Thresholds



Notes: Figure 5 gives details concerning the application of TMO to Bernini et al. (2023). In Panel A, we plot the distribution of the pairwise correlation estimates, $\tilde{\rho}_{i,i'}$, constructed using 60 auxiliary outcomes obtained from the paper's replication package. The solid curve denotes the null density estimate, obtained with the procedure outlined in Section 3.4. Panel B shows the ratio of the TMO standard error to the original standard error as the threshold δ varies. In both panels, the vertical line marks the estimate of the optimal threshold $\hat{\delta}^{\star}$.

TABLE 5. Comparison with Alternative Spatial Standard Errors

	TMO		Conley 150mi		SCPC	
		+State Clusters		+TMO		+TMO
	(1)	(2)	(3)	(4)	(5)	(6)
Cook et al. (2023)	3.56 [3.6%]	3.63 [6.2%]	0.97 [3.1%]	3.37 [6.2%]	1.37	4.01
Caprettini and Voth (2023)	2.10 [1.3%]	2.37 [4.1%]	2.16 [3.2%]	2.48 [4.2%]	1.79	2.98
Esposito et al. (2023)	1.76 [4.1%]	1.87 [7.4%]	1.53 [3.9%]	1.97 [7.6%]	0.87	1.85
Bernini et al. (2023)	1.32 [0.7%]	1.72 [12.3%]	1.41 [9.0%]	1.52 [9.6%]	1.98	2.22
Bazzi et al. (2023)	1.14 [2.1%]	1.43 [5.4%]	1.47 [4.5%]	1.46 [6.0%]	1.12	1.57
Calderon et al. (2023)	1.10 [1.6%]	0.50 [5.6%]	0.95 [5.2%]	0.95 [6.2%]	2.79†	2.40†
Moscona and Sastry (2023)	0.64 [0.3%]	1.00 [3.3%]	0.96 [3.2%]	0.95 [3.4%]	0.94	1.07
Chetty et al. (2014)	0.72 [0.4%]	1.08 [3.2%]	0.98 [2.6%]	1.03 [2.9%]	2.59	2.55
Acemoglu et al. (2019)	1.04 [5.7%]	-	1.03 [3.0%]	1.07 [7.9%]	0.99	1.07

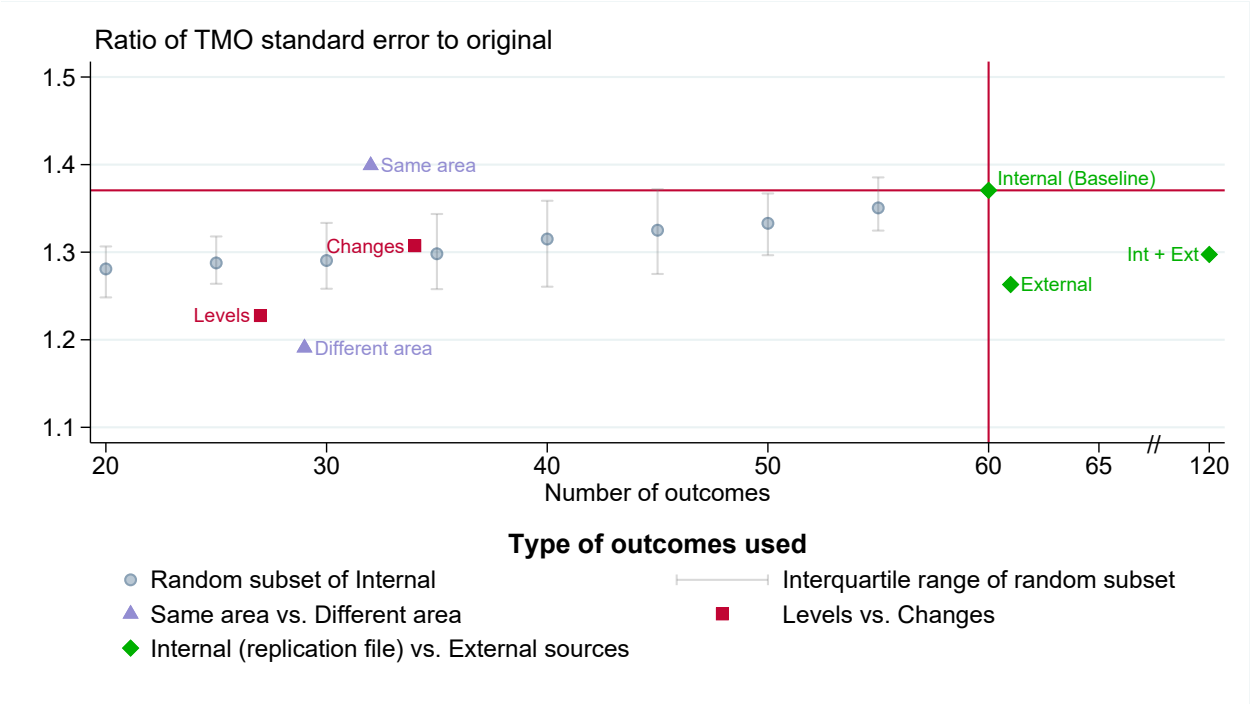
Notes: Table 5 compares the behavior of several methods for adjusting standard errors for spatial correlation. Each entry gives the ratio of the standard error produced by the method denoted in each column to the original standard error. Wherever applicable, the proportion of location pairs that are allowed to correlate is displayed in brackets. Column 1 reports the TMO estimates. Here, TMO does not augment the original level of clustering in the paper (which leads to differences from Column 1 in Table 4). Column 2 gives the analogous ratio when TMO augments U.S. state-level clusters (which does not apply to the country-level Acemoglu et al. (2019) paper). Column 3 reports the adjustment associated with Conley (1999) standard errors using a 150-mile bandwidth. Column 4 combines TMO with the Conley correction. For the country-level Acemoglu et al. (2019) paper, the bandwidth is set to 650 miles, which covers 3% of all country-pairs. Column 5 displays the results from the Müller and Watson (2022, 2023) SCPC method. Column 6 integrates TMO with the SCPC estimates. Columns 5 and 6 do not report the proportion of location pairs that are allowed to correlate, as SCPC does not directly "zero out" terms as in the Equation (3.10) formula. †Calderon et al. (2023) conduct a weighted regression, which is not supported by the current SCPC Stata package; these estimates show the adjustment for the unweighted specification.

correlated. Standard errors in the original regression problem can then be adjusted by allowing for correlation among these pairs.

It is worth emphasizing that we view the TMO adjustment as as a complement, not a substitute, to standard error corrections based on geographic distance. Indeed, in the examples above, combining the TMO adjustment with clustered, Conley (1999), or SCPC standard errors often impacts the TMO adjustment. In practice, combinations of this form are straightforward to implement and are likely to offer the greatest robustness to different forms of spatial correlation.

We conclude by detailing some recommendations for practice. We focus our discussion on the construction of a relevant set of auxiliary outcomes. We again use Bernini et al. (2023) as a running example to contextualize our recommendations. Figure 6 shows the ratio of the TMO standard error to the original standard error across different choices of the set of auxiliary outcomes. We conduct two experiments. First, we compute the TMO standard error using random subsamples of the 60 outcome variables sourced

FIGURE 6. TMO Adjustment for Different Sets of Auxiliary Outcomes



Notes: Figure 6 shows the ratio of the TMO standard error to the original standard error in Bernini et al. (2023), for different choices of auxiliary outcomes. The baseline TMO estimate uses the set of 60 *Internal* outcomes from the Bernini et al. (2023) replication file. The circular points show the average TMO adjustment across 20 randomly drawn subsets of the *Internal* outcomes, where the number of randomly drawn outcomes in each subset is shown on the x-axis. For example, the circular point at $x = 30$ indicates the average adjustment across 20 rounds of drawing 30 of the *Internal* outcomes randomly and running TMO with those 30 outcomes. The intervals correspond to the interquartile range of the TMO adjustment across the 20 rounds. *Same area* refers to the subset of *Internal* outcomes that have a racial, demographic, or political characteristic like the main outcome of interest (the change in the share of black elected officials), while *Different area* represents the subset of the remaining *Internal* outcomes. The *Changes* subset includes the *Internal* outcomes that are expressed in differences, whereas the *Levels* set includes those expressed in levels. The set of *External* outcomes are from external sources such as the Census and are based on data from more recent time periods. *Int + Ext* is the union of the *Internal* and *External* outcomes.

from the replication package associated with this paper. The circular points denote the average adjustment, where the size of the subsample varies along the x-axis. The intervals around each point show the 25-75th percentiles of the adjustment, over 20 random draws. The size of the adjustment increases with the number of outcomes.[20] With fewer outcomes, the null distribution has a higher variance and fewer degrees of freedom, making the optimal threshold higher, and reducing our ability to detect non-nulls. Ultimately, with a limited number of outcomes, the procedure has low power to distinguish true correlations from noise. This exercise highlights the importance of using a sufficiently large number of outcomes, so that the correlation estimates are sufficiently precise to distinguish pairs of locations with correlated errors.

Second, we change various characteristics of the auxiliary outcomes. In Bernini et al. (2023), the outcome of interest is the change in the share of black elected officials in a long-difference regression. Among the 60

[20] Figure A.3 reproduces Panel A of Figure 5 using a random subset of 20 outcomes.

"Internal" outcomes taken from the replication file, around half are "Similar" with either a racial, demographic, or political component (e.g., change in NAACP branches), and the remaining half are "Different" outcomes that cover other areas (e.g., the unemployment rate). Though both sets have roughly 30 outcomes each, using the "Similar" set of outcomes produces a much stronger adjustment. Furthermore, using outcomes that are expressed in changes, like the outcome of interest, also generates a higher adjustment than using outcomes in levels. This exercise demonstrates the importance of identifying outcomes whose underlying spatial correlation is *relevant* for the outcome of interest.

These two experiments illustrate a fundamental trade-off between the number of auxiliary outcomes and their relevance to the outcome of interest. Adding more outcomes can lead to less informative estimates of the standard error if the additional outcomes do not reflect the structure of correlation in the outcome of interest. For example, Figure 6 shows that adding a set of around 60 "External" outcomes (e.g., median age from Census data) reduces the adjustment to the standard error, relative to using only the 60 "Internal" outcomes. In cases where researchers can choose from a large set of outcomes, we recommend that they aim to select the most relevant ones that provide at least 20 degrees of freedom.[21]

Of course, in the application to Bernini et al. (2023), we do not know what the standard error *should* be, i.e., whether larger adjustments are in fact indicative of better performance. Figure A.4 displays the results of a set of experiments analogous to the results displayed in Figure 6, but in the context of the simulation exercise considered in Section 2. Reassuringly, similar patterns emerge: the size of the adjustment increases—toward the true standard error—with both the number of auxiliary outcomes and the relevance of the outcomes to the underlying correlation structure.

We also recommend two diagnostic exercises. The first is to plot the distribution of the correlation estimates and to check that the central mass of this distribution is well-approximated by a Gaussian. Heavily skewed distributions suggest that there is too much correlation across outcomes and that a larger sample of outcomes should be used. The second is to ensure that the function (3.18) varies smoothly with the threshold. Otherwise, the estimate for the optimal threshold may be unreliable. These issues are more prevalent in cases with fewer locations, such as U.S. state or country level analyses. For example, Funke et al. (2023) study the effect of populist leaders on GDP per capita growth rates for 60 countries. Figure A.5 displays these diagnostics implemented with data from this paper. We find that the Gaussian approximation to the null distribution is poor and that the estimate of the optimal threshold is unstable, due to the small number of units.

Finally, we stress that we see more avenues for further research along the lines outlined in this paper. For example, although our estimator can be applied to panel settings, as we have done in some of the examples discussed above, data sets with panel structures have special features which merit their own analysis. Furthermore, there may be alternative ways to use information on spatial correlations obtained from multiple

[21] We have found that larger degrees of freedom are necessary for reliable results in panel data settings.

outcomes. For instance, this information could be used to identify clusters of outcomes which share a similar spatial correlation structure, as opposed to assuming a common correlation among all outcomes considered.

REFERENCES

Abadie, A., Agarwal, A., Dwivedi, R., and Shah, A. (2024). Doubly robust inference in causal latent factor models. *arXiv preprint arXiv:2402.11652*.

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1):265–296.

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

Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34(2):584–653.

Acemoglu, D., Naidu, S., Restrepo, P., and Robinson, J. A. (2019). Democracy does cause growth. *Journal of political economy*, 127(1):47–100.

Adao, R., Kolesár, M., and Morales, E. (2019). Shift-share designs: Theory and inference. *The Quarterly Journal of Economics*, 134(4):1949–2010.

Allen, G. I. and Tibshirani, R. (2012). Inference with transposable data: modelling the effects of row and column correlations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74(4):721–743.

Andrews, I., Stock, J. H., and Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11(1):727–753.

Azriel, D. and Schwartzman, A. (2015). The empirical distribution of a large number of correlated normal variables. *Journal of the American Statistical Association*, 110(511):1217–1228.

Bailey, N., Holly, S., and Pesaran, M. H. (2016). A two-stage approach to spatio-temporal analysis with strong and weak cross-sectional dependence. *Journal of Applied Econometrics*, 31(1):249–280.

Bailey, N., Pesaran, M. H., and Smith, L. V. (2019). A multiple testing approach to the regularisation of large sample correlation matrices. *Journal of Econometrics*, 208(2):507–534.

Barrios, T., Diamond, R., Imbens, G. W., and Kolesár, M. (2012). Clustering, spatial correlations, and randomization inference. *Journal of the American Statistical Association*, 107(498):578–591.

Bazzi, S., Ferrara, A., Fiszbein, M., Pearson, T., and Testa, P. A. (2023). The other great migration: Southern whites and the new right. *The Quarterly Journal of Economics*, 138(3):1577–1647.

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

Bernini, A., Facchini, G., and Testa, C. (2023). Race, representation, and local governments in the us south: the effect of the voting rights act. *Journal of Political Economy*, 131(4):994–1056.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1):249–275.

Besley, T. and Case, A. (1995). Incumbent behavior: vote-seeking, tax-setting, and yardstick competition. *The American Economic Review*, 85(1):25–45.

Bester, C. A., Conley, T. G., Hansen, C. B., and Vogelsang, T. J. (2016). Fixed-b asymptotics for spatially dependent robust nonparametric covariance matrix estimators. *Econometric Theory*, 32(1):154–186.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227.

Bustos, P., Caprettini, B., and Ponticelli, J. (2016). Agricultural productivity and structural transformation: Evidence from brazil. *American Economic Review*, 106(6):1320–1365.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684.

Cai, T. T. and Jin, J. (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *The Annals of Statistics*, 38(1):100–145.

Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Annals Statistics*, 38(1):2118–2144.

Calderon, A., Fouka, V., and Tabellini, M. (2023). Racial diversity and racial policy preferences: the great migration and civil rights. *The Review of Economic Studies*, 90(1):165–200.

California Deprtment of Transportation (2023). Modoc county economic forecast. *Long-Term Socio-Economic Forecasts by County. https://dot.ca.gov/-/media/dot-media/programs/transportation-planning/documents/new-state-planning/transportation-economics/socioeconomic-forecasts/2023/2023-pdf/modoc-2023-a11y.pdf.*

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

Canay, I. A., Romano, J. P., and Shaikh, A. M. (2017). Randomization tests under an approximate symmetry assumption. *Econometrica*, 85(3):1013–1030.

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

Cao, J., Hansen, C., Kozbur, D., and Villacorta, L. (2024). Inference for dependent data with learned clusters. *Review of Economics and Statistics*, pages 1–45.

Caprettini, B. and Voth, H.-J. (2023). New deal, new patriots: How 1930s government spending boosted patriotism during world war ii. *The Quarterly Journal of Economics*, 138(1):465–513.

Case, A. C., Rosen, H. S., and Hines Jr, J. R. (1993). Budget spillovers and fiscal policy interdependence: Evidence from the states. *Journal of public economics*, 52(3):285–307.

Chetty, R., Hendren, N., Kline, P., and Saez, E. (2014). Where is the land of opportunity? the geography of intergenerational mobility in the united states. *The quarterly journal of economics*, 129(4):1553–1623.

Colella, F., Lalive, R., Sakalli, S. O., and Thoenig, M. (2023). acreg: Arbitrary correlation regression. *The Stata Journal*, 23(1):119–147.

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

Conley, T. G. and Kelly, M. (2025). The standard errors of persistence. *Journal of International Economics*, 153:104027.

Cook, L. D., Jones, M. E., Logan, T. D., and Rosé, D. (2023). The evolution of access to public accommodations in the united states. *The Quarterly Journal of Economics*, 138(1):37–102.

Dawid, A. P. (1981). Some matrix-variate distribution theory: notational considerations and a bayesian application. *Biometrika*, 68(1):265–274.

DellaVigna, S. and Kim, W. (2022). Policy diffusion and polarization across us states. Technical report, National Bureau of Economic Research.

Efron, B. (1982). Transformation theory: How normal is a family of distributions? *The Annals of Statistics*, pages 323–339.

Efron, B. (2007). Size, power and false discovery rates. *The Annals of Statistics*, pages 1351–1377.

Efron, B. (2010). Correlated z-values and the accuracy of large-scale statistical estimates. *Journal of the American Statistical Association*, 105(491):1042–1055.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Esposito, E., Rotesi, T., Saia, A., and Thoenig, M. (2023). Reconciliation narratives: The birth of a nation after the us civil war. *The American Economic Review*, 113(6):1461–1504.

Feller, W. (1991). *An introduction to probability theory and its applications, Volume 2*, volume 81. John Wiley & Sons.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521.

Funke, M., Schularick, M., and Trebesch, C. (2023). Populist leaders and the economy. *American Economic Review*, 113(12):3249–3288.

Giroud, X., Lenzu, S., Maingi, Q., and Mueller, H. (2024). Propagation and amplification of local productivity spillovers. *Econometrica*, 92(5):1589–1619.

Greenstone, M., Hornbeck, R., and Moretti, E. (2010). Identifying agglomeration spillovers: Evidence from winners and losers of large plant openings. *Journal of political economy*, 118(3):536–598.

Guggenberger, P., Kleibergen, F., and Mavroeidis, S. (2023). A test for kronecker product structure covariance matrix. *Journal of Econometrics*, 233(1):88–112.

Gupta, A. K. and Nagar, D. K. (1999). *Matrix variate distributions*. Chapman and Hall/CRC.

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

Hanson, D. L. and Wright, F. T. (1971). A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083.

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

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, pages 13–30.

Hoff, P. D. (2016). Limitations on detecting row covariance in the presence of column covariance. *Journal of Multivariate Analysis*, 152:249–258.

Hornstein, M., Fan, R., Shedden, K., and Zhou, S. (2018). Joint mean and covariance estimation with unreplicated matrix-variate data. *Journal of the American Statistical Association*.

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society. Series B (Methodological)*, 15(2):193–232.

