cemmap
centre for microdata methods and practice

Inference for ranks with applications to mobility across neighborhoods and academic achievement across countries

Magne Mogstad
Joseph P. Romano
Azeem M. Shaikh
Daniel Wilhelm

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP17/21

Economic and Social Research Council

Inference for Ranks with Applications to Mobility across Neighborhoods and Academic Achievement across Countries*

Magne Mogstad
Department of Economics
University of Chicago
Statistics Norway & NBER
magne.mogstad@gmail.com

Joseph P. Romano
Departments of Statistics and Economics
Stanford University
romano@stanford.edu

Azeem M. Shaikh
Department of Economics
University of Chicago
amshaikh@uchicago.edu

Daniel Wilhelm
Department of Economics
University College London
d.wilhelm@ucl.ac.uk

April 5, 2021

Abstract

It is often desired to rank different populations according to the value of some feature of each population. For example, it may be desired to rank neighborhoods according to some measure of intergenerational mobility or countries according to some measure of academic achievement. These rankings are invariably computed using estimates rather than the true values of these features. As a result, there may be considerable uncertainty concerning the rank of each population. In this paper, we consider the problem of accounting for such uncertainty by constructing confidence sets for the rank of each population. We consider both the problem of constructing marginal confidence sets for the rank of a particular population as well as simultaneous confidence sets for the ranks of all populations. We show how to construct such confidence sets under weak assumptions. An important feature of all of our constructions is that they remain computationally feasible even when the number of populations is very large. We apply our theoretical results to re-examine the rankings of both neighborhoods in the United States in terms of intergenerational mobility and developed countries in terms of academic achievement. The conclusions about which countries do best and worst at reading, math, and science are fairly robust to accounting for uncertainty. The confidence sets for the ranking of the 50 most populous commuting zones by mobility are also found to be small. However, the mobility rankings become much less informative if one includes all commuting zones, if one considers neighborhoods at a more granular level (counties, Census tracts), or if one uses movers across areas to address concerns about selection.

KEYWORDS: Confidence sets, Directional errors, Familywise error rate, Intergenerational mobility, Multiple testing, PISA, Ranks

JEL classification codes: C12, C14, D31, I20, J62

*The second author acknowledges support from the National Science Foundation (MMS-1949845). The third author acknowledges support from the National Science Foundation (SES-1530661). The fourth author acknowledges support from the ESRC Centre for Microdata Methods and Practice at IFS (RES-589-28-0001) and the European Research Council (Starting Grant No. 852332). We thank Sergei Bazylik and Eyo Herstad for excellent research assistance and the following individuals for helpful comments: Raj Chetty, Nathan Hendren, Helmut Küchenhoff, Imran Rasul, and David Ritzwoller.

1 Introduction

Rankings of different populations according to the value of some feature of each population are ubiquitous. Interest in such rankings stem from their ability to convey succinct answers to various questions, such as whether a particular population is "good" or "bad" in terms of the value of this feature relative to other populations, or which populations are "best" or "worst" in terms of the value of this feature. A prominent example from the recent economics literature is provided by Chetty et al. (2014, 2018) and Chetty and Hendren (2018), in which different populations correspond to different neighborhoods in the United States and the feature by which it is desired to rank them is some measure of intergenerational mobility. A further example of contemporary interest is provided by the Programme for International Student Assessment (PISA), in which different populations correspond to different countries and the feature by which it is desired to rank them is some measure of academic achievement. These rankings are invariably computed using estimates rather than the true values of these features. As a result, there may be considerable uncertainty concerning the rank of each population.

In this paper, we consider the problem of accounting for such uncertainty by constructing confidence sets for the rank of each population. We consider both marginal confidence sets for the rank of a particular population, i.e., random sets that contain the rank of the particular population of interest with probability approximately no less than some pre-specified level, as well as simultaneous confidence sets for the ranks of all populations, i.e., random sets that contain the ranks of all populations with probability approximately no less than some pre-specified level. The former confidence sets provide a way of accounting for uncertainty when answering questions pertaining to the rank of a particular population, whereas the latter confidence sets provide a way of accounting for uncertainty when answering questions pertaining to the ranks of all populations. We show how to construct both types of confidence sets under weak assumptions. An important feature of all of our constructions is that they remain computationally feasible even when the number of populations is very large. We apply our inference procedures to re-examine the rankings of both neighborhoods in the United States in terms of intergenerational mobility and developed countries in terms of academic achievement.

For each of the preceding confidence sets, we first show how they can be constructed using simultaneous confidence sets for differences across the populations in the values of the features. The main requirement underlying our analysis is only that these latter confidence sets for the differences are suitably valid. Our procedure is therefore applicable not only to rankings of "populations" narrowly defined, but rather to rankings of any objects (e.g. also treatments, treatment assignment rules, forecasting rules, etc.) as long as suitably valid confidence sets for the differences of the objects' performance measures can be constructed. We show, however, that it is possible to improve upon this construction using a suitable multiple hypothesis testing problem without imposing any further assumptions. In this sense, the assumptions involved in establishing our formal results are weak. A novel feature of the multiple hypothesis testing problem we consider is that it requires control of the mixed-directional familywise error rate rather than simply the familywise error rate. As the terminology suggests, the distinction between these two error rates is that the former penalizes not only false rejections, like the latter, but also false directional assertions. For further discussion, see Bauer et al. (1986) as well as Sections 3.2.2 and 3.3.2 below.

As a specific example of the way in which the aforementioned confidence sets may be used by researchers, we examine in more depth the question of identifying which populations are among the top (or the bottom).