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

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

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

Jin, J. and Cai, T. T. (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association*, 102(478):495–506.

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

Kim, M. S. and Sun, Y. (2011). Spatial heteroskedasticity and autocorrelation consistent estimation of covariance matrix. *Journal of Econometrics*, 160(2):349–371.

Langaas, M., Lindqvist, B. H., and Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to dna microarray data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(4):555–572.

Lehmann, E. L. and Romano, J. P. (2022). *Testing statistical hypotheses*. Springer.

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

Ludwig, J., Mullainathan, S., and Spiess, J. (2019). Machine-learning tests for effects on multiple outcomes.

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

Madestam, A., Shoag, D., Veuger, S., and Yanagizawa-Drott, D. (2013). Do political protests matter? evidence from the tea party movement. *The Quarterly Journal of Economics*, 128(4):1633–1685.

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

Moscona, J. and Sastry, K. A. (2023). Does directed innovation mitigate climate damage? evidence from us agriculture. *The Quarterly Journal of Economics*, 138(2):637–701.

Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of econometrics*, 32(3):385–397.

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

Müller, U. K. and Watson, M. W. (2022). Spatial correlation robust inference. *Econometrica*, 90(6):2901–2935.

Müller, U. K. and Watson, M. W. (2023). Spatial correlation robust inference in linear regression and panel models. *Journal of Business & Economic Statistics*, 41(4):1050–1064.

Müller, U. K. and Watson, M. W. (2024). Spatial unit roots and spurious regression. *Econometrica*, 92(5):1661–1695.

Muralidharan, O. (2010). Detecting column dependence when rows are correlated and estimating the strength of the row correlation. *Electronic Journal of Statistics*, 4:1527–1546.

Pollmann, M. (2020). Causal inference for spatial treatments. *arXiv preprint arXiv:2011.00373*.

Rudelson, M. and Vershynin, R. (2013). Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(none):1 – 9.

Singull, M. and Koski, T. (2012). On the distribution of matrix quadratic forms. *Communications in Statistics-Theory and Methods*, 41(18):3403–3415.

Song, Y., Chen, X., and Kato, K. (2019). Approximating high-dimensional infinite-order u-statistics: Statistical and computational guarantees. *Electronic Journal of Statistics*, 13(2).

Sun, L., Ben-Michael, E., and Feller, A. (2023). Using multiple outcomes to improve the synthetic control method. *arXiv preprint arXiv:2311.16260*.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

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

Zhou, S. (2014). Gemini: Graph estimation with matrix variate normal instances. *The Annals of Statistics*, 42(2):532–562.

Supplemental Appendix to:

Using Multiple Outcomes to Adjust Standard Errors for Spatial Correlation*

Stefano DellaVigna
UC Berkeley and NBER

Guido Imbens
Stanford University

Woojin Kim
UC Berkeley

David M. Ritzwoller
Stanford University

Contents

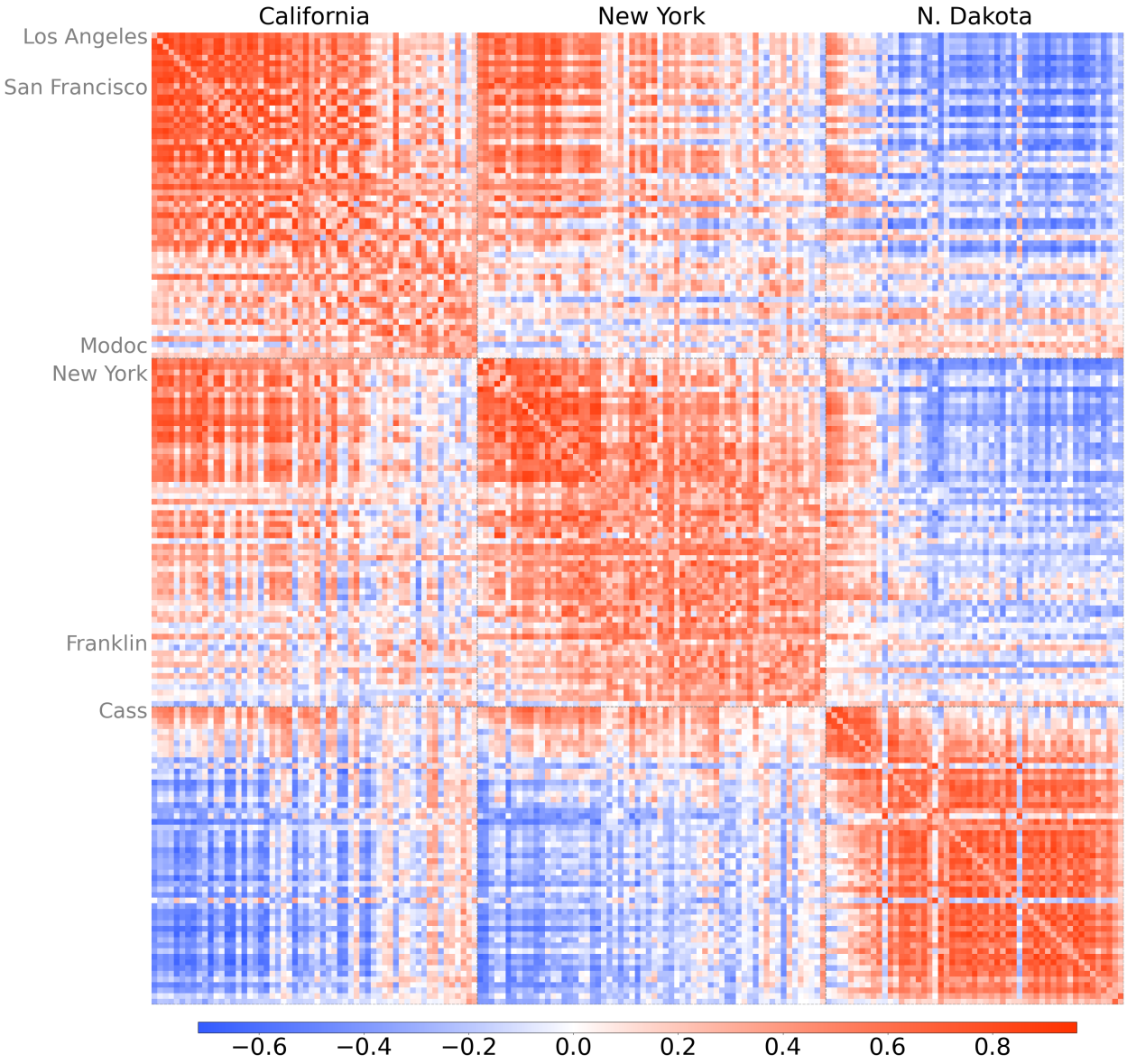
Appendix A. Auxiliary Figures and Tables 1
A.1. Figures 1
A.2. Tables 6
Appendix B. Details and Extensions 7
B.1. Outcome Selection and Cleaning 7
B.2. Simulation Design 13
B.3. Alternative Estimators and Data Structures 13
B.4. Alternative Null Distribution Estimators 15
Appendix C. Proofs for Results Stated in the Main Text 16
C.1. Proof of Theorem 3.1 16
C.2. Proof of Theorem 3.2 17
C.3. Proof of Theorem 4.1 18
C.4. Proof of Theorem 4.2 20
Appendix D. Proofs for Auxiliary Results 21
D.1. Proof of Lemma C.1. Part (i) 21
D.2. Proof of Lemma C.1. Part (ii) 24
D.3. Proof of Lemma C.2 24
Appendix E. Further Details for Empirical Applications 27
E.1. Bazzi et al. (2023) 27
E.2. Bernini et al. (2023) 28
E.3. Caprettini and Voth (2023) 28
E.4. Moscona and Sastry (2023) 29
E.5. Esposito et al. (2023) 30
E.6. Calderon et al. (2023) 31
E.7. Cook et al. (2023) 31
E.8. Chetty et al. (2014) 32
E.9. Acemoglu et al. (2019) 32

Date: March 20, 2025

APPENDIX A. AUXILIARY FIGURES AND TABLES

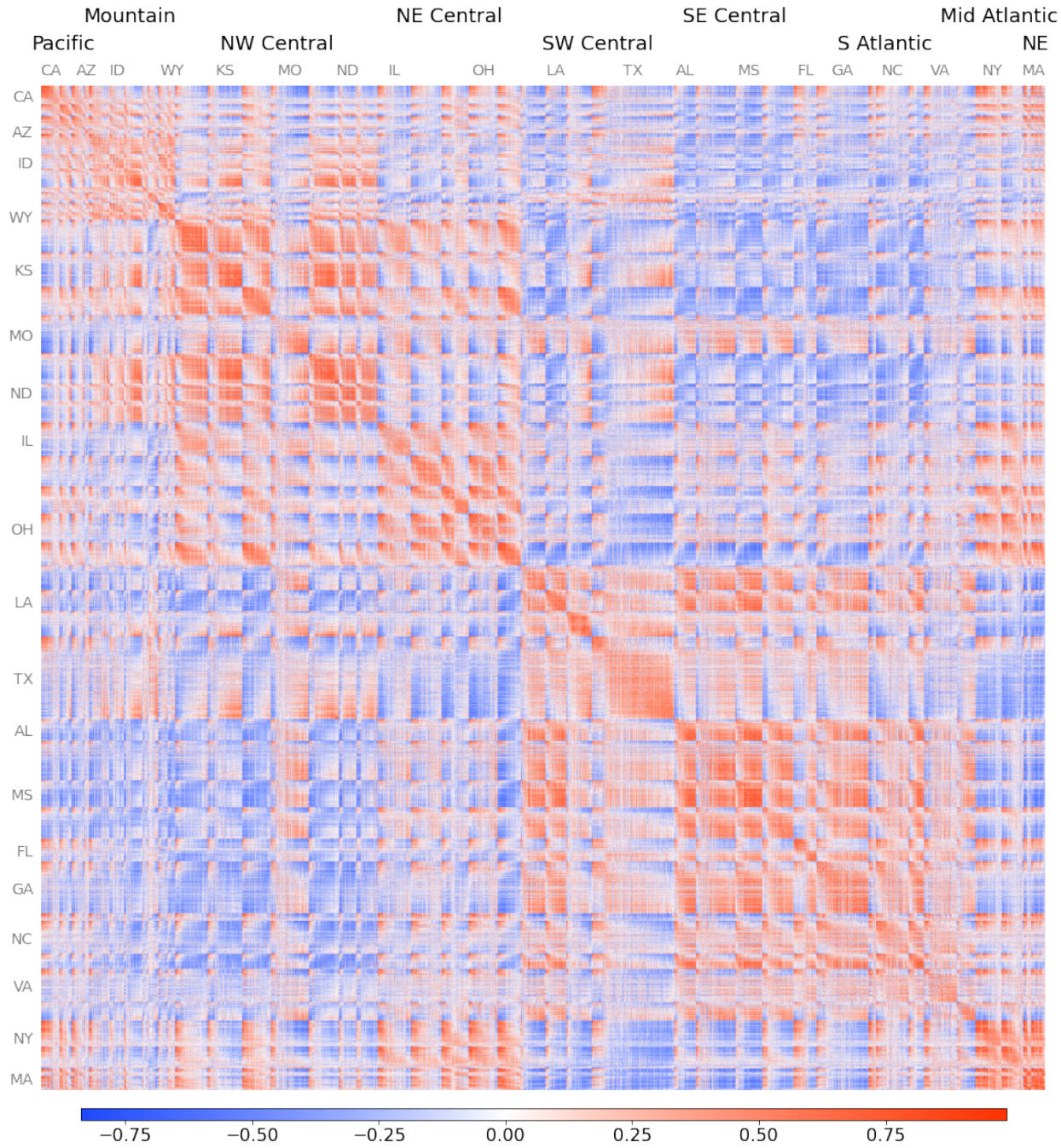
A.1 Figures

FIGURE A.1. Spatial Dependence Across Counties in Selected States



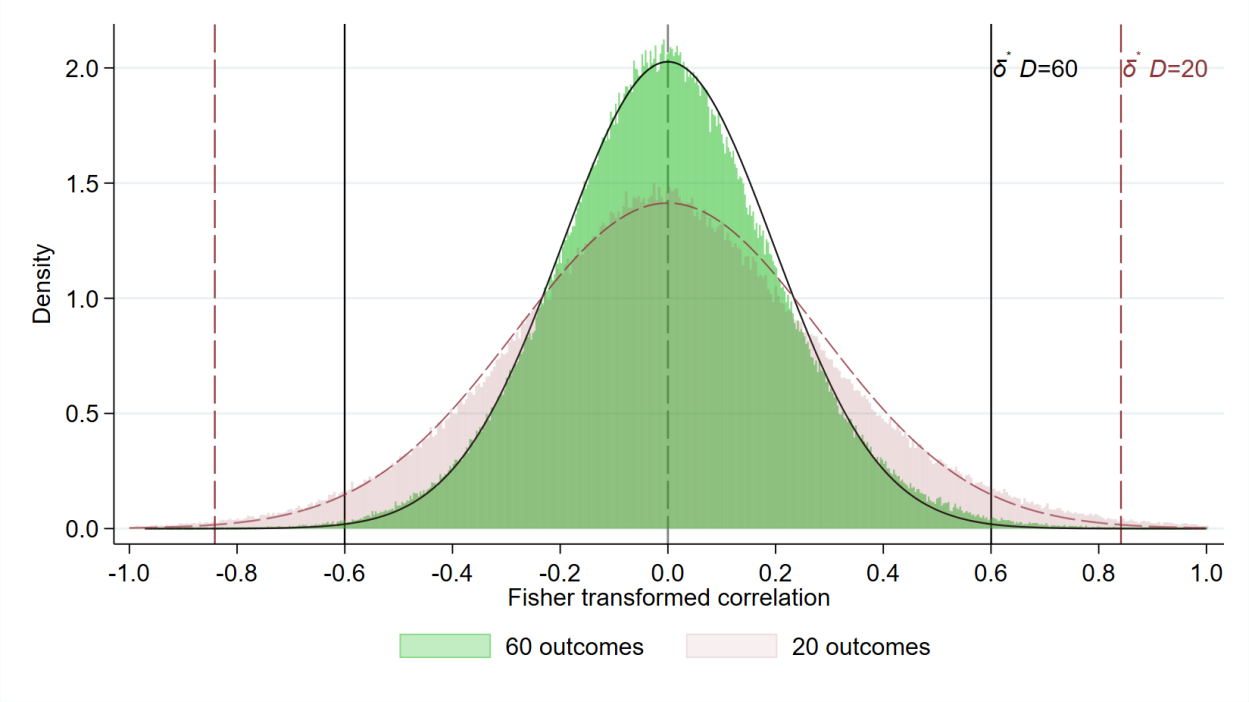
Notes: Figure A.1 displays a correllogram reflecting the measurements $\hat{\rho}_{i,i'}$ for each pair of U.S. counties in California, North Dakota, and New York State. Counties are sorted by population within each state.

FIGURE A.2. Spatial Dependence across U.S. Counties



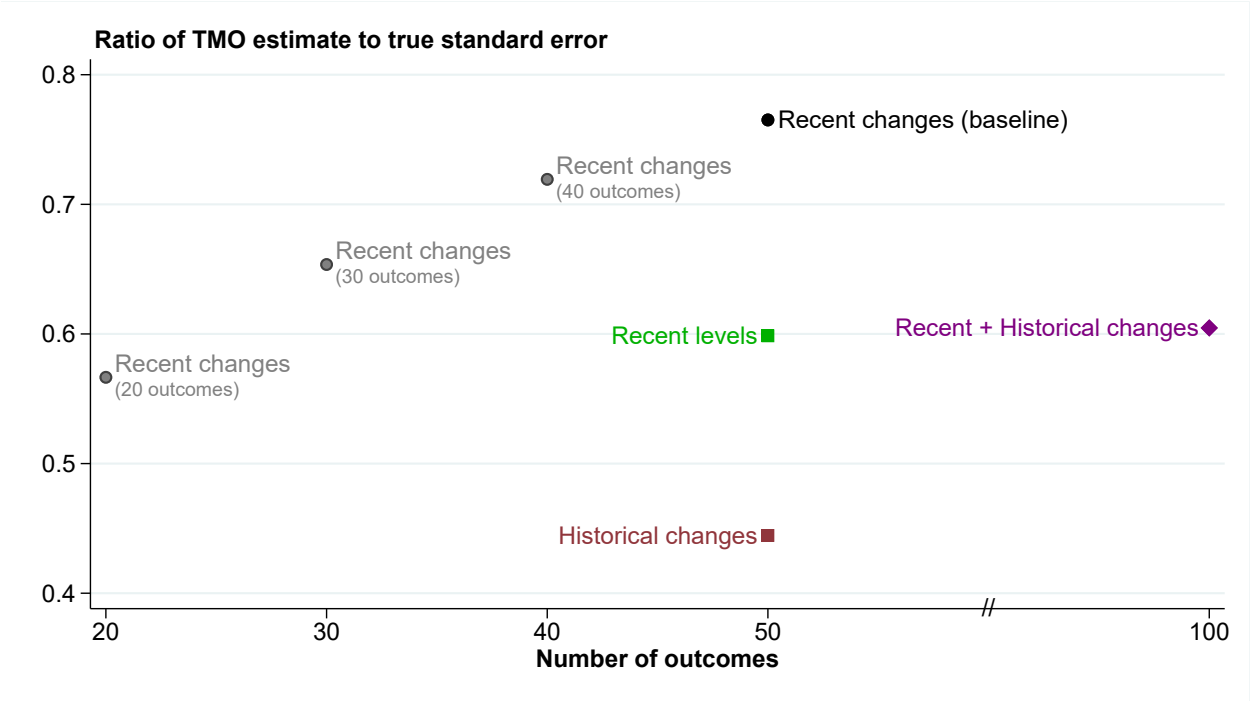
Notes: Figure A.2 displays a correllogram reflecting the measurements $\hat{\rho}_{i,i'}$ for each pair of U.S. counties. Counties are sorted by region, state, and population within each state.

FIGURE A.3. Correlation Between Pairs of Counties (20 vs. 60 Auxiliary Outcomes)



Notes: Figure A.3 builds on Panel A of Figure 5. As in Panel A of Figure 5, we plot the distribution of the pairwise correlation estimates, $\tilde{\rho}_{i,i'}$, in green constructed using 60 auxiliary outcomes obtained from the replication package associated with Bernini et al. (2023). The solid curve denotes the null density estimate, obtained with the procedure outlined in Section 3.4. The analogous histogram, constructed from a subsample of 20 randomly drawn outcomes, is displayed in red. In both cases, the associated optimal thresholds are displayed with vertical lines.

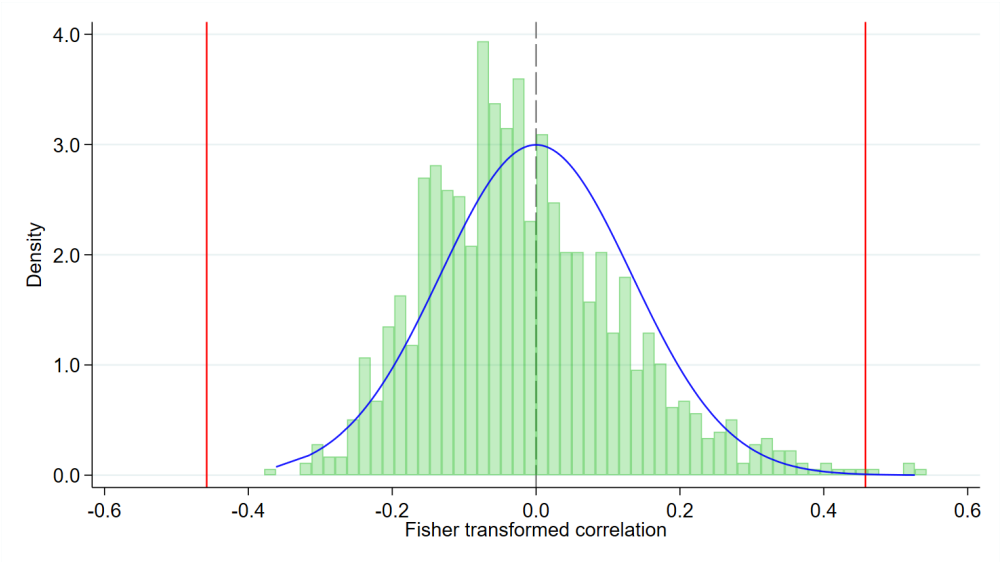
FIGURE A.4. Recommendations for Outcome Selection



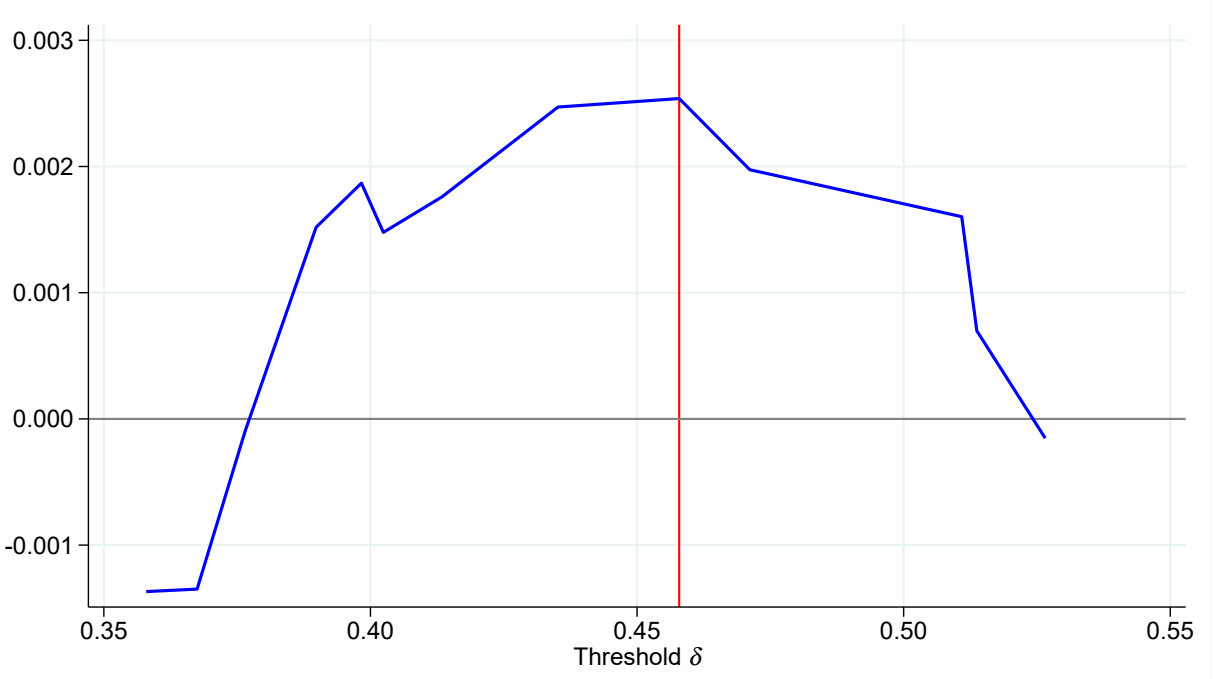
Notes: Figure A.4 shows the average ratio of the TMO standard error to the true standard error in the simulation developed in Section 2. We vary the set of auxiliary outcomes used to construct the TMO estimator. The treatment variable is the change in the percentage of the population in county i that has completed a college degree from 1980 to 2009. In each of the 1000 simulation rounds, the outcome of interest is a randomly drawn vector from a covariance structure based on recent changes in various economic, health, demographic, agricultural, and environmental variables in the last few decades (see Appendix B.1 and Appendix B.2 for details). The baseline *Recent changes* TMO adjustment uses 50 randomly drawn outcomes from the same covariance structure as the outcome of interest. The other *Recent changes* adjustments use fewer randomly drawn outcomes as shown on the x-axis, but still from the same covariance structure. The outcomes for *Recent levels* are drawn from a covariance structure based on the same outcomes listed in Appendix B.1, but the procedure in Appendix B.2 is performed on the average levels of these outcomes across years, rather than the changes between the years. The outcomes for *Historical changes* are drawn from a covariance structure built using the changes in different outcomes from the early to mid 1900s, rather than recent decades. *Recent + Historical changes* is the union of 50 randomly drawn outcomes from the *Recent changes* covariance structure and 50 randomly drawn outcomes from the *Historical changes* covariance structure.

FIGURE A.5. Diagnostic Tests for TMO Assumptions (Funke et al., 2023)

Panel A: Non-normal Distribution of Correlations



Panel B: Approximation to Non-Nulls Minus Nulls



Notes: Figure A.5 is analogous to Figure 5, but is constructed using data from Funke et al. (2023). Panel A plots the distribution of the pairwise correlation estimates, $\tilde{\rho}_{i,i'}$. Panel B shows the shows the value of the estimator (3.18) over a range of values of the threshold δ.

A.2 Tables

TABLE A.1. Predicting Significant and Positive Correlations between Locations in U.S. Applications

			Among county/CZ pairs (i, i') with $\hat{\rho}_{i,i'} > \delta$, percentage in which i ranks among the 10% closest to i' in:						
	$\hat{\delta}^*$	$\% \geqslant \hat{\delta}^\star$	Distance	Population	Urban %	Median income	Non-white %	Vote-share	Any of ←
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Cook et al. (2023)	0.41	2.74	46	22	26	22	21	16	80
Caprettini and Voth (2023)	0.51	0.81	35	16	16	15	16	15	64
Esposito et al. (2023)	0.19	2.70	25	16	15	16	15	14	61
Bernini et al. (2023)	0.54	0.47	34	17	18	20	20	16	68
Bazzi et al. (2023)	0.54	1.17	50	23	20	17	16	23	75
Calderon et al. (2023)	0.55	1.20	57	22	20	15	21	20	81
Moscona and Sastry (2023)	0.67	0.19	55	31	28	34	31	23	87
Chetty et al. (2014)	0.70	0.26	38	54	48	38	35	27	90

Notes: Table A.1 displays the results of exercise analogous to results reported in Table 4. Column 1 reports the optimal threshold estimated for each paper. Column 2 shows the percent of county/CZ-pairs that have a positive correlation (i.e., not in absolute value) above the threshold. Columns 3-9 show the percent of county/CZ-pairs (i, i') in which county/CZ i is within the top 10% of those closest to i' in the observable characteristic listed in the column header, among the pairs that have a correlation greater than a threshold $\hat{\delta}^\star$.

APPENDIX B. DETAILS AND EXTENSIONS

B.1 Outcome Selection and Cleaning

We construct a dataset of 91 county variables that reflect typical variables used in empirical research. We take these variables from the following sources:

- The Social Determinants of Health Database from the Agency for Healthcare Research https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html#download
- Bailey et al. (2016). "U.S. County-Level Natality and Mortality Data, 1915-2007." Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E100229V4
- U.S. Census Bureau. (2012). "Consolidated File: County Data, 1947-1977." Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR07736.v2
- U.S. Census Bureau. (2011). "USA Counties." Statistical Compendia. https://www.census.gov/library/publications/2011/compendia/usa-counties-2011.html

The set of variables are listed in Tables B.2a-B.2e. They cover a range of policy areas, including economics (e.g., income and labor force participation), health (e.g., life expectancy and health insurance coverage), demographics (e.g., population density and percent black), agriculture (e.g., cropland and percent employed in agriculture), and environment (e.g., water use and air pollution).

We perform the following data processing steps. For variables that scale with population (e.g., the number of black residents, public school enrollment, and manufacturing establishments), we normalize by the county population. We take the logarithm for all positive variables that do not have negative values or are not in percentages. Then for each variable, we keep the first and last years for which there are data for at least 3000 counties. We then standardize the variable to have mean zero and unit standard deviation within each year and winsorize at the 0.1 and 99.9 percentiles to prevent results being skewed by outliers. Finally, to construct the long-difference outcomes, we take the difference between the first and last years for which there are data.

TABLE B.2A. County outcomes used in simulation

Description	Source	First year	Last year	No. counties
1. Annual mean of Particulate Matter (PM2.5) concentration (g/m3)	AHRQ	2009	2018	3106
2. Average household size	AHRQ	2009	2018	3105
3. Gini index of income inequality	AHRQ	2009	2018	3105
4. Median distance in miles to the nearest emergency department, calculated using population weighted tract centroids in the county	AHRQ	2009	2018	3105
5. Median distance in miles to the nearest obstetrics department, calculated using population weighted tract centroids in the county	AHRQ	2009	2018	3105
6. Median distance in miles to the nearest pediatric ICU, calculated using population weighted tract centroids in the county	AHRQ	2009	2018	3105
7. Median selected monthly owner costs for houses with a mortgage (dollars)	AHRQ	2009	2018	3103
8. Percentage of families with children that are single-parent families	AHRQ	2009	2018	3106
9. Percentage of households with same-sex unmarried partner	AHRQ	2009	2018	3106
10. Percentage of limited English speaking households	AHRQ	2009	2018	3106
11. Percentage of occupied housing units with utility gas heating	AHRQ	2009	2018	3106
12. Percentage of population reporting American Indian and Alaska Native race alone	AHRQ	2009	2018	3106
13. Percentage of population reporting multiple races	AHRQ	2009	2018	3106
14. Percentage of population that speaks Spanish (ages 5 and over)	AHRQ	2009	2018	3106
15. Percentage of unmarried partner households that received food stamps/SNAP benefits	AHRQ	2009	2018	3102
16. Total cardiovascular disease deaths per 100,000 population (ages 35 and over)	AHRQ	2009	2018	3106
17. Total expenditure (Dollars) per student	AHRQ	2009	2018	3100
18. Total number of days with daily maximum heat index, absolute threshold: 90F	AHRQ	2009	2018	3106
19. Total number of hospital beds per 1,000 population	AHRQ	2009	2018	3106
20. Total number of hospitals per 1,000 population	AHRQ	2009	2018	3106

Sources: AHRQ: Agency for Healthcare Research and Quality, Bailey et al. (2016): www.doi.org/10.3886/E100229V4, Country Data Book: www.doi.org/10.3886/ICPSR07736.v2, IHME: Institute For Health Metrics and Evaluation, U.S. Census Bureau: USA Counties.

TABLE B.2B. County outcomes used in simulation

Description	Source	First year	Last year	No. counties
21. Total standardized Medicare costs, fee for service (dollars)	AHRQ	2009	2018	3104
22. Total stroke deaths per 100,000 population (ages 35 and over)	AHRQ	2009	2018	3102
23. Birth Weight by Residence (2500 grams or less)	Bailey et al. (2016)	1982	1988	3108
24. Nonmarital births by place of residence	Bailey et al. (2016)	1968	1988	3097
25. Civilian Labor Force.Pct.Male	County Data Book	1960	1970	3094
26. Divorce Rate Per 1000 Pop.	County Data Book	1970	1975	3044
27. Employed.Pct.In Manfgr	County Data Book	1950	1970	3053
28. Loc.Gov.Dir.Gen.Exp.Pct.Educ	County Data Book	1967	1972	3090
29. Loc.Gov.Dir.Gen.Exp.Pct.Hghway	County Data Book	1967	1972	3075
30. Loc.Gov.Prop.Tax.Per.Capita USD	County Data Book	1962	1972	3088
31. Manfgr.Estab.	County Data Book	1947	1972	3088
32. Marriage Rate Per 1000 Pop.	County Data Book	1970	1975	3101
33. Oasdhi Payments Per Mon. USD1000	County Data Book	1971	1976	3104
34. Ou.Pct.Owner Occupied	County Data Book	1940	1970	3087
35. Population Pct Foreign Stock	County Data Book	1960	1970	3095
36. Population Pct. Urban	County Data Book	1960	1970	3026
37. Population Rank	County Data Book	1950	1960	3092
38. Pub.Assis.Recipients.Afdc.	County Data Book	1972	1976	3079
39. Retail Trade Estab.	County Data Book	1954	1972	3096
40. Retail Trade Estab.Sales USD1000	County Data Book	1948	1972	3076

Sources: AHRQ: Agency for Healthcare Research and Quality, Bailey et al. (2016): www.doi.org/10.3886/E100229V4, Country Data Book: www.doi.org/10.3886/ICPSR07736.v2, IHME: Institute For Health Metrics and Evaluation, U.S. Census Bureau: USA Counties.

TABLE B.2C. County outcomes used in simulation

Description	Source	First year	Last year	No. counties
41. Workers.Pct.Used Pub.Trans	County Data Book	1960	1970	3095
42. Life expectancy	IHME	1999	2019	3113
43. All persons 18 to 64 years without health insurance, percent	U.S. Census Bureau	2005	2007	3108
44. All persons under 18 years without health insurance, percent	U.S. Census Bureau	2005	2007	3108
45. Average age of farm operators (NAICS)	U.S. Census Bureau	2002	2007	3064
46. Average travel time to work for workers 16 years and over who did not work at home	U.S. Census Bureau	1990	2009	3107
47. Average value of land and buildings per acre (NAICS)	U.S. Census Bureau	2002	2007	3058
48. Average value of land and buildings per farm (NAICS)	U.S. Census Bureau	2002	2007	3058
49. Births per 1,000 population	U.S. Census Bureau	1970	2007	3102
50. Black population	U.S. Census Bureau	1990	2010	3106
51. Civilian labor force	U.S. Census Bureau	1990	2010	3106
52. Civilian labor force unemployment rate	U.S. Census Bureau	1990	2010	3107
53. Commercial banks and savings institutions (FDIC-insured) - total deposits	U.S. Census Bureau	1980	2010	3103
54. Cropland - harvested cropland (NAICS) (acres)	U.S. Census Bureau	2002	2007	2963
55. Cropland - total (NAICS) (acres)	U.S. Census Bureau	2002	2007	3051
56. Deaths per 1,000 population	U.S. Census Bureau	1970	2007	3102
57. Earnings in retail trade (NAICS 44-45)	U.S. Census Bureau	2001	2007	2989
58. Educational attainment - persons 25 years and over - percent bachelor's degree or higher	U.S. Census Bureau	1980	2009	3105
59. Educational attainment - persons 25 years and over - percent high school graduate or higher	U.S. Census Bureau	1980	2009	3105
60. Employment in farming (NAICS)	U.S. Census Bureau	2001	2007	3077

Sources: AHRQ: Agency for Healthcare Research and Quality, Bailey et al. (2016): www.doi.org/10.3886/E100229V4, Country Data Book: www.doi.org/10.3886/ICPSR07736.v2, IHME: Institute For Health Metrics and Evaluation, U.S. Census Bureau: USA Counties.

TABLE B.2D. County outcomes used in simulation

Description	Source	First year	Last year	No. counties
61. Employment in government - state and local (NAICS)	U.S. Census Bureau	2001	2007	3078
62. Employment in retail trade (NAICS 44-45)	U.S. Census Bureau	2001	2007	2988
63. Federal Government direct loans FY	U.S. Census Bureau	1983	2010	3087
64. Hospital insurance and/or supplemental medical insurance (Medicare) - aged persons enrolled	U.S. Census Bureau	1998	2007	3084
65. Infant deaths per 1,000 live births	U.S. Census Bureau	1990	2007	3107
66. Land in farms (NAICS) (acres)	U.S. Census Bureau	2002	2007	3026
67. Median contract rent of specified renter-occupied housing units paying cash rent	U.S. Census Bureau	1980	2009	3105
68. Median household income	U.S. Census Bureau	1995	2009	3107
69. Median selected monthly owner costs of specified owner-occupied housing units with a mortgage	U.S. Census Bureau	1980	2000	3106
70. Median value of specified owner-occupied housing units	U.S. Census Bureau	1980	2009	3105
71. New private housing units authorized by building permits - total	U.S. Census Bureau	2004	2010	3107
72. People of all ages in poverty - percent	U.S. Census Bureau	1995	2009	3107
73. People under age 18 in poverty - percent	U.S. Census Bureau	1995	2009	3107
74. Per capita personal income	U.S. Census Bureau	1969	2000	3075
75. Persons 16 to 19 years not enrolled in school and not a high school graduate	U.S. Census Bureau	1990	2000	3108
76. Population per square mile	U.S. Census Bureau	1980	2010	3105
77. Private nonfarm establishments	U.S. Census Bureau	1990	2009	3107
78. Private nonfarm establishments - arts, entertainment and recreation (NAICS 71)	U.S. Census Bureau	2002	2009	3108
79. Public school enrollment Fall	U.S. Census Bureau	1987	2009	3088
80. Related children age 5 to 17 in families in poverty - percent	U.S. Census Bureau	1995	2009	3107

Sources: AHRQ: Agency for Healthcare Research and Quality, Bailey et al. (2016): www.doi.org/10.3886/E100229V4, Country Data Book: www.doi.org/10.3886/ICPSR07736.v2, IHME: Institute For Health Metrics and Evaluation, U.S. Census Bureau: USA Counties.

TABLE B.2E. County outcomes used in simulation

Description	Source	First year	Last year	No. counties
81. Renter-occupied housing units	U.S. Census Bureau	1990	2010	3106
82. Resident population under 18 years, percent	U.S. Census Bureau	2000	2009	3108
83. Resident population: Black alone, percent	U.S. Census Bureau	2000	2009	3059
84. Resident population: Median age	U.S. Census Bureau	1980	2010	3105
85. Resident population: total females, percent	U.S. Census Bureau	2000	2009	3108
86. Total physicians	U.S. Census Bureau	2004	2009	3108
87. Valuation of new private housing units authorized by building permits	U.S. Census Bureau	2004	2010	3108
88. Value of farm products sold - total (NAICS)	U.S. Census Bureau	2002	2007	3003
89. Vehicles available per occupied housing unit	U.S. Census Bureau	1990	2000	3108
90. Vote cast for president- percent Republican	U.S. Census Bureau	1980	2008	3105
91. Water use: per capita use	U.S. Census Bureau	1990	2005	3107

Sources: AHRQ: Agency for Healthcare Research and Quality, Bailey et al. (2016): www.doi.org/10.3886/E100229V4, Country Data Book: www.doi.org/10.3886/ICPSR07736.v2, IHME: Institute For Health Metrics and Evaluation, U.S. Census Bureau: USA Counties.