For concreteness, we define a population to be among the top if its rank is less than or equal to a pre-specified value τ. In order to account for uncertainty when answering this question, it may be desired to construct what we subsequently refer to as a confidence set for the τ-best populations, i.e., a random set that contains the identities of these populations with probability approximately no less than some pre-specified level. While it is possible to use simultaneous confidence sets for the ranks of all populations to construct such confidence sets, we show that it is possible to improve upon this construction without imposing any further assumptions.

In order to illustrate the widespread applicability of our inference procedure, we use it to re-examine the rankings of both neighborhoods in the United States in terms of intergenerational mobility and developed countries in terms of academic achievement. The former application uses data from Chetty et al. (2014, 2018), while the latter application uses data from the 2018 PISA test. In each application, we apply our methodology to compute (i) the marginal confidence sets for the rank of a given place, (ii) the simultaneous confidence sets for the ranks of all places, and (iii) the confidence sets for the τ-best (or the τ-worst) places.

Before describing our empirical results, we emphasize that (i)–(iii) answer distinct economic questions. Consider, for example, the application to intergenerational mobility and neighborhoods. Marginal confidence sets answer the question of whether a given place has relatively high or low income mobility compared to other places. Thus, (i) is relevant if one is interested in whether a particular place is among the worst or the best places to grow up in terms of income mobility. Simultaneous confidence sets allow such inferences to be drawn simultaneously across all places. Thus, (ii) is relevant if one is interested in broader geographic patterns of income mobility across the United States. By comparison, confidence sets for the τ-best (or τ-worst) answer the more specific question of which places cannot be ruled out as being among the areas with the most (least) income mobility. In other words, (iii) is relevant if one is interested in only the top (or bottom) of a league table of neighborhoods by income mobility.

In our analysis of data from the 2018 PISA test, we find that the conclusions about which developed countries do best and worst at reading, math, and science are fairly robust to accounting for uncertainty. Both the marginal and simultaneous confidence sets are relatively narrow, especially for the countries at the top and the bottom of the PISA league tables. Indeed, only a small set of countries cannot be ruled out as being among the top or bottom three in terms of scholastic performance.

In our analysis of data from Chetty et al. (2014, 2018), we find that several celebrated findings about intergenerational income mobility in the United States are not robust to taking uncertainty into account. The key outputs from these studies were "local statistics" on upward mobility across commuting zones or counties.[1] The stated goal was to draw the attention of policymakers to low-mobility neighborhoods that need improvement and to help low-income families move to high-mobility neighborhoods. We examine how informative these local statistics are about a given neighborhood having relatively high or low income mobility compared to other neighborhoods.

The most robust findings are obtained if we restrict attention to the 50 most populous commuting zones or counties. In that case, both the marginal and simultaneous confidence sets are relatively narrow, and few places cannot be ruled out as being among the top or bottom five. By comparison, in the national ranking of all commuting zones or counties by income mobility, it is often not possible to determine with

[1] A key contribution of Chetty et al. (2014, 2018) is the granular estimates of intergenerational mobility. For a description of of intergenerational mobility across broader regions of the U.S., see for example Connor and Storper (2020) and the references therein.

statistical confidence whether a given place has relatively high or low income mobility compared to other places. Another key finding is that the rankings of even the most populous commuting zones or counties become largely uninformative if one uses movers across areas to address concerns about selection and draw causal conclusions.

In order to illustrate the policy relevance of these findings, we revisit the recent Creating Moves to Opportunity Experiment (CMTO) of Bergman et al. (2019). With the aim of helping families move to neighborhoods with higher mobility rates, the authors conduct a randomized controlled trial with housing voucher recipients in Seattle and King County. A treatment group of low-income families were offered assistance and financial support to find and lease units in areas that were classified as high upward-mobility neighborhoods within the county. The authors define high upward-mobility neighborhoods as Census tracts with point estimates of upward mobility among the top one-third of the tracts in the county. Since no data on outcomes is yet available, the authors predict the impacts of the moves induced by the CMTO program on children's future outcomes using the point estimates of upward mobility of the individual tracts. We show that the classification of a given tract as a high upward-mobility neighborhood may simply reflect statistical uncertainty, not that mobility is particularly high in that neighborhood. We discuss the implication of this finding for the assumptions needed to be statistically confident, prior to the experiment, that CMTO would actually help families move to high opportunity neighborhoods.

Our paper is most closely related to a recent paper by Klein et al. (2020), who consider the problem of constructing confidence sets analogous to ours. The main difference between their constructions and ours is that they rely upon simultaneous confidence sets for the values of the features for all populations, whereas, as mentioned previously, we exploit simultaneous confidence sets for differences in the values of the features for certain pairs of populations. In Remark 3.11 and Appendix B, we show that their confidence sets are always at least as large as ours when there are only two populations or in the homogeneous case with common variances and sample sizes when there are more than two populations. More importantly, we show that their method cannot in general produce smaller confidence sets with positive probability uniformly across populations. While it is unknown if even one component may be smaller with positive probability, we find in our simulations that their approach generally leads to confidence sets that are much larger than ours for all populations.

Other related work includes Goldstein and Spiegelhalter (1996), who propose the use of resampling methods such as the bootstrap to account for the type of uncertainty with which we are concerned. In the context of the PISA study, for instance, such a bootstrap procedure has been used to report "range of ranks" (see OECD (2019, Annex A3)). As explained by Hall and Miller (2009) and Xie et al. (2009), however, such methods perform poorly when some populations have features whose values are "close" to one another. In Remark 3.7 and Appendix A, we show that the bootstrap does not satisfy the coverage requirement when there are more than two populations. Motivated by these observations, Xie et al. (2009) propose an alternative method for accounting for uncertainty based on combining resampling with a smooth estimator of the rank which requires, among other things, delicate choices of user-specified "bandwidths". Our constructions, by contrast, require no such tuning parameters.

Finally, we note that the problems studied in this paper are distinct from those of two recent papers in econometrics, Andrews et al. (2018) and Gu and Koenker (2020). To explain the differences consider the example of intergenerational mobility in the U.S.. Andrews et al. (2018) develop methods for inference on the value of mobility in the neighborhood with the highest estimated mobility. In contrast, we develop

methods for inference on the rank of a neighborhood, not on the value of mobility that was used to rank neighborhoods; see Remarks 5.1 and 5.3 for more discussion. Gu and Koenker (2020) develop optimal decision rules for selecting the best neighborhoods, which is related to a literature in subset selection (see Gupta and Panchapakesan (1979) for a review), but complementary to our inference-based approach of selecting the τ-best; see Remark 3.17 for more details.

The remainder of our paper is organized as follows. In Section 2, we illustrate the logic underlying our inference procedures in a stylized example using a subset of the data from one of our empirical applications. Section 3 then introduces our general setup, including a formal description of the confidence sets we consider. We first discuss the construction of a marginal confidence set for the rank of a particular population and then turn our attention to the construction of simultaneous confidence sets for the ranks of all populations. As mentioned previously, in each case, we begin by describing a simple construction that relies on simultaneous confidence sets for certain pairs of populations before showing how to improve upon this construction using an appropriately chosen multiple hypothesis testing problem. In Section 4, we examine the finite-sample behavior of our inference procedure via a simulation study, including a comparison with the method proposed by Klein et al. (2020). Finally, in Section 5, we apply our inference procedures to re-examine the rankings of both developed countries in terms of academic achievement and neighborhoods in the United States in terms of intergenerational mobility.

2 Inference for Ranks in a Stylized Example

Suppose it is desired to rank five commuting zones (CZs) in the United States by a measure of upward intergenerational mobility. Denote by r_j the rank of CZ j based on the mobility measure θ_j. Panel A of Figure 1 shows estimated mobility measures $\hat{\theta}_j$ with 95% marginal confidence intervals (estimates plus or minus twice the standard error) for five CZs from our dataset in Section 5.2. Linton and Albany have the highest and lowest mobility estimates among these five CZs and thus the smallest ($\hat{r}_j = 1$ for $j =$ Linton) and highest ($\hat{r}_j = 5$ for $j =$ Albany) estimated ranks, respectively. Since $\hat{\theta}_j$ is an estimate of θ_j, the estimated rank $\hat{r}_j$ may not equal the true rank r_j. In particular, Linton need not have the highest mobility and Albany need not have the lowest mobility.

Table 1 summarizes the results of accounting for uncertainty in the ranks of these five CZs using (i) marginal confidence sets for the rank of a single CZ, (ii) simultaneous confidence sets for the ranks of all CZs (i.e., for the entire ranking), and (iii) confidence sets containing the τ-best CZs. We first report the estimated ranks as well as the point estimates and their standard errors. As explained further below, these data are all that is required to compute (i)–(iii). The sixth column reports the first set of results, marginal confidence sets for the rank of each CZ. The second set of results is reported in the seventh column, which displays simultaneous confidence sets for the ranks of all CZs. In general, the simultaneous confidence sets are at least as large as the marginal ones, but in this example they are identical. The last set of results is reported in the final column, showing the number of CZs contained in the confidence set for the τ-best, where τ varies from one to five across the rows. For instance, with at least 95% confidence, there is only one CZ that can be the best and there are four CZs that can be among the top two.

The remainder of this section describes how we arrive at the three set of results in Table 1 in the context of this example.

					95% CS		
Rank	τ	CZ	$\hat{\theta}_j$	SE	marg.	simul.	τ-best
1	1	Linton	0.608	0.014	[1, 1]	[1, 1]	1
2	2	Gordon	0.443	0.010	[2, 4]	[2, 4]	4
3	3	Trenton	0.433	0.010	[2, 4]	[2, 4]	4
4	4	Jordan	0.413	0.050	[2, 5]	[2, 5]	5
5	5	Albany	0.331	0.002	[4, 5]	[4, 5]	5

Table 1: Commuting zones (CZs) ranked by the estimated intergenerational mobility measure $\hat{\theta}_j$. "SE" refers to the standard error of $\hat{\theta}_j$. "95% CS (marg.)" refers to the 95% marginal confidence set for the rank, "95% CS (simul.)" to the 95% simultaneous confidence set for all ranks, and "τ-best" refers to the size of the 95% confidence set for the "τ-best" CZs.