B.2 Simulation Design

For the simulation considered in Section 2.2, we construct covariance matrix Σ for the residual vector ε. Our aim is to ensure that the off-diagonal elements of Σ primarily consist of "non-null" pairs of counties, while maintaining positive definiteness.

To meet these objectives, we take as input the correlation matrix of the standardized outcomes. Next, we set a threshold δ, above which, we believe most pairs of counties with larger correlations should be non-nulls. After some experimentation, we set $\delta = 0.45$. We then find a "central" county that has the highest number of correlations with other counties that are above 0.45 in absolute value. That central county and all the counties with which it is sufficiently correlated are designated as a cluster. We find the next "central" county, among those that have not been assigned to a cluster, that has the highest number of correlations above 0.45 with other counties that have also yet to be assigned to a cluster. Note that once a county has been assigned to a cluster (as the central one or not), then it cannot be assigned to another cluster. We repeat the process, and once all the correlations above 0.45 have been exhausted, we set all correlations between counties outside the clusters to 0, and keep all the correlations within clusters intact. The resulting correlation matrix, ordered by clusters, has a block-diagonal structure that fulfills the objectives.

B.3 Alternative Estimators and Data Structures

In this section, we detail several extensions to the method proposed in Section 3. We keep the same notation. That is, we consider the linear model

$$Y_i^{(0)} = \alpha + \tau^{(0)} W_i + \theta^\top X_i \varepsilon_i^{(0)}, \quad i = 1, \dots, n\,, \tag{B.1}$$

where $Y_i^{(0)}$ measures some outcome of interest, W_i denotes some treatment of interest, and X_i denotes a k-vector of covariates. The variables $Y_i^{(1)}, \dots, Y_i^{(d)}$ denote d auxiliary, post-treatment outcomes. The variables $\varepsilon_i^{(j)}$ and $\hat{\varepsilon}_i^{(j)}$ denote the population and empirical residuals associated with the regression of the outcome $Y_i^{(j)}$ on the treatment W_i, respectively.

B.3.1 Covariates. In the main text, for ease of exposition, we treat the case where there are no covariates. The extension to the case with covariates is straightforward, by the Frisch-Waugh-Lovell theorem. In particular, let $\tilde{W}_i$ and $\tilde{Y}_i^{(j)}$ denote the residuals from regressing W_i and $\tilde{Y}_i^{(j)}$ on X_i, respectively. A standard error for $\hat{\tau}$ can then be constructed as before, by using these data.

B.3.2 Weighted regression. Throughout the main text, we study the construction of standard errors for ordinary least squares estimators. The proposed method can be easily adapted to weighted least squares estimators. In particular, suppose that we are interested in the estimator $\hat{\tau}_h^{(0)}$ defined by

$$\hat{\tau}_h^{(0)} = \arg\min_{\tau} \left\{ \sum_{i=1}^{n} h_i (Y_i^{(0)} - \tau W_i)^2 \right\}\,, \tag{B.2}$$

where $h = (h_i)_{i=1}^n$ is an exogenous weight vector. Observe that, in this case, we have that

$$\begin{aligned} \mathrm{Var}(\hat{\tau}_h^{(0)} \mid W) &= S_n(h)^{-2}(W^\top H \Sigma^0 H W), \quad \text{where} \\ S_n(h) &= W^\top H W \quad \text{and} \quad H = \mathsf{diag}(h)\,. \end{aligned} \tag{B.3}$$

In other words, the problem again reduces to the estimation of the residual covariance matrix $\Sigma^0 = \mathrm{Var}(\varepsilon^{(0)})$ and our method for estimating this matrix applies.

B.3.3 Instrumental Variables Regression. For each unit i, suppose that we observe the instrumental variable Z_i. If we are interested in constructing a standard error for the two-stage least squares estimator, we can simply apply our method to the second-stage regression. We caution that this approach is not robust to weak identification. The construction of identification robust standard error estimates for settings with spatial correlation is an interesting direction for future research. See Andrews et al. (2019) for a recent review of identification robust standard errors for linear instrumental variables regression in non-homoskedastic data.

B.3.4 Panel Regressions. Suppose that we observe the collection of outcomes across t time periods. That is, we observe $(Y_{i,s}^{(0)}, Y_{i,s}^{(1)}, \ldots, Y_{i,s}^{(d)})$ for each unit i in $1, \ldots, n$ and period s in $1, \ldots, t$. Here, we are often interested in constructing standard errors for coefficients in regressions of the form

$$Y_{i,s}^{(0)} = \alpha + \tau^{(0)} W_{i,s} + \theta^\top X_{i,s} \varepsilon_{i,s}^{(0)}\,, \tag{B.4}$$

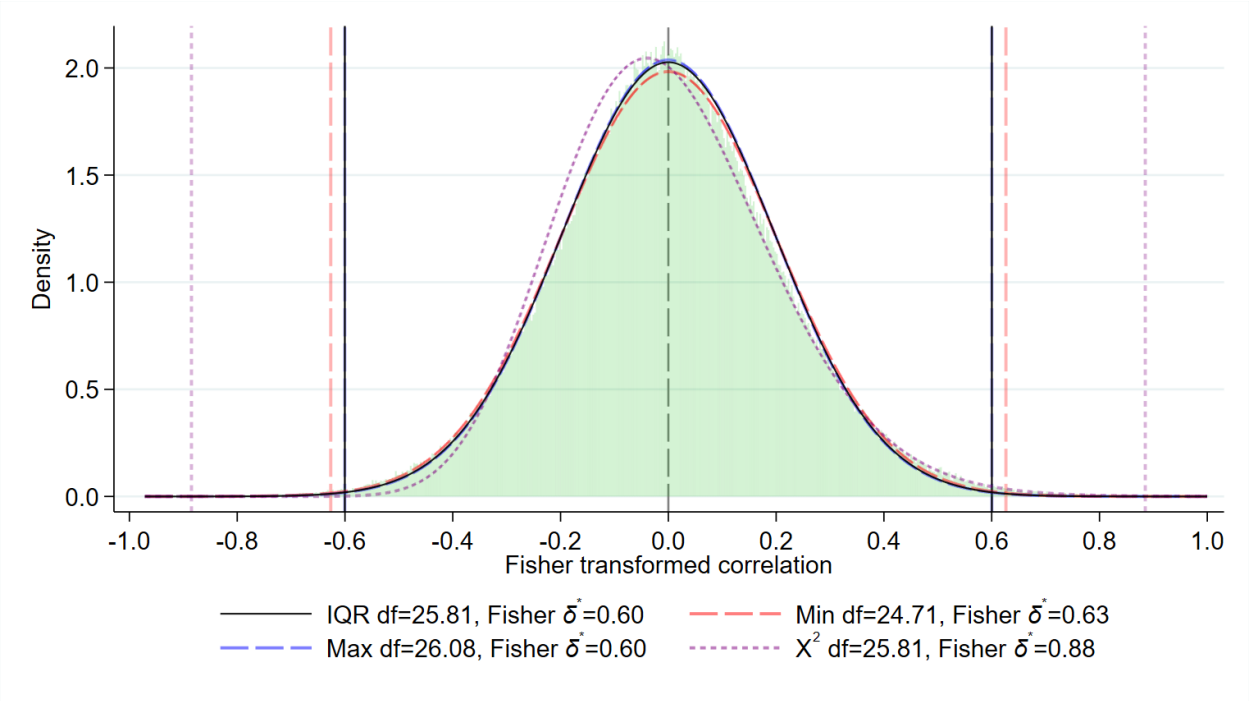
where, in many cases, the covariates $X_{i,s}$ contain various fixed effects. To adjust standard errors for correlation across units, there are at least two options. First, we might account for the correlation between the units i and i' only *within* each time period. That is, the residuals $(\varepsilon_{i,s}^{(j)}, \varepsilon_{i',s}^{(j)})$ are allowed to be correlate, but the residuals $(\varepsilon_{i,s}^{(j)}, \varepsilon_{i',s'}^{(j)})$, for $s \neq s'$, are not. We note that an analogous assumption is widely applied in treatments of two-way or multi-way clustered data (Menzel, 2021). An alternative option would be to allow for correlation *across* time periods as well. For the purposes of this paper, we adopt the latter approach. A more thorough consideration of the construction of standard errors for panel data, that accommodates spatial correlation across units is useful direction for further research.

In our treatment, we estimate the covariance between residuals associated with units i and i' with the estimator

$$\hat{\lambda}_{i,i'} = \frac{1}{dt} \sum_{j=1}^{d} \sum_{s=1}^{t} (\tilde{\varepsilon}_{i,t}^{(j)} - \overline{\varepsilon}_i)(\tilde{\varepsilon}_{i',t}^{(j)} - \overline{\varepsilon}_{i'}) \tag{B.5}$$

That is, in effect, we treat each outcome-period pair as a distinct outcome. These estimates are then be used to determine which pairs of units are "non-nulls" with the procedure developed in the main text. Standard errors are then adjusted, as before, allowing for correlation across units and across time periods.

FIGURE B.6. Robustness to Alternative Null Distribution Estimators



Notes: Figure B.6 builds on Panel A of Figure 5. As in Panel A of Figure 5, we plot the distribution of the pairwise correlation estimates, $\tilde{\rho}_{i,i'}$ using 60 auxiliary outcomes obtained from the replication package associated with Bernini et al. (2023). The solid black curve denotes the null density estimate, obtained with the procedure outlined in Section 3.4. The blue and red curves display the fits using the methods outlined in Appendix B.4 that give the largest and smallest threshold. A χ^2 approximation is also displayed.

B.4 Alternative Null Distribution Estimators

In this section, we consider alternative methods for estimating the null distribution. The methods we consider use the entire vector of off-diagonal correlations between the q and $1-q$ percentiles, where we vary $q \in \{0.1, 0.2, 0.25\}$. We group the correlations into granular bins. For a candidate variance v, we calculate the "distance" between the empirical and null distributions in one of two ways. The first computes the difference in the density at the center of each bin versus the null density at the midpoints. The second calculates the difference in the mass within each bin versus the mass within the endpoints of each bin under the null. Differences are calculated using either ℓ^1, ℓ^2, or ℓ^∞ norms. We compute the variance v that minimizes the difference. Figure B.6 is analogous to Panel A of Figure 5, but additionally displays the fits to the null distribution and maximize and minimize the resultant threshold. We also display a approximation to the null distribution based on a χ^2 distribution. The quality of the approximation, and the resultant threshold are insensitive to how the null distribution is estimated.

APPENDIX C. PROOFS FOR RESULTS STATED IN THE MAIN TEXT

C.1 Proof of Theorem 3.1

Throughout, we let the operator $\sim$ denote equality in distribution. Fix any constant $b > 0$ and define the matrix

$$\Sigma_b = b \cdot R\,, \quad \text{where} \quad R = W(W^\top W)^{-1}W^\top\,. \tag{C.1}$$

Suppose that, conditional on W, each pair of population residual vectors $\varepsilon^{(j)}$ and $\varepsilon^{(j')}$ had the joint distribution

$$\begin{pmatrix}\varepsilon^{(j)}\\ \varepsilon^{(j')}\end{pmatrix} \sim \mathsf{N}\left(\begin{pmatrix}\mathbf{0}_n\\ \mathbf{0}_n\end{pmatrix}, \begin{pmatrix}\Sigma_b & \Sigma_b\\ \Sigma_b & \Sigma_b\end{pmatrix}\right). \tag{C.2}$$

In this case, the data Y are jointly Gaussian conditional on W and Assumption 3.1 holds. Fix $\tau^{(j)} = 0$ for all j in $0, 1, ..., d$. The parameter of interest is given by

$$V(\Sigma_b) = (W^\top W)^{-1}W^\top\Sigma_b W(W^\top W)^{-1} = b(W^\top W)^{-1}\,. \tag{C.3}$$

Observe that the collection

$$(Y, W) = (Y^{(0)}, Y^{(1)}, \ldots, Y^{(d)}, W) \tag{C.4}$$

can be reconstructed from

$$(\hat{Y}, \hat{\varepsilon}, W) = (\hat{Y}^{(0)}, \hat{\varepsilon}^{(0)}, \hat{Y}^{(1)}, \hat{\varepsilon}^{(1)}, \ldots, \hat{Y}^{(d)}, \hat{\varepsilon}^{(d)}, W) \tag{C.5}$$

where

$$\hat{Y}^{(j)} = RY^{(j)} \quad \text{and} \quad \hat{\varepsilon}^{(j)} = (I_n - R)Y^{(j)}\,. \tag{C.6}$$

Moreover, conditional on W, we have that

$$(\hat{Y}^{(j)}, \hat{Y}^{(j')}, \hat{\varepsilon}^{(j'')}, \hat{\varepsilon}^{(j''')})^\top \sim \mathsf{N}\left(\begin{pmatrix}\mathbf{0}_{2n}\\ \mathbf{0}_{2n}\end{pmatrix}, \begin{pmatrix}bR \otimes \mathbf{1}_{2\times 2} & \mathbf{0}_{2n\times 2n}\\ \mathbf{0}_{2n\times 2n} & \mathbf{0}_{2n\times 2n}\end{pmatrix}\right), \tag{C.7}$$

where we have used the facts that

$$M\Sigma_b^{(0)}M = bMRM = \mathbf{0}_{n\times n} \quad \text{and} \quad bR^\top RR = bR\,. \tag{C.8}$$

Here, the operator $\otimes$ denotes the Kronecker product. Thus, there exists some locally bounded, measurable function $\breve{V}'(\cdot)$ such that

$$\breve{V}'(\hat{Y}, W) = \breve{V}(Y, W) \tag{C.9}$$

almost surely.

Let $\overline{r}$ denote the component of R that is largest in absolute value. Observe that there exists a constant c, that does not depend on n, d, or W such that the set

$$A(b) = \left\{\hat{y} = (\hat{y}^{(j)})_{j=0}^d \in \mathbb{R}^{(d+1)\times n} : \max_{i=1\ldots n}\max_{j=0,\ldots,d}\{|\hat{y}_i^{(j)}|\} \leq c\sqrt{b\overline{r}\log(dn)}\right\} \tag{C.10}$$

satisfies

$$P\left\{\hat{Y} \in A(b)\right\} \geq 1 - \eta \,. \tag{C.11}$$

for all $b > 0$. Define the functions

$$M(b) = \sup\{\check{V}'(\hat{y}, W) : \hat{y} \in A(b)\} \quad \text{and} \quad m(b) = \inf\{\check{V}'(\hat{y}, W) : \hat{y} \in A(b)\} \,, \tag{C.12}$$

respectively. Both functions are finite, as $\check{V}'(\cdot)$ is locally bounded.

Suppose that for each $b > 0$, we have both

$$b(W^\top W)^{-1} - M(b) \leq K \quad \text{and} \quad m(b) - b(W^\top W)^{-1} \leq K \,. \tag{C.13}$$

In this case, it would hold that

$$\check{V}'(\hat{y}, W) \in b(W^\top W)^{-1} \pm K \tag{C.14}$$

for each $\hat{y} \in A(b)$ and every $b > 0$, where we recall that $V(\Sigma_b) = b(W^\top W)^{-1}$. Now, for any fixed $\hat{y}_0$, the parameter b_0 can be chosen such that $\hat{y}_0 \in A(b)$ for all b greater than b_0. But then, it would be the case that

$$\check{V}'(\hat{y}_0, W) \in b(W^\top W)^{-1} \pm K \tag{C.15}$$

for all b greater than b_0. However, this must fail for some sufficiently large b, giving a contradiction. Consequently, there exists $b > 0$ such that

$$b(W^\top W)^{-1} - M(b) \leq K \quad \text{or} \quad m(b) - b(W^\top W)^{-1} \leq K \,. \tag{C.16}$$

In the first case, for all $\hat{y} \in A(n)$, we have that $\check{V}'(\hat{y}_0, W) < V(\Sigma_b) - K$, and so

$$P\left\{\check{V}(Y, W) - V(\Sigma_b) < -K\right\} \geq P\left\{A(b)\right\} \geq 1 - \eta \,. \tag{C.17}$$

In the second case, for all $\hat{y} \in A(n)$, we have that $\check{V}'(\hat{y}_0, W) > V(\Sigma_b) + K$, and so

$$P\left\{\check{V}(Y, W) - V(\Sigma_b) > K\right\} \geq P\left\{A(b)\right\} \geq 1 - \eta \,, \tag{C.18}$$

which completes the proof. ■

C.2 Proof of Theorem 3.2

Consider the loss

$$L(\delta) = (1 - p_0) \int_0^\delta t \, \mathrm{d}f_1(t) + p_0 \int_\delta^\infty t \, \mathrm{d}f_0(t) \,. \tag{C.19}$$

As we have assumed that the functions $f_0(t)$ and $f_1(t)$ are continuously differentiable, we can evaluate

$$\frac{\mathrm{d}}{\mathrm{d}\delta}\left((1 - p_0) \int_0^\delta t \, \mathrm{d}f_1(t)\right) = (1 - p_0)\delta f_1(\delta) \quad \text{and} \quad \frac{\mathrm{d}}{\mathrm{d}\delta}\left(p_0 \int_\delta^\infty t \, \mathrm{d}f_0(t)\right) = -p_0 \delta f_0(\delta) \,, \tag{C.20}$$

by Leibniz's rule. Thus, we find that

$$\frac{\mathrm{d}L(\delta)}{\mathrm{d}\delta} = \delta((1 - p_0)f_1(\delta) - p_0 f_0(\delta)) \,. \tag{C.21}$$

For $\delta > 0$, equalizing the derivative (C.21) with zero gives

$$(1 - p_0) f_1(\delta) = p_0 f_0(\delta) \,, \tag{C.22}$$

or equivalently

$$\mathsf{fdr}(\delta) = \frac{p_0 f_0(\delta)}{p_0 f_0(\delta) + (1 - p_0) f_1(\delta)} = \frac{1}{2} \,. \tag{C.23}$$

Hence, it will suffice to show that $\mathsf{fdr}(\delta)$ is strictly decreasing and continuous, as, in this case, there is a unique $\delta^\star$ that solves the first-order condition (C.23).