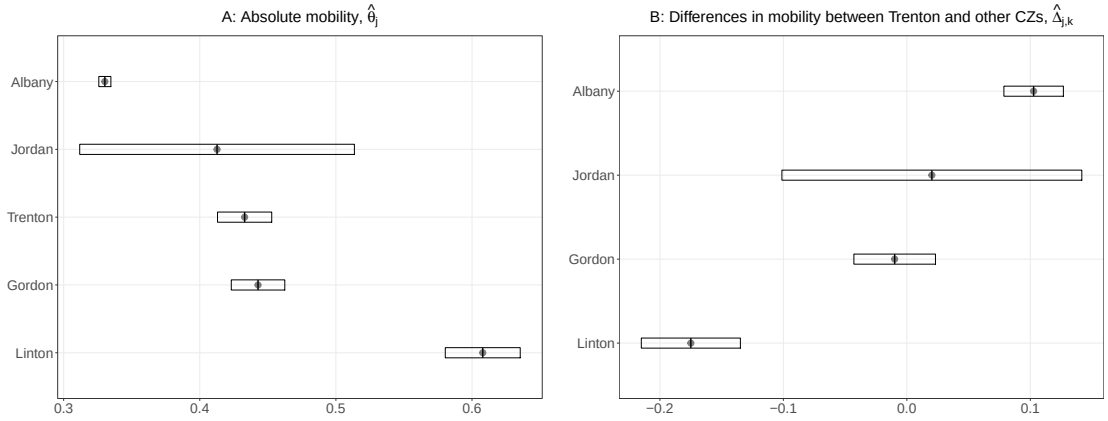

Figure 1: Panel A shows the estimated mobility with 95% (marginal) confidence sets (estimates plus or minus twice the standard error) for five CZs. Panel B shows the estimated differences in mobility between Trenton and all other CZs, together with 95% (simultaneous) confidence sets. Each marginal confidence set covers a single mobility measure with probability 95% whereas the simultaneous confidence sets simultaneously cover all differences in mobility measures with probability 95%.

Inference on the rank of a particular CZ

Suppose we are interested in the rank of Trenton. From Panel A of Figure 1, we see that its estimated rank is three, but the mobility estimate is close to that of Gordon and Jordan's mobility estimate has a large standard error, so one might be uncertain whether Trenton's rank is in fact larger or smaller than three. In order to move beyond this conjecture, we use the following two-step procedure to construct a confidence set for Trenton's rank.

First, we consider the differences in mobility estimates between Trenton and all other CZs. It is clear that only the signs of the differences in mobility estimates between Trenton and all other CZs being positive or negative determine Trenton's rank. These differences are displayed in Panel B of Figure 1 together with 95% simultaneous confidence sets. Simultaneous coverage of this confidence set is important. In order to explain the simultaneous coverage property, it is useful to introduce some further notation. To this end, let $\hat{\Delta}_{j,k}$ be the estimator of the difference in mobility $\Delta_{j,k} \equiv \theta_j - \theta_k$ for $j =$ Trenton and $k \in \{\text{Linton, Gordon, Jordan, Albany}\}$. The confidence set in Panel B of Figure 1 is the product of four confidence

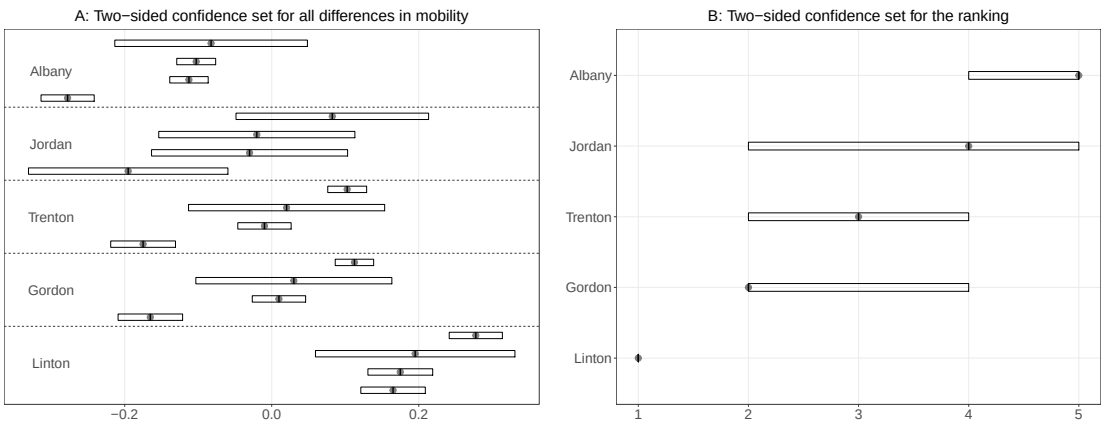

Figure 2: Panel A shows all estimated differences in mobility, together with 95% (simultaneous) two-sided confidence sets. Panel B shows estimated ranks together with 95% (simultaneous) two-sided confidence sets.

sets so the probability of it simultaneously covering all four differences $\Delta_{j,k}$ for j = Trenton and $k \in \{\text{Linton}, \text{Gordon}, \text{Jordan}, \text{Albany}\}$ is at least 95%. The bounds for the simultaneous confidence sets depend on quantiles from the distribution of the maximum (over k) of the differences $\hat{\Delta}_{j,k} - \Delta_{j,k}$. In Section 3.1, we explain how such quantiles may be approximated using the bootstrap, but other constructions are also possible.

Second, given the simultaneous confidence set for the differences in mobility, we count how many of the individual confidence sets lie entirely above and below zero. The first confidence set, which is for the difference in mobility between Trenton and Linton, lies entirely below zero. Therefore, we can conclude that Linton has significantly higher mobility than Trenton and thus must be ranked strictly better than Trenton. The differences in mobility between Trenton and either Gordon and Jordan are not significantly different from zero, so these three CZs cannot be ranked relative to each other. The confidence set for the difference in mobility between Trenton and Albany lies entirely above zero, so that Albany must be ranked strictly worse than Trenton. Using the notation of the subsequent sections, there is one CZ that must be ranked strictly better, $|N_j^-| = 1$, and one CZ that must be ranked strictly worse, $|N_j^+| = 1$. The confidence set for the rank of Trenton among the $p = 5$ CZs is therefore

$$R_{n,j} = \{|N_j^-| + 1, \ldots, p - |N_j^+|\} = \{2, 3, 4\} \ .$$

By virtue of the simultaneous coverage property for the differences described above, this set contains the rank r_j of Trenton with probability at least 95%.

While simple in nature, the preceding procedure illustrates the logic underlying all of our constructions. In Section 3.2.2, we show that the confidence set $R_{n,j}$ can be improved through the use of a suitable stepwise multiple testing procedure. In the first step of the procedure, some CZs are determined to be ranked higher or lower than the CZ of interest in exactly the manner described above; in subsequent steps, further CZs are possibly determined to be ranked higher or lower than the CZ of interest by appropriately accounting for those that were determined to be ranked higher or lower in previous steps. This process continues until no further CZs can be determined to be ranked higher or lower than the CZ of interest.

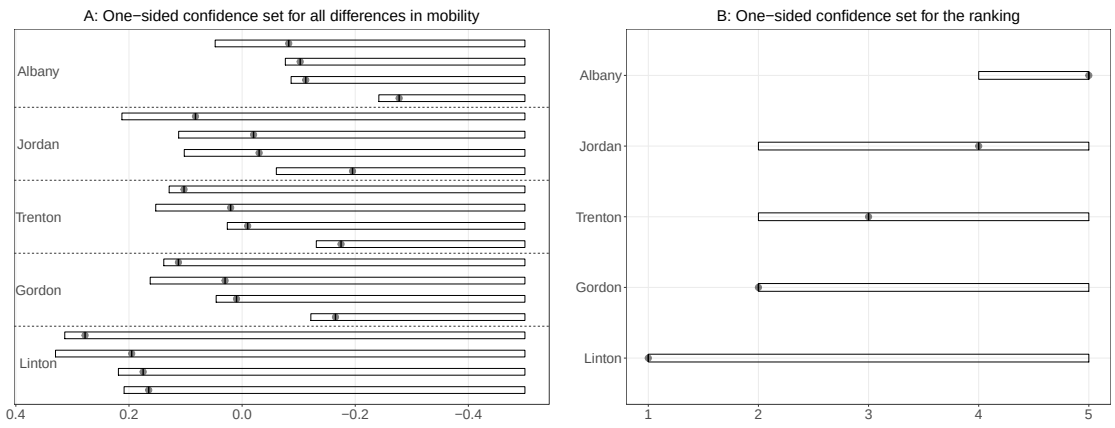

Figure 3: Panel A shows all estimated differences in mobility, together with 95% (simultaneous) one-sided confidence sets. Panel B shows estimated ranks together with 95% (simultaneous) one-sided confidence sets.

Inference on the entire ranking

In order to construct a simultaneous confidence set for the entire ranking of all five CZs, rather than only for Trenton, the approach is modified in the following fashion. We begin by computing every possible difference in mobility estimates between all CZs, not only those involving Trenton. These differences are shown in Panel A of Figure 2 together with simultaneous confidence sets. In this case, the confidence sets simultaneously cover all differences $\Delta_{j,k}$, for all $j, k \in \{\text{Linton}, \text{Gordon}, \text{Trenton}, \text{Jordan}, \text{Albany}\}$ with $j \neq k$. For each CZ j, we then count how many confidence sets k lie above and below zero. For instance, for $j = \text{Trenton}$, we obtain the same result as above, namely that one confidence set lies entirely below and one lies entirely above zero, so $|N_j^-| = 1$ and $|N_j^+| = 1$. For $j = \text{Linton}$, all confidence sets lie above zero, so $|N_j^-| = 0$ and $|N_j^+| = 4$. The confidence sets for each CZ are then constructed using these counts just as above.

The result of this procedure is shown in Panel B of Figure 2. The confidence set for Linton contains only rank one, but the confidence sets for the other CZs contain two to four values. By virtue of the simultaneous coverage property for the differences, the product of the five CZ-specific confidence sets for the ranks simultaneously covers the ranks of all CZs, i.e., r_j for all $j \in \{\text{Linton}, \text{Gordon}, \text{Trenton}, \text{Jordan}, \text{Albany}\}$, with probability at least 95%.

In Section 3.3.2, we show that this simple construction can also be improved through the use of a multiple testing procedure that parallels the one that we use for the marginal confidence set.

Confidence sets containing the τ-best CZs

Suppose it is desired to determine which of the five CZs could be among the $\tau = 2$ best CZs. From Panel A of Figure 1, we see that Linton and Gordon have ranks one and two, but the mobility estimate of Trenton is close to that of Gordon and Jordan's mobility estimate has large standard errors, so one might consider that Trenton or Jordan could also be among the top two CZs. The following procedure provides a means of accounting for this uncertainty formally.

Denote the set of CZs which are among the two best as $J_0^{2-\text{best}} \equiv \{j : r_j \leq 2\}$. This set contains at least two CZs and strictly more than two when CZs are tied at rank one or two. We want to construct a set

$J_n^{2-\text{best}}$ that contains the set $J_0^{2-\text{best}}$ with probability at least 95%.