Observe that the continuity of $\mathsf{fdr}(\delta)$ follows immediately from the continuity of $f_0(t)$ and $f_1(t)$. To verify that it is strictly decreasing, observe that

$$\frac{\mathrm{d}}{\mathrm{d}\delta} \mathsf{fdr}(\delta) = \frac{p_0(1 - p_0)(f_0'(\delta) f_1(\delta) - f_0(\delta) f_1'(\delta))}{(p_0 f_0(\delta) - (1 - p_0) f_1(\delta))^2} \,. \tag{C.24}$$

The assumption that the distribution of the non-nulls is stochastically larger than the distribution of the nulls, implies that the likelihood ratio $f_0(\delta)/f_1(\delta)$ is decreasing. But notice that

$$\frac{\mathrm{d}}{\mathrm{d}\delta} \log \left(\frac{f_0(\delta)}{f_1(\delta)} \right) = \frac{f_0'(\delta)}{f_0(\delta)} - \frac{f_1'(\delta)}{f_1(\delta)} \,. \tag{C.25}$$

and that the sign of (C.24) is determined by the term

$$f_0'(\delta) f_1(\delta) - f_0(\delta) f_1'(\delta)) = f_0(\delta) f_1(\delta) \left(\frac{f_0'(\delta)}{f_0(\delta)} - \frac{f_1'(\delta)}{f_1(\delta)} \right) \,. \tag{C.26}$$

Hence, the function $\mathsf{fdr}(\delta)$ is decreasing, which completes the proof. ■

C.3 Proof of Theorem 4.1

Let Y and denote the $n \times d$ matrix whose (i, j)th element is given by $Y_i^{(j)}$. Throughout, we let $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|_{\mathrm{op}}$ denote the matrix Frobenius and ℓ_2-operator norms, respectively. Theorem 4.1 requires a slightly strengthened version of Assumption 3.1, stated as follows.

Assumption C.1 (Sub-Gaussian Proportionality)**.** *Let Z denote an $n \times d$ matrix whose components are independent, mean-zero, and sub-Gaussian with parameter $M \geq 1$. There exist positive semi-definite matrices $\Lambda_n = (\lambda_{i,i'})_{i,i'=1}^n$ and $\Gamma_d = (\gamma_{j,j'})_{j,j'=1}^d$ such that*

$$Y = \mathbf{1}_n \tau^\top + \Lambda_n^{1/2} Z \Gamma_n^{1/2} \,, \tag{C.27}$$

where $A^{1/2}$ denotes the square-root of the matrix A.

In addition to the restriction

$$\mathrm{Cov}(Y_i^{(j)}, Y_{i'}^{(j')}) = \lambda_{i,i'} \gamma_{j,j'} \tag{C.28}$$

specified by Assumption 3.1, Assumption C.1 implies that (i) the elements of Y are sub-Gaussian and that (ii) Y can be normalized to have fully independent, rather than just uncorrelated, entries by taking

$\Lambda_n^{-1/2}(Y - \mathbf{1}_n\tau^\top)\Gamma_n^{-1/2}$. Analogous versions of this assumption appear elsewhere in the literature (see e.g., Hoff, 2016).

Define the quantities

$$\overline{\gamma} = \max_{j\in[d]} \gamma_{j,j}, \quad \underline{\gamma} = \min_{j\in[d]} \gamma_{j,j}, \quad \overline{\lambda} = \max_{i\in[n]} \lambda_{i,i}, \quad \text{and} \quad \underline{\lambda} = \max_{i\in[n]} \lambda_{i,i}\ . \tag{C.29}$$

To ease exposition, in the statement of Theorem 4.1, the constants c and C are allowed to depend on the constants (C.29) as well as the sub-Gaussian constant M defined in Assumption C.1. For the sake of completeness, we track the dependence on these terms throughout the proof.

The result follows from the following two large deviation bounds. Throughout, we write $\xi_j = 1/\gamma_{j,j}$ and $\hat{\xi}_j = \hat{\gamma}_0/\hat{\gamma}_j$. Recall that we have normalized $\gamma_{0,0} = 1$. Let Ξ_d and $\widehat{\Xi}_d$ denote the diagonal matrices with entries $(\xi_j)_{j=1}^d$ and $(\hat{\xi}_j)_{j=1}^d$, respectively. Let $\widehat{\Lambda}_{n,d}$ denote the matrix whose (i,i')th component is $\hat{\gamma}_0\hat{\lambda}_{i,i'}$. Define the analogous infeasible estimator

$$\tilde{\Lambda}_{n,d} = \frac{1}{d} Y \Xi_d Y^\top \tag{C.30}$$

and let $\tilde{\lambda}_{i,i'}$ denote its (i,i')th component.

Lemma C.1. *Fix a constant $0 < \phi < 1$.*

(i) Suppose that Assumptions 4.1 and C.1 hold. There exist constants c and C such that if

$$\frac{\overline{\lambda}^{1-q/2}}{\underline{\lambda}} \frac{\overline{\gamma}^2}{\underline{\gamma}} \sqrt{\frac{\kappa_n}{n} \log(d/\phi)} < c \tag{C.31}$$

then the inequality

$$\max_{j\in[d]} |\hat{\xi}_j - \xi_j| \le C \frac{\overline{\lambda}^{1-q/2}}{\underline{\lambda}} \sqrt{M \frac{\overline{\gamma}}{\underline{\gamma}} \frac{\kappa_n}{n} \log(d/\phi)} \tag{C.32}$$

holds with probability greater than $1 - C\phi$.

(ii) Suppose that Assumption C.1 holds. The inequality

$$\max_{i,i'\in[n]} |\tilde{\lambda}_{i,i'} - \lambda_{i,i'}| \le C \frac{1}{d} \|\Omega_d\|_F \log(n/\phi) \tag{C.33}$$

holds with probability greater than $1 - C\phi$.

We apply Lemma C.1 to bound each of the terms in the decomposition

$$\widehat{\Lambda}_n - \Lambda_n = (\widehat{\Lambda}_{n,d} - \tilde{\Lambda}_{n,d}) - (\tilde{\Lambda}_{n,d} - \Lambda_n)\ . \tag{C.34}$$

We begin by considering the first term. Observe that

$$\begin{aligned} \max_{i,i'\in[n]} |\hat{\gamma}_0\hat{\lambda}_{i,i'} - \tilde{\lambda}_{i,i'}| &= \max_{i,i'\in[n]} |\frac{1}{d} \sum_{j=1}^d (\hat{\xi}^{(j)} - \xi^{(j)}) Y_i^{(j)} Y_{i'}^{(j)}| \\ &\le \left(\max_{j\in[d]} |\hat{\xi}_j - \xi_j| \right) \left(\max_{i,i'\in[n]} |Y_i^{(j)} Y_{i'}^{(j)}| \right)\ . \end{aligned} \tag{C.35}$$

The random variable $Y_i^{(j)}Y_{i'}^{(j)}$ has expectation smaller than $\overline{\lambda}\overline{\gamma}$ by Assumption C.1. Moreover, it has has sub-Exponential norm smaller than $(\overline{\lambda}\overline{\gamma}M)^2$ (see e.g., Lemma 2.7.6 of Vershynin 2018). Hence, it holds that

$$\max_{i,i'\in[n]} |Y_i^{(j)}Y_{i'}^{(j)}| \leq \overline{\lambda}\overline{\gamma} + 2(\overline{\lambda}\overline{\gamma}M)^2 \log(n/\phi) \tag{C.36}$$

with probability greater than $1-\phi$. Thus, Lemma C.1, Part (i), and the condition $\varphi_{n,d} < c$ (allowing the constant c to depend on the terms (C.29) and M) imply that

$$\max_{i,i'\in[n]} |\hat{\gamma}_0\hat{\lambda}_{i,i'} - \tilde{\lambda}_{i,i'}| \leq C\left(\frac{\overline{\lambda}^{2-q/2}\overline{\gamma}^{3/2}}{\underline{\lambda}\underline{\gamma}^{1/2}} + \frac{\overline{\lambda}^{3-q/2}\overline{\gamma}^{5/2}}{\underline{\lambda}\underline{\gamma}^{1/2}}\right) M^{5/2}\sqrt{\frac{\kappa_n}{n}}\log^{3/2}(dn/\phi)\,. \tag{C.37}$$

with probability greater than $1-C\phi$. To control the second term in the decomposition (C.34), we can apply Lemma C.1, Part (ii), directly. Putting the pieces together, we find that

$$\begin{aligned}&\max_{i,i'\in[n]} |\hat{\gamma}_0\hat{\lambda}_{i,i'} - \lambda_{i,i'}| \\ &\leq C\left(\frac{\overline{\lambda}^{2-q/2}\overline{\gamma}^{3/2}}{\underline{\lambda}\underline{\gamma}^{1/2}} + \frac{\overline{\lambda}^{3-q/2}\overline{\gamma}^{5/2}}{\underline{\lambda}\underline{\gamma}^{1/2}}\right) M^{5/2}\left(\frac{1}{d}\|\Omega_d\|_{\mathrm{F}} + \sqrt{\frac{\kappa_n}{n}}\right)\log^{3/2}(dn/\phi)\,.\end{aligned} \tag{C.38}$$

with probability greater than $1-C\phi$. Absorbing the terms (C.29) and M into C gives the desired result. ■

C.4 Proof of Theorem 4.2

Recall the definition of the matrix Frobenius norm $\|\cdot\|_{\mathrm{F}}$. Observe that

$$v^\star = \frac{\lambda^2}{d^2}\sum_{j=1}^{d}\eta_j^2 = \frac{\lambda^2}{d^2}\|\Omega_d\|_{\mathrm{F}}^2. \tag{C.39}$$

We are interested in establishing the bound

$$\begin{aligned}&\left|\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{i'<i}\mathbb{I}\{\hat{\gamma}_0\hat{\lambda}_{i,i'}\leq\delta\} - \Phi\left(\frac{d}{\lambda}\frac{1}{\|\Omega_d\|_{\mathrm{F}}}\delta\right)\right| \\ &\qquad\leq C\left(\frac{\sum_{j=1}^d\eta_j^3}{(\sum_{j=1}^d\eta_j^2)^{3/2}} + \sqrt{\frac{\log^3(dn/\phi)}{n}}\right).\end{aligned} \tag{C.40}$$

To ease exposition, we provide the details for the proof of the upper bound encoded in (C.40). The analogous lower bound will follow from a similar argument. Recall the definition of the infeasible estimator $\tilde{\lambda}_{i,i'}$ introduced in Appendix C.3. Observe that, on the event that

$$\max_{i,i'\in[n]} |\hat{\gamma}_0\hat{\lambda}_{i,i'} - \tilde{\lambda}_{i,i'}| \leq t\,, \tag{C.41}$$

we have that

$$\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{i'<i}\mathbb{I}\{\hat{\gamma}_0\hat{\lambda}_{i,i'}\leq\delta\} - \Phi\left(\frac{d}{\lambda}\frac{1}{\|\Omega_d\|_{\mathrm{F}}}\delta\right)$$

$$\leq \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{i'<i} \mathbb{I}\{\tilde{\lambda}_{i,i'} \leq \delta + t\} - \Phi\left(\frac{d}{\lambda} \frac{1}{\|\Omega_d\|_{\mathrm{F}}} \delta + t\right)$$
$$+ \Phi\left(\frac{d}{\lambda} \frac{1}{\|\Omega_d\|_{\mathrm{F}}} \delta + t\right) - \Phi\left(\frac{d}{\lambda} \frac{1}{\|\Omega_d\|_{\mathrm{F}}} \delta\right)$$
$$\leq \left(\frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{i'<i} \mathbb{I}\{\tilde{\lambda}_{i,i'} \leq \delta + t\} - \Phi\left(\frac{d}{\lambda} \frac{1}{\|\Omega_d\|_{\mathrm{F}}} \delta + t\right)\right) + t\,, \tag{C.42}$$

where the second inequality follows from the fact that the Gaussian cumulative distribution function has Lipschitz constant less than one. Thus, the result will follow by bounding the first term in (C.42) and choosing a suitable value of t such that the the event (C.41) holds with high probability.

To this end, we obtain a bound for the term (C.42) from the following Lemma.

Lemma C.2. *Suppose that Assumption 3.1 holds, that $\Lambda_n = \lambda I_n$ for some constant λ, and that the data $Y_i^{(j)}$ are Gaussian. Let $\eta_1, \ldots, \eta_d$ denote the eigenvalues of Ω_d. Fix a constant $0 < \phi < 1$. It holds that*

$$\left| \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{i'<i} \mathbb{I}\{\tilde{\lambda}_{i,i'} \leq \delta\} - \Phi\left(\frac{d}{\lambda} \frac{1}{\|\Omega_d\|_F} \delta\right) \right| \leq C \left(\frac{\sum_{j=1}^{d} \eta_j^3}{(\sum_{j=1}^{d} \eta_j^2)^{3/2}} + \frac{\log(1/\phi)}{\sqrt{n}} \right) \tag{C.43}$$

with probability greater than $1 - \phi$.

Moreover, observe that as $\Lambda_n = \lambda I_n$ and the sub-Gaussian norm of a standard Gaussian random variable is less than 2, the inequality (C.37) implies that, by setting

$$t = C \left(\frac{\lambda^{1-q/2} \overline{\gamma}^{3/2}}{\underline{\gamma}^{1/2}} + \frac{\lambda^{2-q/2} \overline{\gamma}^{5/2}}{\underline{\gamma}^{1/2}} \right) \sqrt{\frac{\log^3(dn/\phi)}{n}}\,, \tag{C.44}$$

the event (C.41) holds with probability greater than $1 - C\phi$. Hence, the decomposition (C.42) and Lemma C.2 imply that the bound (C.40) holds with probability greater than $1 - C\phi$, where we have subsumed λ and $\overline{\gamma}$ into the constant C. ■

Appendix D. Proofs for Auxiliary Results

D.1 Proof of Lemma C.1. Part (i)

The result follows from an application of the Hanson and Wright (1971) inequality for sub-Gaussian quadratic forms. See e.g., Rudelson and Vershynin (2013) for a modern treatment.

Lemma D.1 (Theorem 1.1, Rudelson and Vershynin 2013)**.** *Let $X = (X_1, \ldots, X_m)$ be a random vector with independent, centered, and M-sub-Gaussian components. Let A be an $m \times m$ deterministic matrix. There exists a constant $C > 0$ such that*

$$P\left\{ |X^\top A X - \mathbb{E}\left[X^\top A X\right]| \geq t \right\} \leq 2 \exp\left(-C \min\left\{ \frac{t^2}{M^4 \|A\|_F^2}, \frac{t}{M^2 \|A\|_{op}} \right\} \right) \tag{D.1}$$

for every $t \geq 0$.

In particular, we apply Lemma D.1 to the statistic $\hat{\gamma}_j$. To this end, collect the observations of the jth outcome into the vector $Y^{(j)} = (Y_i^{(j)})_{i=1}^n$. Recall the matrix Z defined in Assumption C.1. Observe that the matrix

$$\tilde{Z} = Z\Gamma_d^{1/2} \tag{D.2}$$

has independent rows and has components whose sub-Gaussian norm is at most $M\overline{\gamma}$. Let $\tilde{Z}^{(j)}$ denote its jth column. Observe that we can write

$$\hat{\gamma}_j = \frac{1}{n}(Y^{(j)})^\top Y^{(j)} = \frac{1}{n}(\tilde{Z}^{(j)})^\top \Lambda_n \tilde{Z}^{(j)} \,, \tag{D.3}$$

by Assumption C.1. Moreover, we can evaluate

$$\mathbb{E}[\hat{\gamma}_j] = \gamma_{j,j} L_n \,, \quad \text{where} \quad L_n = \frac{1}{n}\sum_{i=1}^n \lambda_{i,i} \,. \tag{D.4}$$

Thus, Lemma D.1 implies that

$$P\left\{|\hat{\gamma}_j - \gamma_{j,j}L_n| \geq t\right\} \leq 2\exp\left(-C\min\left\{\frac{t^2}{\overline{\gamma}^4 M^4 \|\frac{1}{n}\Lambda_n\|_{\mathrm{F}}^2}, \frac{t}{\overline{\gamma}^2 M^2 \|\frac{1}{n}\Lambda_n\|_{\mathrm{op}}}\right\}\right) \tag{D.5}$$

for all $t \geq 0$. Now, observe that Assumption 4.1 implies that

$$\|\Lambda_n\|_{\mathrm{op}} \leq \max_{i\in[n]} \frac{1}{n}\sum_{i'=1}^n |\lambda_{i,i'}| \leq \frac{\overline{\lambda}^{1-q}}{n}\kappa_n \quad \text{and} \tag{D.6}$$

$$\|\Lambda_n\|_{\mathrm{F}}^2 \leq \frac{1}{n^2}\sum_{i=1}^n\sum_{i'=1}^n \lambda_{i,i'}^2 \leq \overline{\lambda}^{2-q}\frac{1}{n^2}\sum_{i=1}^n\sum_{i'=1}^n |\lambda_{i,i'}|^q \leq \frac{\overline{\lambda}^{2-q}}{n}\kappa_n, \tag{D.7}$$

respectively. Consequently, the inequalities (D.5), (D.6), and (D.7) imply that

$$P\left\{|\hat{\gamma}_j - \gamma_{j,j}L_n| \geq t\right\} \leq 2\exp\left(-\frac{C}{\overline{\gamma}^4 M^4 \overline{\lambda}^{2-q}}\frac{t^2 n}{\kappa_n}\right) \tag{D.8}$$

for all $t \leq \overline{\gamma}^2 M^2 \overline{\lambda}$. Thus, so long as $n^{-1}\kappa_n \log(d/\phi) < 1$, by choosing

$$t = c\overline{\lambda}^{1-q/2}\overline{\gamma}^2 M^2 \sqrt{\frac{\kappa_n}{n}\log(d/\phi)} \tag{D.9}$$

for a sufficiently small constant c, the inequality (D.8) implies that

$$\max_{j\in[d]} \left|\hat{\gamma}_j - \gamma_{j,j}L_n\right| \leq c\overline{\lambda}^{1-q/2}\overline{\gamma}^2 M^2 \sqrt{\frac{\kappa_n}{n}\log(d/\phi)} \tag{D.10}$$

with probability $1-\phi$, by a union bound.