A simple approach is based on one-sided simultaneous confidence sets for all ranks. Figure 3 repeats the computations for Figure 2 except the simultaneous confidence sets for the differences are one-sided (upper bounds) and the resulting simultaneous confidence sets for the ranks are therefore also one-sided (lower bounds). Let $R_{n,j}^{\text{joint}}$ be the jth dimension of the one-sided confidence set for the ranking, i.e., the confidence set for r_j in Panel B of Figure 3. In order to construct a set with the desired coverage property, it suffices to collect all CZs j for which $R_{n,j}^{\text{joint}}$ contains the value two, i.e.,

$$J_n^{2-\text{best}} = \{j \colon 2 \in R_{n,j}^{\text{joint}}\} = \{\text{Linton, Gordon, Trenton, Jordan}\} \ .$$

By virtue of the coverage property of the simultaneous confidence set for the ranking, this set covers the set of the two best CZs, $J_0^{2-\text{best}}$, with probability at least 95%. While this "projection" approach for constructing the confidence set is parsimonious and intuitive, improvements may be possible by realizing that a CZ can be among the top two if and only if at most one other CZ (without regard to its identity) has higher mobility. By comparison, the one-sided simultaneous confidence sets for all ranks R_n^{joint} encodes some information about which CZs have higher mobility than another. In Section 3.4, we propose a more "direct" procedure based on exploiting the insight and show through simulations that it leads to smaller confidence sets.

Key features of the inference approach

Section 3 formally shows that, under weak assumptions, the above three constructions of confidence sets asymptotically control the probability of covering the objects of interest at the desired level (95% in the example of this section) uniformly over a large class of possible distributions for the observed data. The following two aspects of the theoretical results are especially important in our empirical applications and can already be understood in the context of the example in Panel A of Figure 1.

First, in our applications, we see that many estimates $\hat{\theta}_j$ are close to one other, such as the mobility estimates of Gordon and Trenton in the preceding example. It is therefore important to develop inference methods that do not break down when some (or even all) measures θ_j are (close to) equal to one another. Formally, our confidence sets achieve this goal because we show that they guarantee coverage uniformly over a large family of distributions for the observed data, and hence uniformly over all configurations of measures $\theta_1, \ldots, \theta_p$, irrespectively of whether some (or even all) of them are (close to) equal to each other.

Second, our confidence sets satisfy the uniform coverage requirement under weak conditions. In particular, the distributions of $\hat{\theta}_j - \theta_j$ are allowed to vary across j. Such heterogeneity is salient in our empirical applications and its importance can already be seen in Panel A of Figure 1: Trenton's mobility estimate has much smaller standard error than that of Jordan, but much larger than that of Albany.

3 General Setup and Main Results

3.1 Setup and Notation

Let $j \in J \equiv \{1, \ldots, p\}$ index populations of interest. Denote by P_j distributions characterizing the different populations and by $\theta(P_j)$ the associated features by which it is desired to rank them. In the example of

Section 2, j denotes a county, $\theta(P_j)$ is a measure of intergenerational mobility in county j, and P_j is the distribution from which we observe data for estimation of the feature $\theta(P_j)$. The rank of population j is defined as

$$r_j(P) \equiv 1 + \sum_{k \in J} \mathbb{1}\{\theta(P_k) > \theta(P_j)\},$$

where P is a distribution with marginals P_j for $j \in J$, and $\mathbb{1}\{A\}$ is equal to one if the event A holds and equal to zero otherwise. Let $\theta(P) \equiv (\theta(P_1), \ldots, \theta(P_p))'$ and $r(P) \equiv (r(P_1), \ldots, r_p(P_p))'$. Before proceeding, a simple example illustrates the way in which ties are handled with this definition of ranks: if $\theta(P) = (4, 1, 1, 3, 3, 3, 6)'$, then $r(P) = (2, 6, 6, 3, 3, 3, 1)'$.

The primary goal is to construct confidence sets for the rank of a particular population or for the ranks of all populations simultaneously. More precisely, for a given value of $\alpha \in (0, 1)$, we use a sample of observations from P to construct (random) sets $R_{n,j}$ such that

$$\liminf_{n \to \infty} \inf_{P \in \mathbf{P}} P\{r_j(P) \in R_{n,j}\} \geq 1 - \alpha \tag{1}$$

for a pre-specified population $j \in J$, where $\mathbf{P}$ denotes a "large" nonparametric family of distributions. Here, n denotes a measure of the size of the sample, typically the minimum sample size across populations. We also construct (random) sets $R_n^{\text{joint}} \equiv \prod_{j \in J} R_{n,j}^{\text{joint}}$ such that

$$\liminf_{n \to \infty} \inf_{P \in \mathbf{P}} P\left\{r(P) \in R_n^{\text{joint}}\right\} \geq 1 - \alpha \ . \tag{2}$$

In all of our constructions, $R_{n,j}$ and $R_{n,j}^{\text{joint}}$ are subsets of J, allowing for the possibility that the lower endpoint is 1 or the upper endpoint is p to permit both one-sided and two-sided inference. Below, sets satisfying (1) are referred to as *marginal confidence sets for the rank of a single population* and sets satisfying (2) as *simultaneous confidence sets for the ranks of all populations.*

In addition, we consider the goal of constructing confidence sets for the identities of all populations whose rank is less than or equal to a pre-specified value $\tau \in J$, i.e, for a given value of $\alpha \in (0, 1)$, we construct (random) sets $J_n^{\tau-\text{best}}$ that are subsets of J and satisfy

$$\liminf_{n \to \infty} \inf_{P \in \mathbf{P}} P\left\{J_0^{\tau-\text{best}}(P) \subseteq J_n^{\tau-\text{best}}\right\} \geq 1 - \alpha \ , \tag{3}$$

where

$$J_0^{\tau-\text{best}}(P) \equiv \{j \in J : r_j(P) \leq \tau\} \ .$$

Sets satisfying (3) are referred to as *confidence sets for the τ-best populations.*

Much of the analysis relies upon confidence sets $C_n(1 - \alpha, S)$ for sets of pairwise differences,

$$\Delta_S(P) \equiv (\Delta_{j,k}(P) : (j, k) \in S) \ ,$$

where $\Delta_{j,k}(P) \equiv \theta(P_j) - \theta(P_k)$ and $S \subseteq \{(j, k) \in J \times J : j \neq k\}$. We require these to be rectangular in the sense that

$$C_n(1 - \alpha, S) = \prod_{(j,k) \in S} C_n(1 - \alpha, S, (j, k)) \tag{4}$$

for suitable sets $\{C_n(1-\alpha, S, (j,k)) : (j,k) \in S\}$. Furthermore, we assume that they satisfy

$$\liminf_{n\to\infty} \inf_{P\in\mathbf{P}} P\{\Delta_S(P) \in C_n(1-\alpha, S)\} \geq 1-\alpha \ . \tag{5}$$

We now describe some examples of confidence sets that satisfy these two conditions. Let $\hat{\theta}_1, \ldots, \hat{\theta}_p$ be estimators of the features $\theta(P_1), \ldots, \theta(P_p)$ and $\hat{\sigma}^2_{j,k}$ an estimator of the variance of $\hat{\theta}_j - \hat{\theta}_k$. For $S \subseteq \{(j,k) \in J \times J : j \neq k\}$, define the following cumulative distribution functions:

$$L_{\text{lower},n}(x, S, P) \equiv P\left\{ \max_{(j,k)\in S} \frac{\hat{\theta}_j - \hat{\theta}_k - \Delta_{j,k}(P)}{\hat{\sigma}_{j,k}} \leq x \right\} , \tag{6}$$

$$L_{\text{upper},n}(x, S, P) \equiv P\left\{ \max_{(j,k)\in S} \frac{\Delta_{j,k}(P) - (\hat{\theta}_j - \hat{\theta}_k)}{\hat{\sigma}_{j,k}} \leq x \right\} , \tag{7}$$

$$L_{\text{symm},n}(x, S, P) \equiv P\left\{ \max_{(j,k)\in S} \frac{|\hat{\theta}_j - \hat{\theta}_k - \Delta_{j,k}(P)|}{\hat{\sigma}_{j,k}} \leq x \right\} . \tag{8}$$

Further consider estimators of (6) – (8) using an estimate $\hat{P}_n$ of P to define the following confidence sets:

$$C_{\text{lower},n}(1-\alpha, S) \equiv \prod_{(j,k)\in S} \left[\hat{\theta}_j - \hat{\theta}_k - \hat{\sigma}_{j,k} L^{-1}_{\text{lower},n}(1-\alpha, S, \hat{P}_n), \infty \right), \tag{9}$$

$$C_{\text{upper},n}(1-\alpha, S) \equiv \prod_{(j,k)\in S} \left(-\infty, \hat{\theta}_j - \hat{\theta}_k + \hat{\sigma}_{j,k} L^{-1}_{\text{upper},n}(1-\alpha, S, \hat{P}_n) \right], \tag{10}$$

$$C_{\text{symm},n}(1-\alpha, S) \equiv \prod_{(j,k)\in S} \left[\hat{\theta}_j - \hat{\theta}_k \pm \hat{\sigma}_{j,k} L^{-1}_{\text{symm},n}(1-\alpha, S, \hat{P}_n) \right], \tag{11}$$

$$C_{\text{equi},n}(1-\alpha, S) \equiv C_{\text{lower},n}\left(1-\frac{\alpha}{2}, S\right) \bigcap C_{\text{upper},n}\left(1-\frac{\alpha}{2}, S\right) . \tag{12}$$

Here, it is understood that, for a cumulative distribution function $F(x)$ on the real line, the quantity $F^{-1}(1-\alpha)$ is defined to be $\inf\{x \in \mathbf{R} : F(x) \geq 1-\alpha\}$; it is also understood that, for real numbers a and b, $[a \pm b]$ is defined to be $[a-b, a+b]$. If the estimators $\hat{\theta}_1, \ldots, \hat{\theta}_p$ are jointly asymptotically normally distributed, then the quantiles $L^{-1}_{t,n}(1-\alpha, S, \hat{P}_n)$, $t \in \{\text{lower}, \text{upper}, \text{symm}\}$, can be computed from the limiting distributions of the max-statistics in (6)–(8), e.g., through simulation. Alternatively, resampling methods such as the bootstrap may be employed.

The four confidence sets in (9)–(12) can be viewed as nonparametric generalizations of Tukey (1953)'s method for all pairwise comparisons and Dunnett (1955)'s method for comparisons with a control. The classical methods rely on the assumptions of normal populations and equal variances, under which critical values can be computed using Tukey's studentized range distribution or Dunnet's two-sided range distribution. We do not impose either of these assumptions, but rather only require an estimate $\hat{P}_n$ of P so that the resulting confidence set satisfies (5). The argument establishing this condition determines how $\mathbf{P}$ and $\hat{P}_n$ should be defined. For example, suppose we observe an i.i.d. sample $X_1, \ldots, X_n$, where $X_i \equiv (X_{i,1}, \ldots, X_{i,p})'$ has distribution P. When $\mathbf{P}$ is the set of distributions on $\mathbb{R}^p$ satisfying a uniform integrability condition, then the bootstrap leads to confidence sets satisfying (5) when $\theta(P)$ is the population mean vector and $\hat{\theta}_n$ is the sample mean vector. For other parameters and estimators, see Romano and Shaikh (2012). This result

may also be adapted to the case in which, for each population $j \in J$, we observe n_j realizations from a distribution P_j and the populations are independent of each other, i.e., $X_{i,j}$ is independent of $X_{k,l}$ for all i, j, k, l such that $j \neq l$.