To conclude the proof, we translate our large deviation bound for $\hat{\gamma}_j$ into a large deviation bound for $\hat{\gamma}_j$. To this end, observe that, on the event

$$\max_{j\in[d]}\left\{\left|\frac{\hat{\gamma}_j}{\gamma_{j,j}L_n} - 1\right|\right\} \leq \frac{1}{2} \tag{D.11}$$

it holds that

$$\begin{aligned}\left|\frac{\hat{\gamma}_1}{\hat{\gamma}_j}-\frac{1}{\gamma_{j,j}}\right| &\leq \frac{L_n}{\hat{\gamma}_j}\left|\frac{\hat{\gamma}_1}{L_n}-1\right|+\frac{1}{\gamma_{j,j}}\left|\frac{\gamma_{j,j}L_n}{\hat{\gamma}_j}-1\right| \\ &\leq 2\left(\frac{L_n}{\hat{\gamma}_j}\left|\frac{\hat{\gamma}_1}{L_n}-1\right|+\frac{1}{\gamma_{j,j}}\left|\frac{\hat{\gamma}_j}{\gamma_{j,j}L_n}-1\right|\right) \\ &\leq 2\left(\left(\frac{1}{\gamma_{j,j}}+\left|\frac{1}{\gamma_{j,j}}-\frac{L_n}{\hat{\gamma}_j}\right|\right)\left|\frac{\hat{\gamma}_1}{L_n}-1\right|+\frac{1}{\gamma_{j,j}}\left|\frac{\hat{\gamma}_j}{\gamma_{j,j}L_n}-1\right|\right) \\ &\leq 2\left(\left(\frac{1}{\gamma_{j,j}}+\frac{1}{\gamma_{j,j}}\left|1-\frac{\gamma_{j,j}L_n}{\hat{\gamma}_j}\right|\right)\left|\frac{\hat{\gamma}_1}{L_n}-1\right|+\frac{1}{\gamma_{j,j}}\left|\frac{\hat{\gamma}_j}{\gamma_{j,j}L_n}-1\right|\right) \\ &\leq 4\left(\left(\frac{1}{\gamma_{j,j}}+\frac{1}{\gamma_{j,j}}\left|1-\frac{\hat{\gamma}_j}{\gamma_{j,j}L_n}\right|\right)\left|\frac{\hat{\gamma}_1}{L_n}-1\right|+\frac{1}{\gamma_{j,j}}\left|\frac{\hat{\gamma}_j}{\gamma_{j,j}L_n}-1\right|\right) \\ &\leq \frac{4}{\gamma_{j,j}}\left(\left|\frac{\hat{\gamma}_1}{L_n}-1\right|+\left|1-\frac{\hat{\gamma}_j}{\gamma_{j,j}L_n}\right|+\left|\frac{\hat{\gamma}_1}{L_n}-1\right|\left|1-\frac{\hat{\gamma}_j}{\gamma_{j,j}L_n}\right|\right)\,, \end{aligned} \tag{D.12}$$

where we have applied the elementary inequality

$$|z^{-1}-1| \leq 2|z-1|\,, \quad \text{for} \quad |z-1| \leq 1/2\,, \tag{D.13}$$

to obtain the second and fifth inequalities. Now, observe that

$$\begin{aligned}\left|\frac{\hat{\gamma}_j}{\gamma_{j,j}L_n}-1\right| &\leq c\frac{\overline{\lambda}^{1-q/2}\overline{\gamma}^2}{\gamma_{j,j}L_n}\sqrt{M\frac{\kappa_n}{n}\log(d\phi)} \\ &\leq c\frac{\overline{\lambda}^{1-q/2}}{\underline{\lambda}}\frac{\overline{\gamma}^2}{\underline{\gamma}}\sqrt{M\frac{\kappa_n}{n}\log(d/\phi)} \end{aligned} \tag{D.14}$$

with probability greater than $1-\phi/d$, by the inequality (D.10). The condition (C.31) implies that

$$c\frac{\overline{\lambda}^{1-q/2}}{\underline{\lambda}}\frac{\overline{\gamma}^2}{\underline{\gamma}}\sqrt{M\frac{\kappa_n}{n}\log(d/\phi)} \leq \frac{1}{2} \tag{D.15}$$

with probability greater than $1-\phi/d$. Consequently, the union bound implies that the event (D.11) holds for every j in $[d]$ with probability greater than $1-\phi$. Thus, the inequality (D.12) implies that

$$\max_{j\in[d]}|\hat{\xi}_j-\xi_j| \leq \max_{j\in[d]}\left\{C\gamma_{j,j}\left(\left|\frac{\hat{\gamma}_1}{L_n}-1\right|+\left|1-\frac{\hat{\gamma}_j}{\gamma_{j,j}L_n}\right|+\left|\frac{\hat{\gamma}_1}{L_n}-1\right|\left|1-\frac{\hat{\gamma}_j}{\gamma_{j,j}L_n}\right|\right)\right\}\,, \tag{D.16}$$

with probability greater than $1-\phi$. Appealing, again, to the fact that the event (D.11) holds for every j in $[d]$ with probability greater than $1-\phi$, and applying the bound (D.14), we can conclude that the inequality

$$\max_{j\in[d]}|\hat{\xi}_j-\xi_j| \leq C\frac{\overline{\lambda}^{1-q/2}}{\underline{\lambda}}\sqrt{M\frac{\overline{\gamma}}{\underline{\gamma}}\frac{\kappa_n}{n}\log(d/\phi)} \tag{D.17}$$

holds with probability greater than $1-C\phi$, as desired. ■

D.2 Proof of Lemma C.1. Part (ii)

The result again follows from an application of Lemma D.1. Recall the matrix Z defined in Assumption C.1. Observe that the matrix

$$\breve{Z} = \Lambda_n^{1/2} Z \tag{D.18}$$

has independent columns and has components whose sub-Gaussian norm is at most $M\overline{\lambda}$. Let $\breve{Z}_i$ denote its ith column. By Assumption C.1, we have that

$$\tilde{\lambda}_{i,i'} = \frac{1}{d} Y_i \Xi_d Y_i^\top = \frac{1}{d} \breve{Z}_i \Lambda_d^{1/2} \Xi_d \Lambda_d^{1/2} \breve{Z}_i^\top \ . \tag{D.19}$$

Thus, Lemma D.1 and a union bound imply that

$$P\left\{ \max_{i,i' \in [n]} |\tilde{\lambda}_{i,i'} - \lambda_{i,i'}| \geq t \right\} \leq 2n^2 \exp\left(-Ctd \min\left\{ \frac{td}{\|\Omega_d\|_{\mathrm{F}}^2}, \frac{1}{\|\Omega_d\|_{\mathrm{op}}} \right\} \right) \ . \tag{D.20}$$

Re-expressing this bound, we find that the inequality

$$\begin{aligned} \max_{i,i' \in [n]} |\tilde{\lambda}_{i,i'} - \lambda_{i,i'}| &\leq C \frac{1}{d} \left(\|\Omega_d\|_{\mathrm{F}} \sqrt{\log(n/\phi)} + \|\Omega_d\|_{\mathrm{op}} \log(n/\phi) \right) \\ &\leq C \frac{1}{d} \|\Omega_d\|_{\mathrm{F}} \log(n/\phi) \end{aligned} \tag{D.21}$$

holds with probability greater than $1 - C\phi$, as desired. ■

D.3 Proof of Lemma C.2

Throughout, we let the operator $\sim$ denote equality in distribution. The result relies on the following distributional decomposition of Gaussian quadratic forms. Let $\mathcal{W}_q(\Sigma, p)$ denote the q-variate Wishart distribution, with parameter Σ and degrees of freedom p.

Lemma D.2 (Corollary 2.3, Singull and Koski (2012))**.** *Fix an $n \times n$ positive semi-definite matrix $\Lambda_n = (\lambda_{i,i'})_{i,i'=1}^n$ and a $d \times d$ full-rank matrix $\Gamma_d = (\gamma_{j,j'})_{j,j'=1}^d$. Let X denote a mean-zero, Gaussian random matrix, whose (i,j)th component is given by $X_i^{(j)}$ and that satisfies*

$$\mathrm{Cov}(X_i^{(j)}, X_{i'}^{(j')}) = \lambda_{i,i'} \gamma_{j,j'} \ . \tag{D.22}$$

Let A denote a deterministic, symmetric, full-rank $d \times d$ matrix. The statistic $Q = XAX^\top$ has the distribution

$$Q \sim \sum_{j=1}^d \eta_j S_j \ , \tag{D.23}$$

where $\eta_1, \ldots, \eta_d$ are the eigenvalues of $\Gamma_d^{1/2} A \Gamma_d^{1/2}$ and $S_1, \ldots, S_d$ are independent random matrices with distribution $\mathcal{W}_d(\Lambda_n, 1)$.

As the data are Gaussian, and satisfy Assumption 3.1, Lemma D.2 implies that

$$\tilde{\Lambda}_{n,d} = \frac{1}{d} Y \Xi_d Y^\top \sim \frac{1}{d} \sum_{j=1}^{d} \eta_j S_j \,, \tag{D.24}$$

where $\eta_1, \ldots, \eta_d$ are the eigenvalues of $\Omega_d = \Gamma_d^{1/2} \Xi_d \Gamma_d^{1/2}$ and $S_1, \ldots, S_d$ are independent random matrices with distribution $\mathcal{W}_d(\Lambda_n, 1)$.

Let Z denote a mean-zero, Gaussian random matrix whose components are mutually independent. Let $Z^{(j)}$ denote the jth column of Z and set $\mathrm{Var}(Z^{(j)}) = \lambda I_n$. Likewise, let $Z_i^{(j)}$ and Z_i denote the (i,j)th component and ith row of Z, respectively. By the definition of the Wishart distribution, it holds that

$$S_j \sim Z^{(j)} (Z^{(j)})^\top \tag{D.25}$$

for each j in $[d]$. Thus, the characterization (D.24) implies that

$$\tilde{\lambda}_{i,i'} \sim \frac{1}{d} \sum_{j=1}^{d} \eta_j Z_i^{(j)} Z_{i'}^{(j)} \tag{D.26}$$

for each i, i' in $[n]$. Therefore, the statistic

$$\frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{i'<i} \left(\mathbb{I}\{\tilde{\lambda}_{i,i'} \le \delta\} - \Phi\left(\frac{d}{\lambda} \frac{1}{\|\Omega_d\|_{\mathrm{F}}} \delta \right) \right) \tag{D.27}$$

is equi-distributed with

$$\frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{i'<i} \psi_\delta(Z_i, Z_{i'}) \tag{D.28}$$

where

$$\psi_\delta(Z_i, Z_{i'}) = \mathbb{I}\left\{ \frac{1}{d} \sum_{j=1}^{d} \eta_j Z_i^{(j)} Z_{i'}^{(j)} \le \delta \right\} - \Phi\left(\frac{d}{\lambda} \frac{1}{\|\Omega_d\|_{\mathrm{F}}} \delta \right) \,. \tag{D.29}$$

The random variable (D.28) can be recognized as a U-statistic of order 2.

Thus, we apply the following large-deviation bound, due to Hoeffding (1963).

Lemma D.3 (Lemma A.5, Song et al. (2019))**.** *Let $X_1, \ldots, X_n$ denote a collection of independent and identically distributed random variables. Suppose that $\psi(\cdot, \cdot)$ is a real-valued function that is symmetric in its two components and satisfies*

$$\mathbb{E}\left[\psi(X_1, X_2)\right] = 0 \,, \quad \mathrm{Var}(\psi(X_1, X_2)) = \nu \,, \quad \textit{and} \quad |\psi(X_1, X_2)| \le \theta \tag{D.30}$$

almost surely. Fix a constant $0 < \phi < 1$. The inequality

$$\left| \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{i'<i} \psi(X_i, X_{i'}) \right| \le C \left(\sqrt{\frac{\nu \log(1/\phi)}{n}} + \frac{\theta \log(1/\phi)}{n} \right) \tag{D.31}$$

holds with probability greater than $1 - \phi$.

Observe that, as $\Lambda_n = \lambda I_n$ and the components of Z are independent, the random variables $Z_1, \ldots, Z_n$ are independent and identically distributed. Moreover, we have that

$$\mathrm{Var}(\psi_\delta(Z_i, Z_{i'})) \leq 1 \quad \text{and} \quad |\psi_\delta(Z_i, Z_{i'}) - \mathbb{E}\left[\psi_\delta(Z_i, Z_{i'})\right]| \leq 1 \tag{D.32}$$

almost surely. Thus, Lemma D.3 implies that the inequality

$$\left| \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{i'<i} (\psi_\delta(Z_i, Z_{i'}) - \mathbb{E}\left[\psi_\delta(Z_i, Z_{i'})\right]) \right| \leq C \frac{\log(1/\phi)}{\sqrt{n}} \tag{D.33}$$

holds with probability greater than $1 - \phi$.

Hence, it suffices to bound the expectation

$$\mathbb{E}\left[\psi_\delta(Z_i, Z_{i'})\right] = P\left\{ \frac{1}{d} \sum_{j=1}^{d} \eta_j Z_i^{(j)} Z_{i'}^{(j)} \leq \delta \right\} - \Phi\left(\frac{d}{\lambda} \frac{1}{\|\Omega_d\|_{\mathrm{F}}} \delta \right) . \tag{D.34}$$

To do this, we apply the following Berry-Esseen inequality for non-identically distributed sums

Lemma D.4 (Theorem XVI.5.2, Feller (1991))**.** *Let $X_1, \ldots, X_n$ be independent variables such that*

$$\mathbb{E}\left[X_k\right] = 0, \quad \mathbb{E}\left[|X_k|^2\right] = \sigma_k^2, \quad \textit{and} \quad \mathbb{E}\left[|X_k|^3\right] = \theta^k . \tag{D.35}$$

Define the sequences

$$s_n^2 = \sum_{i=1}^{n} \sigma_i^2 \quad \textit{and} \quad r_n = \sum_{i=1}^{n} \theta_i \tag{D.36}$$

For all δ and n, the inequality

$$\left| P\left\{ \frac{1}{s_n} \sum_{i=1}^{n} X_k \leq \delta \right\} - \Phi\left(\delta\right) \right| \leq C \frac{r_n}{s_n^3} \tag{D.37}$$

holds.

Consider the sum

$$\sum_{j=1}^{d} \frac{1}{d} \eta_j Z_i^{(j)} Z_{i'}^{(j)} . \tag{D.38}$$

Observe that

$$\begin{aligned} \mathbb{E}\left[\frac{1}{d} \eta_j Z_i^{(j)} Z_{i'}^{(j)} \right] &= 0 , \\ \mathbb{E}\left[|\frac{1}{d} \eta_j Z_i^{(j)} Z_{i'}^{(j)}|^2 \right] &= \frac{\eta_j^2}{d^2} \lambda^2 , \quad \text{and} \\ \mathbb{E}\left[|\frac{1}{d} \eta_j Z_i^{(j)} Z_{i'}^{(j)}|^3 \right] &\geq C \frac{\eta_j^3}{d^3} \lambda^3 , \end{aligned} \tag{D.39}$$

where we have used facts about the second and third moments of Gaussian distributions for the second and third relations. Consequently, Lemma D.4 implies that

$$\begin{aligned}
&\left|P\left\{\frac{1}{d}\sum_{j=1}^{d}\eta_j Z_i^{(j)} Z_{i'}^{(j)} \leq \delta\right\} - \Phi\left(\frac{d}{\lambda}\frac{1}{\|\Omega_d\|_{\mathrm{F}}}\delta\right)\right| \\
&\leq \left|P\left\{\frac{d}{\lambda\sqrt{\sum_{j=1}^{d}\eta_j^2}}\frac{1}{d}\sum_{j=1}^{d}\eta_j Z_i^{(j)} Z_{i'}^{(j)} \leq \frac{d}{\lambda\sqrt{\sum_{j=1}^{d}\eta_j^2}}\delta\right\} - \Phi\left(\frac{d}{\lambda\sqrt{\sum_{j=1}^{d}\eta_j^2}}\delta\right)\right| \\
&\leq \frac{\sum_{j=1}^{d}\eta_j^3}{(\sum_{j=1}^{d}\eta_j^2)^{3/2}}\,. \qquad \text{(D.40)}
\end{aligned}$$

Hence, by combining the inequalities (D.33) and (D.40), we can conclude that

$$\left|\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{i'<i}\left(\mathbb{I}\{\tilde{\lambda}_{i,i'} \leq \delta\} - \Phi\left(\frac{d}{\lambda}\frac{1}{\|\Omega_d\|_{\mathrm{F}}}\delta\right)\right)\right| \leq C\left(\frac{\sum_{j=1}^{d}\eta_j^3}{(\sum_{j=1}^{d}\eta_j^2)^{3/2}} + \frac{\log(1/\phi)}{\sqrt{n}}\right) \qquad \text{(D.41)}$$

holds with probability greater than $1-\phi$, as required. ■

APPENDIX E. FURTHER DETAILS FOR EMPIRICAL APPLICATIONS

E.1 Bazzi et al. (2023)

Bazzi et al. (2023) examine the impact of the migration of millions of Southern whites in the twentieth century. They use cross-sectional data concerning the demographic and political characteristics of 1,888 U.S. counties. The main finding is that the Southern white migration in the early twentieth century is associated with significantly higher Republican vote shares in the twenty-first century. Bazzi et al. (2023) consider the model

$$\text{Vote}_c = \beta \cdot \text{Southern Whites}_{c,1940} + X_c + \alpha_s + \varepsilon_c\,, \qquad \text{(E.1)}$$

where c indexes counties and s indexes states. The main outcome, Vote_c, is the vote share received by Donald Trump in 2016. The regressor of interest, $\text{Southern Whites}_{c,1940}$, is the Southern-born white population share in 1940. The regression is instrumented by a shift-share variable constructed with predetermined Southern white migration networks as of 1900 and predicted aggregate migration flows out of the South for each decade from 1900 to 1940. The control variables X_c are historical economic factors (population density, manufacturing employment, average farm values), ideological factors (Union Army enlistment, mortality rates from the U.S. Civil War), and the vote share for Woodrow Wilson in 1912. The variable α_s denotes a state fixed effect. Standard errors are clustered across counties in 60×60 mile grid cells.

The main result are displayed in their Table 2, Panel A, Column (4).[22] The authors implement several adjustments for spatial correlation. They report alternative standard errors, constructed with the methods

[22]The authors do not explicitly indicate the preferred specification. However, on page 1591, they write "Our main estimating equation takes the following form." Thus, we infer that Table 2 displays their "main" results. On Page 1595, they write "Secondary specifications control for additional potential sorting correlates," and so columns 5 and 7 are not the main specifications. Therefore, we use Panel A column 4 as the main specification.

proposed by Conley (1999), Colella et al. (2023), and Adao et al. (2019), as well as through a wild bootstrap with clustering by state. The corrections of the main specification are shown in their Column (4) of Table A.3. The standard error based on the Conley (1999), with a cutoff at 500 km, is the largest and is 68% larger than the baseline result.

To construct a collection of outcomes for the application of the TMO variance estimator, we collect all feasible variables available in the replication package. We exclude outcomes that are very similar to the main outcome, the regressor of interest, and the primary controls variables. We also exclude outcomes that are missing for over 50% of observations or are collinear with right hand side of the model.

E.2 Bernini et al. (2023)

Bernini et al. (2023) study the impact of the Voting Rights Act (VRA) on the racial makeup of local governments in the U.S. South. They consider cross-sectional data consisting of 971 counties in 11 states. They find that federal scrutiny over Southern states post VRA had a sizable impact on the extent to which enfranchisement led to Black office holding. Bernini et al. (2023) consider the long differences regression model

$$\begin{aligned}\Delta\text{Share Black Elected}_{cs} = \gamma\text{Percent Black}_{1960} + \theta\text{Percent Black}_{1960} * \text{Covered}_{cs} \\ + X_{cs} + X_{cs} * \text{Covered}_{cs} + I_s + \epsilon_{cs} \ ,\end{aligned}$$

where c denote county and s denotes state. The main outcome, ΔShare Black Elected$_{cs}$, is the change in the share of Black elected officials between 1964 and 1980. The regressor of interest, Percent Black$_{1960}$ $*$ Covered$_{cs}$, is the interaction of the Black share in 1960 and a dummy for federal scrutiny. Control variables X_{cs} are pre-VRA county characteristics (unemployment rate, the percent of families below the poverty line, the percent of unskilled workers, agricultural productivity, population (ln), percent urban population, cotton share, pro- and anti-Black activism, Republican share), I_s is state dummies. Standard errors are clustered by judicial division. Results are shown in Table 2 Column 4, which is stated as the paper's preferred specification[23].

To construct a collection of outcomes for the application of the TMO variance estimator, we collect all feasible variables available in the replication package. We exclude outcomes that are very similar to the main outcome, the regressor of interest, and the primary controls variables. We also exclude outcomes that are missing for over 50% of observations or are collinear with right hand side of the model.

E.3 Caprettini and Voth (2023)

Caprettini and Voth (2023) study the complementarity between patriotism in World War II and public-good provision in the New Deal. They consider cross-sectional data considering 2329 counties. They find that higher government spending at the county level in the 1930s is positively correlated with more patriotic

[23] On page 1019-1020, Bernini et al. (2023) write "Our preferred specification in column 4 of table 2"

actions during World War II. Caprettini and Voth (2023) consider the OLS regression model

$$\text{WWII Patriotism}_i = \beta \text{New Deal Grants}_i + \gamma WWIPatriotism_i + \delta X_i + \xi_s + u_i \ ,$$

where i denotes county. The main outcome, WWII Patriotism$_i$ denotes the average purchases of war bonds per capita in 1944. The regressor of interest, New Deal Grants$_i$ denotes the New Deal grants per 1930 population. Control variables WWI Patriotism$_i$ and X_i are WWI volunteering rate, WWI medals per 1000 inhabitants, socioeconomic characteristics in 1930 (log of population, unemployment rate, share of veterans, share of African Americans, and share of people born in major Axis countries), share of people living on a farm in 1930, 1929 farm income, an urban dummy, 1898-1928 average Democratic vote share, value of WWII war contracts per capita, average 1939 wage of employees, and 1930 share of men. The variables ξ_s denote state fixed effects. Heteroskedasticity-robust standard errors are used.

Results are shown in Table 2 Panel A Column 1.[24]. Caprettini and Voth (2023) discuss spatial error correlation. They report Conley (1999) standard errors in Table 2 Panel B and Table 3. The Conley (1999) standard errors are between 0% and 171% larger than the heteroskedasticity-robust standard errors.

To construct a collection of outcomes for the application of the TMO variance estimator, we collect all feasible variables available in the replication package. We exclude outcomes that are very similar to the main outcome, the regressor of interest, and the primary controls variables. We also exclude outcomes that are missing for over 50% of observations or are collinear with right hand side of the model.

E.4 Moscona and Sastry (2023)

Moscona and Sastry (2023) study how innovation shapes the economic impact of climate change. They consider panel data consisting of 3000 counties between 1960 and 2010. They find that that counties' exposure to climate-change-induced innovation significantly decreases the local economic damage from extreme temperatures. Moscona and Sastry (2023) consider the regression model

$$\begin{aligned}\text{log Agr Land Price}_{i,t} = \delta_i + \alpha_{s(i),t} + \beta \text{Extreme Exposure}_{i,t} + \gamma \text{Innovation Exposure}_{i,t} \\ + \phi(\text{Extreme Exposure}_{i,t} * \text{Innovation Exposure}_{i,t}) + \epsilon_{i,t}\end{aligned}$$

where i denotes county, s denotes state, and t denotes time (1959 or 2017). The main outcome, log Agr Land Price$_{i,t}$, is the price per acre of cultivated land. The regressor of interest, Extreme Exposure$_{i,t}$*Innovation Exposure$_{i,t}$, is the interaction of measures of a county's exposures to extreme temperature and innovation, respectively. The variable δ_i denote county fixed effects. The variables $\alpha_{s(i),t}$ denote state by year fixed effects. Standard errors are double clustered at the county and state-by-year levels. Results are shown in Table 3 Column 1

[24]On Page 481, Caprettini and Voth (2023) write "Figure I summarizes our main result". Similarly, on Page 483 they write "to go beyond the graphical evidence", suggest results in this section (Table 2) are the main results. The first three columns report results for three equally important measures of patriotism, so we randomly choose the first as the main specification

in the paper.[25] Moscona and Sastry (2023) report Conley (1999) standard errors in Table A21. The Conley (1999) standard errors are smaller than the clustered standard errors reported in the main text.

To construct a collection of outcomes for the application of the TMO variance estimator, we collect all feasible variables available in the replication package, as well as outcomes from other papers. We exclude outcomes that are very similar to the main outcome, the regressor of interest, and the primary controls variables. We also exclude outcomes that are missing for over 50% of observations, are collinear with right hand side of the model, or are measured out of the time window of 1940 to 2010.

E.5 Esposito et al. (2023)

Esposito et al. (2023) study the effect of the Lost Cause narrative—a revisionist and racist retelling of the U.S. Civil War—on national reunification and racial discrimination. They consider a panel dataset consisting of 786 counties between 1910 to 1920. They find that screenings of *The Birth of a Nation* shifted public discourse toward more patriotic and less divisive language, increased military enlistment, and fostered cultural convergence. This paper considers an instrumental variables regression model

$$\text{Reconciliation}_{ct} = \beta \text{BON}_{ct} + X_{ct} + \alpha_c + \alpha_t + \epsilon_{ct}$$

where c is county and t is year-month. The main outcome, $\text{Reconciliation}_{ct}$, represents the relative log frequencies of patriotic keywords to divisive ones in local newspapers. The regressor of interest, BON_{ct}, is equal to 1 after the screening of *The Birth of a Nation*. The regression is instrumented with the screening of an another movie, *The Million Dollar Mystery*. The control, X_{ct}, is total number of newspaper pages available in data. Standard errors are clustered at the county level. Results are shown in Table 3 Panel A Column 3.[26] We choose the specification reported in Panel A Column 3 because the authors specify that this specification is their preferred model.[27] The paper also clusters the standard errors at the state level and applies the estimator proposed in Colella et al. (2023) (see Table B34 Panel C Column 1). Clustering at the state level increases the standard error by 34%, while the correction based on Colella et al. (2023) decreases the standard error.

To construct a collection of outcomes for the application of the TMO variance estimator, we collect all feasible variables available in the replication package, as well as outcomes from other papers. We exclude outcomes that are very similar to the main outcome, the regressor of interest, and the primary controls

[25] The authors do not explicitly indicate the preferred specification. However, on page 645, they write "Sections IV and V present our main results." Thus, we conclude that Table 3 displays the paper's main county level results. Within Table 3, we pick Column 1 as the main specification, because the authors choose this specification to perform down-stream analysis. For example, on Page 680, they write "Figure VI reports the marginal effect of exposure to extreme heat (y-axis) for quantiles of the innovation exposure distribution (x-axis), using the specification from column (1)".

[26] On page 1479, the authors write "In our baseline analysis...." Likewise, on page 1479, they write "we compute our main dependent variable, $\text{Reconciliation}_{ct}$."

[27] On page 1472, Esposito et al. (2023) write "our preferred solution is to instrument the treatment in a two-stage least squares (2SLS) version of equation (1)."

variables. We also exclude outcomes that are missing for over 50% of observations, are collinear with right hand side of the model, or are measured out of the time window of 1910 to 1930.

E.6 Calderon et al. (2023)

Calderon et al. (2023) study the political effects of the migration of four million African Americans from the South to the North. They consider with a panel data consisting of 1263 non-southern counties from 1940 to 1960. Calderon et al. (2023) find that the great migration increased support for the Democratic Party, increased Congress members' propensity to promote civil rights legislation, and encouraged pro-civil rights activism outside the US South. This paper considers the instrumental variables regression model

$$\Delta \text{demsh}_{c\tau} = \delta_{s\tau} + \beta \Delta \text{Bl}_{c\tau} + \gamma X_{c\tau} + u_{c\tau} \ ,$$

where c is county and τ is decade. The main outcome, $\Delta \text{demsh}_{c\tau}$, is the change in the Democratic vote share during decade τ. The regressor of interest, $\Delta \text{Bl}_{c\tau}$, is the change in the Black population share. The regression is instrumented with a shift-share predictor of Black inflow. Control variables $X_{c\tau}$ are interactions between decade dummies and 1940 county characteristics (Black population share and Democratic incumbency in Congressional elections). The variable $\delta_{s\tau}$ collects interactions between decade and state dummies. The regression is weighed by 1940 county population. Standard errors are clustered at the county level. Results are displayed in Table 2 Column 6. The authors note that this is the paper's preferred specification [28]. Calderon et al. (2023) additionally clusters their standard errors at the commuting zone level and reports the confidence interval constructed with the method of Adao et al. (2019). These results are displayed in Table D.21 and D.22. Clustering at the commuting zone level increases the standard errors by 1%.

To construct a collection of outcomes for the application of the TMO variance estimator, we collect all feasible variables available in the replication package, as well as outcomes from other papers. We exclude outcomes that are very similar to the main outcome, the regressor of interest, and the primary controls variables. We also exclude outcomes that are missing for over 50% of observations, are collinear with right hand side of the model, or are measured out of the time window of the paper,

E.7 Cook et al. (2023)

Cook et al. (2023) study factors correlated with the provision of nondiscriminatory services. They consider a panel dataset consisting of 3092 counties between 1939 to 1955. They find that declines white population led to increases in the number of nondiscriminatory businesses. This paper considers the linear regression model

$$\text{Asinh}(N_{0ct}) = \beta_0 + \beta_1 \text{Asinh}(\text{casualties}_c) * \text{postWWII}_t + \phi_c + \xi_t + \epsilon_{ct}$$

where c denotes county and t denotes year. The main outcome, N_{0ct}, is the number of Green Book establishments in county c at time t. The regressor of interest, $\text{Asinh}(\text{casualties}_c) * \text{postWWII}_t$, is the

[28] On page 179, Calderon et al. (2023) write "especially for our preferred specification (Column 6)"

interaction of the number of white casualties in World War II and a post-war indicator. The variables ϕ_c and ξ_t denote county and year fixed effects. Standard errors are clustered at the county level. Results are shown in Table 3 Panel A Column 4 in the paper, which is stated as the paper's preferred specification.[29]

To construct a collection of outcomes for the application of the TMO variance estimator, we collect all feasible variables available in the replication package, as well as outcomes from other papers. We exclude outcomes that are very similar to the main outcome, the regressor of interest, and the primary controls variables. We also exclude outcomes that are missing for over 50% of observations, are collinear with right hand side of the model, or are measured out of the time window of 1930 to 1960.

E.8 Chetty et al. (2014)

Chetty et al. (2014) study features of inter-generational mobility in the United States. One of the findings of this paper is that the absolute upward mobility (the expected income rank of children from families at the 25 percentile of the national parent income distribution) is negatively correlated with the proportion in a commuting zone that is African American. This paper considers the linear regression model

$$\text{absmob}_c = \beta_0 + \beta_1 \text{frac black}_c + \delta_s + \epsilon_c$$

where s denotes state and c denotes commuting zone. The main outcome, absmob_c, is the measure of absolute upward mobility in commuting zone c. The regressor of interest, frac black_c is the share of African American people in commuting zone c. Standard errors are clustered at the state level. Results are shown in the first row of Figure VIII in the paper[30].

To construct a collection of outcomes for the application of the TMO variance estimator, we collect all feasible variables available in the replication packages. We exclude outcomes that are missing for over 50% of observations or are collinear with right hand side of the model.

E.9 Acemoglu et al. (2019)

Acemoglu et al. (2019) estimates the impact of democracy on economic growth. They consider panel data of 175 countries between 1960 to 2010. They find that democratization increases GDP per capita in the long run. This paper considers the linear regression model

$$\text{lgdppc}_{c,t} = \beta \text{demo}_{c,t} + \sum_{j=1}^{4} \gamma_j \text{lgdppc}_{c,t-j} + \alpha_c + \delta_t + \epsilon_{c,t}$$

where c denotes country and t denotes year. Results are shown in Table 2 Column 3, which is stated as the paper's preferred specification[31]. The main outcome, $\text{lgdppc}_{c,t}$, is the log of GDP per capita. The regressor of interest, $\text{demo}_{c,t}$, is a dichotomous measure of democracy. Control variables are the lags of log GDP per

[29] On page 77, the authors write "our preferred specification uses county and year fixed effects".

[30] We pick this specification because the authors write "Perhaps the most obvious pattern from the maps in Figure VI is that inter-generational mobility is lower in areas with larger African American populations, such as the Southeast" on Page 1605

[31] On page 59, Acemoglu et al. (2019) write "Column 3, which is our preferred specification, includes four lags of GDP per capita"

capita. The quantities δ_t and α_c are year and country fixed effects. Standard errors are clustered at the country level.

To construct a collection of outcomes for the application of the TMO variance estimator, we collect all feasible variables available in the replication packages. We exclude outcomes that are missing for over 50% of observations or are collinear with right hand side of the model.

FIGURE E.7. Distribution of Correlations Between Locations in Applications

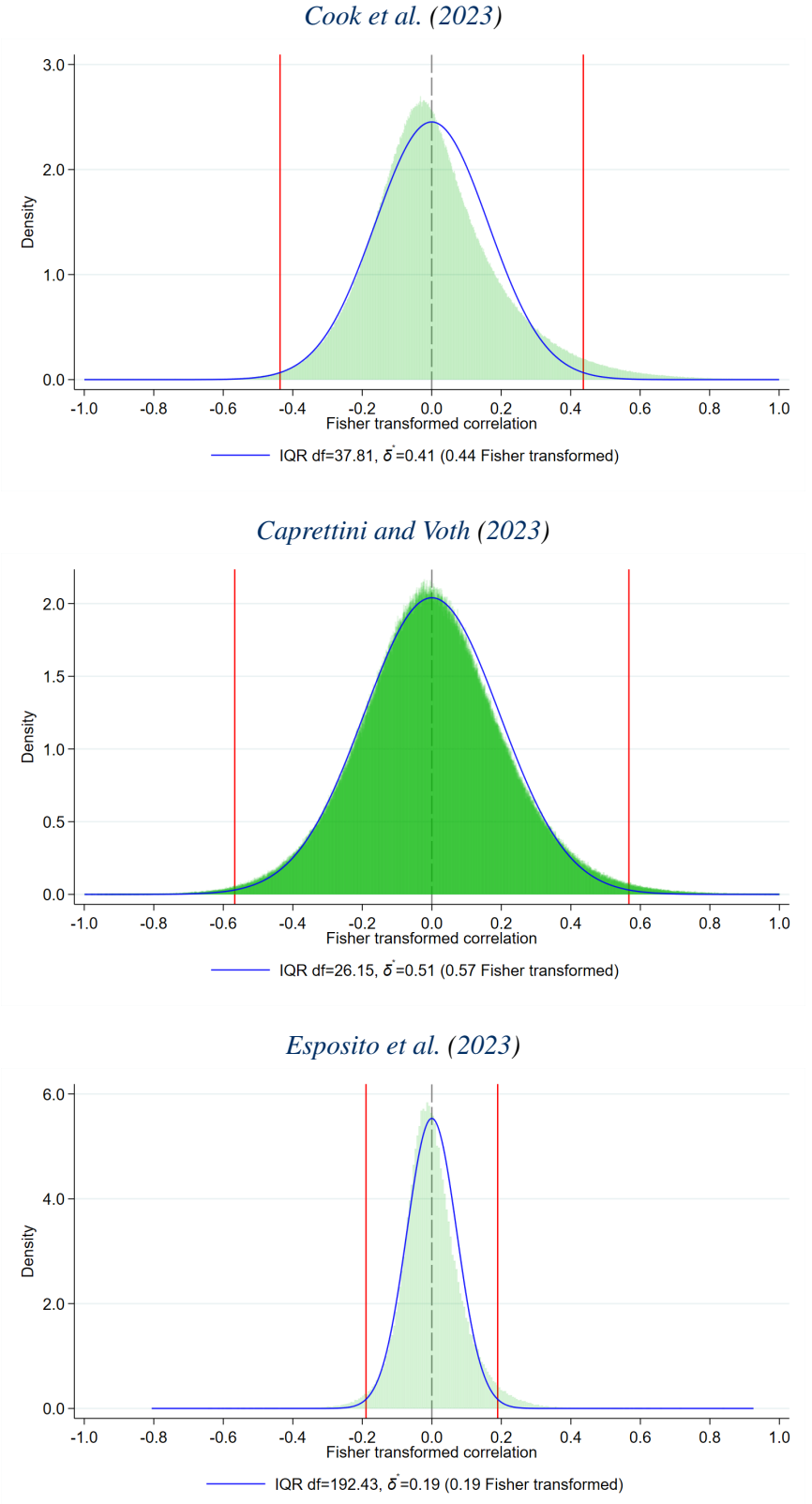
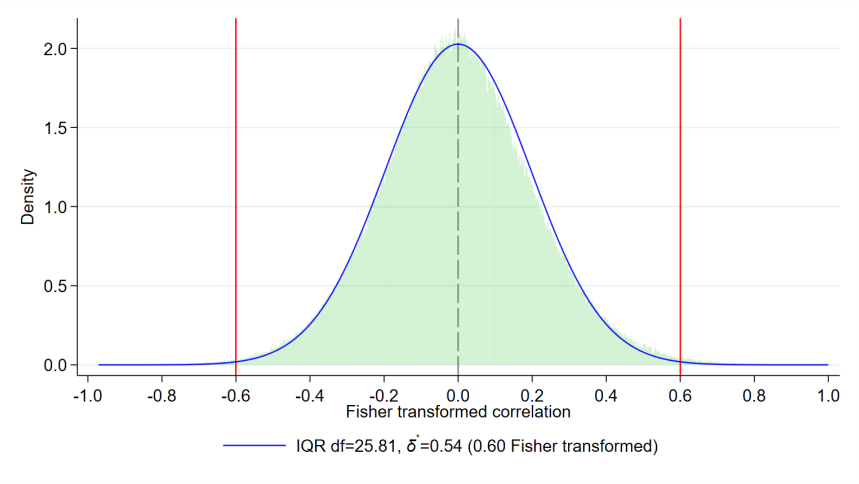
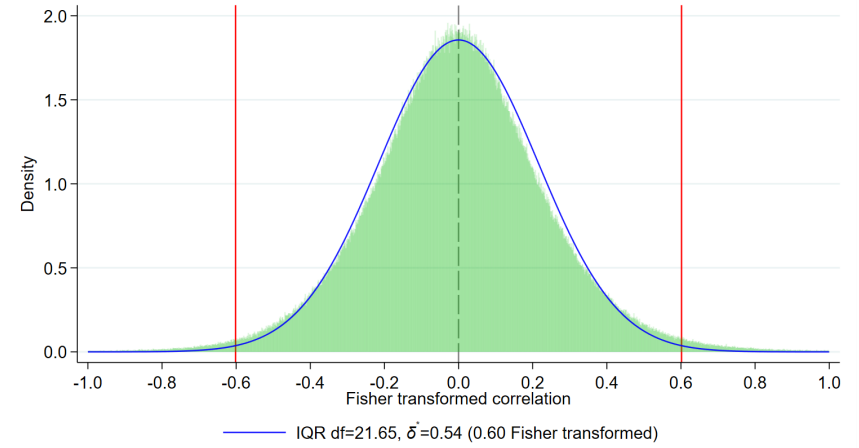


FIGURE E.7. Distribution of Correlations Between Locations in Applications

Bernini et al. (2023)



Bazzi et al. (2023)



Calderon et al. (2023)

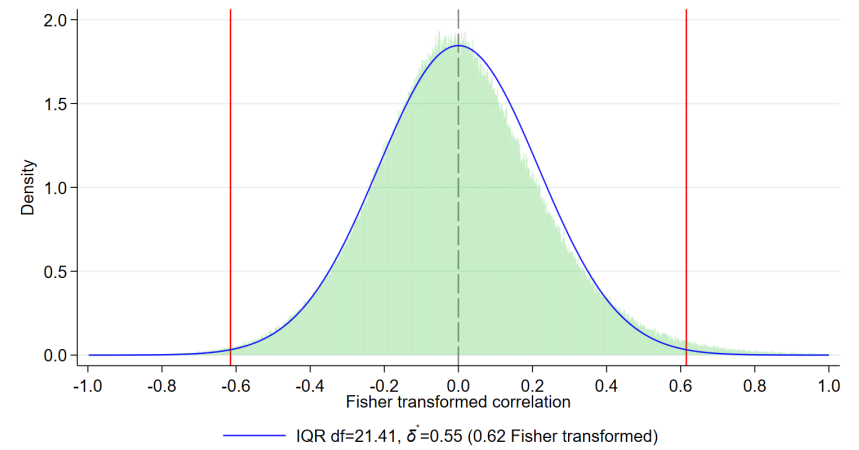
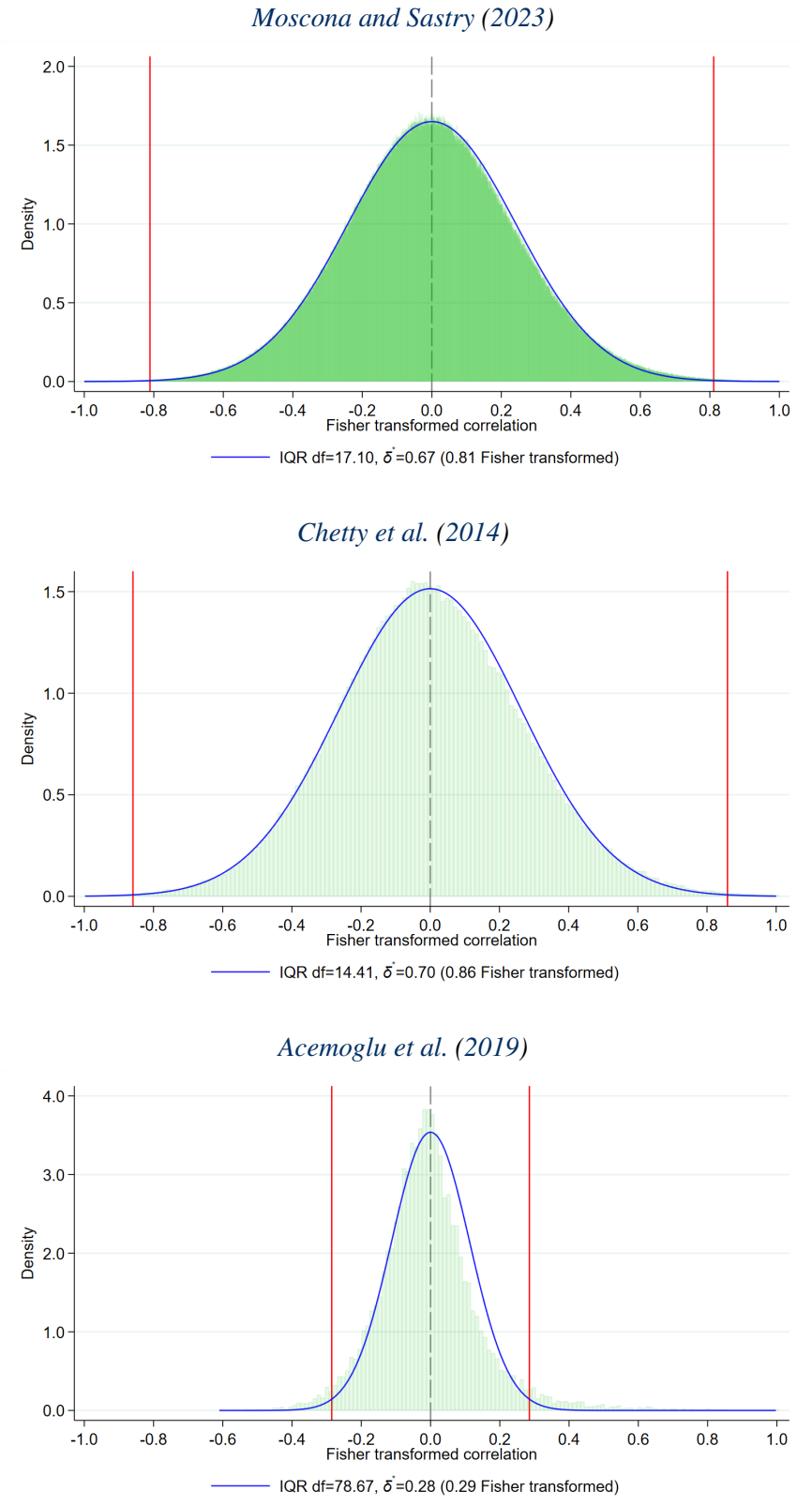


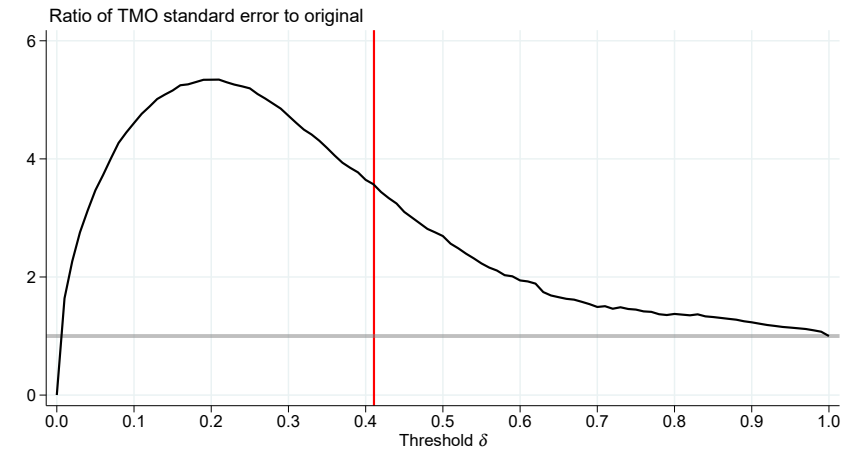
FIGURE E.7. Distribution of Correlations Between Locations in Applications



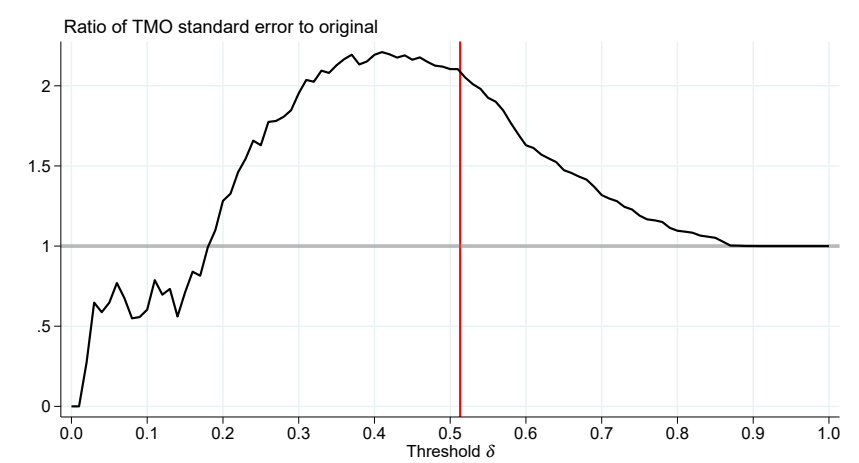
Notes: Figure E.8 displays a histogram of the Fisher-transformed correlations $\tilde{\rho}_{i,i'}$ between each pair of locations in each of the empirical applications in Table 4. The density curve denotes the null density estimate, obtained with the procedure outlined in Section 3.4. The vertical line marks the optimal threshold $\hat{\delta}^*$.

FIGURE E.8. TMO Relative to Original Standard Error across Thresholds in Applications

Cook et al. (2023)



Caprettini and Voth (2023)



Esposito et al. (2023)

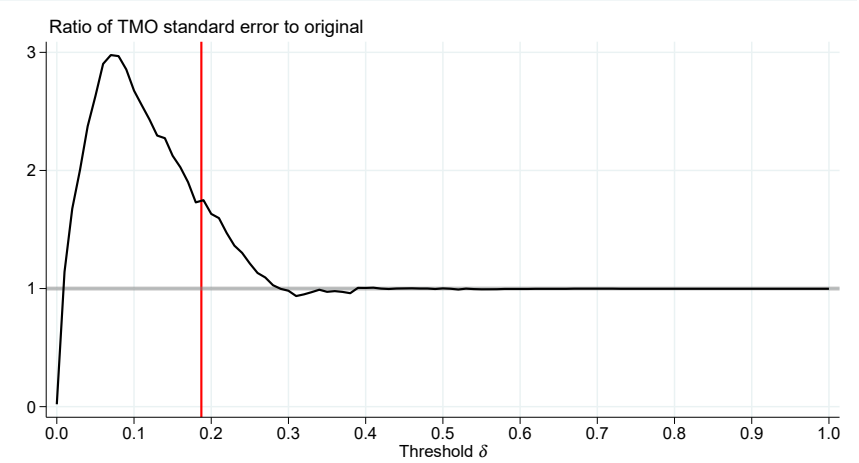
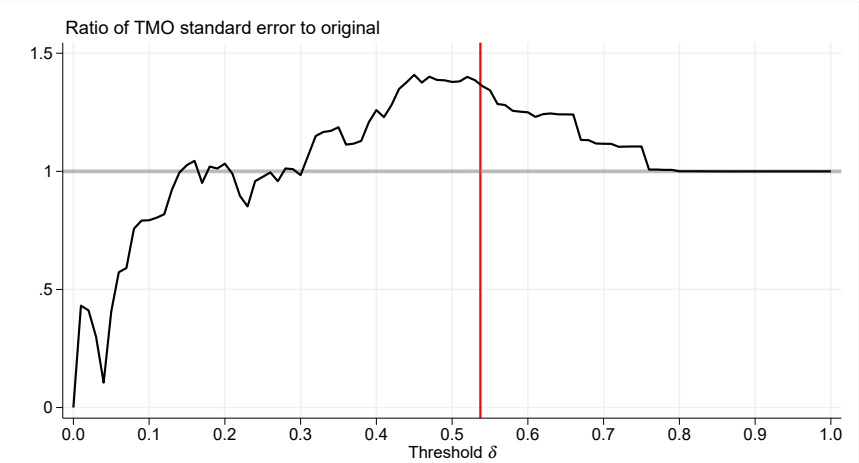
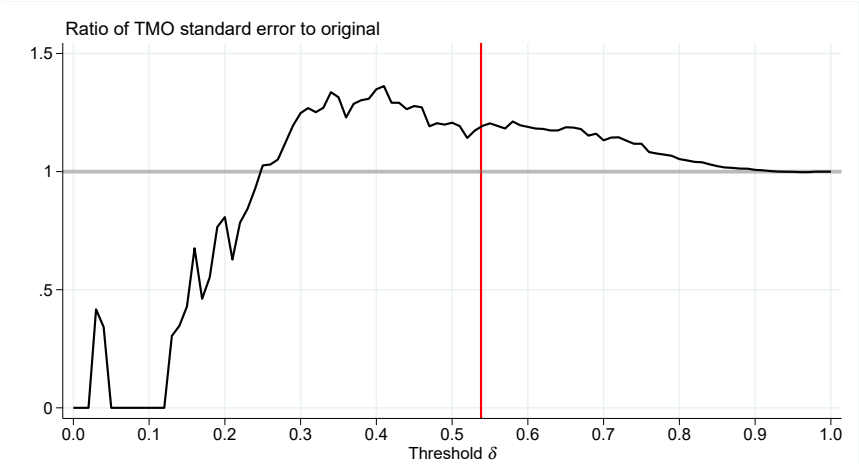


FIGURE E.8. Distribution of Correlations Between Locations in Applications

Bernini et al. (2023)



Bazzi et al. (2023)



Calderon et al. (2023)

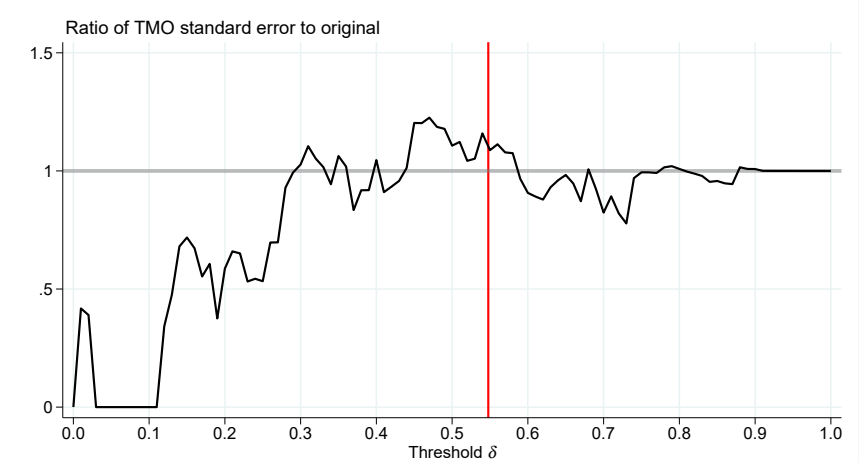
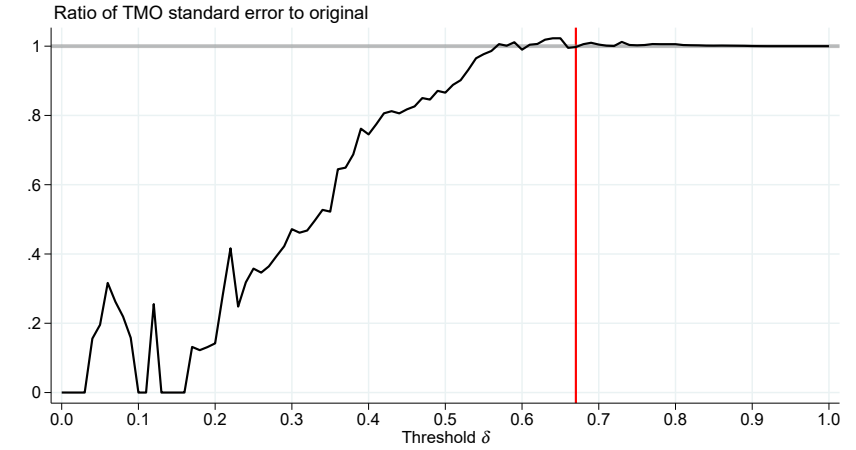
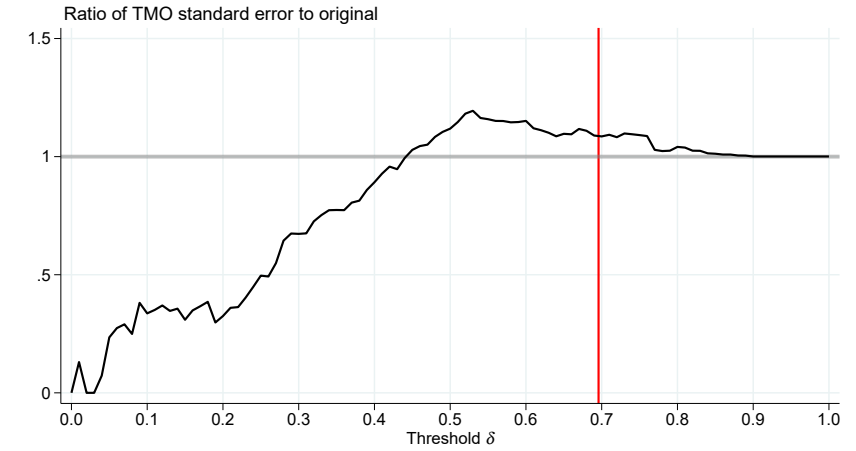


FIGURE E.8. Distribution of Correlations Between Locations in Applications

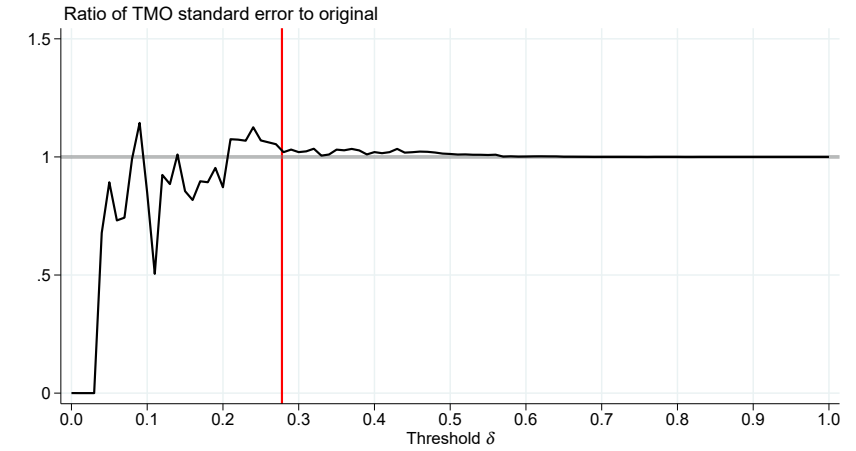
Moscona and Sastry (2023)



Chetty et al. (2014)



Acemoglu et al. (2019)



Notes: Figure E.8 plots the ratio of the TMO standard error relative to the original standard error in the papers as the threshold δ varies. The vertical line marks the estimate of the optimal threshold $\hat{\delta}^*$.