Remark 3.1. In light of the above discussion, whether the estimators $\hat{\theta}_1, \ldots, \hat{\theta}_p$ are dependent or not does not pose any conceptual challenges to constructing confidence sets $C_n(1-\alpha, S)$ satisfying (5). ■

Remark 3.2. Romano and Shaikh (2012) provide a general theory for establishing (5) in the case when the number of observations in each population diverges and the number of populations p is fixed. However, the results can be extended to the high-dimensional case in which p diverges, using, for instance, high-dimensional central limit theorems in Chernozhukov et al. (2013, 2017, 2019). For relevant results in the case in which each $\hat{\theta}_j$ is a sample mean, see Bai et al. (2019). ■

3.2 Marginal Confidence Sets for the Rank of a Single Population

3.2.1 A Simple Construction

Suppose we want to construct a confidence set for the rank of population $j \in J$. Define $S_j \equiv \{(j,k) \colon k \in J \setminus \{j\}\}$. For a confidence region $C_n(1-\alpha, S_j)$ for $\Delta_{S_j}(P)$ that is rectangular in the sense of (4) with $\{C_n(1-\alpha, S_j, (j,k)) : (j,k) \in S_j\}$, define

$$\begin{aligned} N_j^- &\equiv \{k \in J \setminus \{j\} \colon C_n(1-\alpha, S_j, (j,k)) \subseteq \mathbf{R}_-\}, \\ N_j^+ &\equiv \{k \in J \setminus \{j\} \colon C_n(1-\alpha, S_j, (j,k)) \subseteq \mathbf{R}_+\}, \end{aligned}$$

where $\mathbf{R}_+ \equiv (0, \infty)$ and $\mathbf{R}_- = (-\infty, 0)$. Using this notation, we have the following result:

Theorem 3.1. *If $C_n(1-\alpha, S)$ satisfies (4) with $S = S_j$, then, for any P,*

$$P\left\{|N_j^-| + 1 \leq r_j(P) \leq p - |N_j^+|\right\} \geq P\{\Delta_{S_j}(P) \in C_n(1-\alpha, S_j)\}.$$

If, in addition, $C_n(1-\alpha, S)$ also satisfies (5) with $S = S_j$, then the confidence set

$$R_{n,j} \equiv \left\{|N_j^-| + 1, \ldots, p - |N_j^+|\right\} \tag{13}$$

satisfies (1).

The lower bound of the confidence set involves the number confidence sets for the differences $\Delta_{S_j}(P)$ which lie entirely below zero, $|N_j^-|$. This quantity indicates the number of features $\theta(P_k)$ that are significantly larger than that of population j. The rank of j must therefore be strictly larger than $|N_j^-|$. Similarly, $|N_j^+|$ is the number of confidence sets that lie entirely above zero, so that there are $|N_j^+|$ populations with features $\theta(P_k)$ strictly smaller than that of population j. The rank of j can therefore be at most $p - |N_j^+|$.

The theorem shows that the confidence set $R_{n,j}$ covers the rank of population j with probability converging to at least $1-\alpha$, uniformly over distributions $P \in \mathbf{P}$. As mentioned previously, Romano and Shaikh (2012) provide conditions on $\mathbf{P}$ such that $C_n(1-\alpha, S)$ satisfies (5). The confidence set therefore asymptotically covers the rank of population j with probability no less than $1-\alpha$ even under sequences of distributions

P_n with each $P_n \in \mathbf{P}$. In particular, $R_{n,j}$ covers the rank of j with probability converging to at least $1-\alpha$ even under sequences where some (or all) of $\theta(P_{k,n})$ with $k \neq j$ approach $\theta(P_{j,n})$ as $n \to \infty$. In this sense, our results do not require the features $\theta(P_k)$ to be well separated from that of population j.

Remark 3.3. Choosing a one-sided (two-sided) confidence set $C_n(1-\alpha, S_j)$ for the differences $\Delta_{S_j}(P)$ leads to a one-sided (two-sided) confidence set $R_{n,j}$ for the rank. For instance, suppose $C_n(1-\alpha, S_j)$ is a lower bound such as (9). In that case, none of the confidence sets $C_n(1-\alpha, S_j, (j,k))$ can lie entirely below zero, so that $|N_j^-| = 0$ and the resulting confidence set for the rank is an upper bound: $R_{n,j} = \{1, \ldots, p - |N_j^+|\}$. Similarly, choosing $C_n(1-\alpha, S_j)$ to be an upper bound such as (10) leads to the one-sided confidence set $R_{n,j} = \{|N_j^-| + 1, \ldots, p\}$ on the rank. Finally, the equi-tailed confidence set in (12) leads to an equi-tailed confidence set $R_{n,j}$ in the sense that the asymptotic probability with which the true rank lies above the confidence interval is bounded above by $\alpha/2$, and similarly for the asymptotic probability that the true rank lies below. ■

Remark 3.4. Suppose $C_n(1-\alpha, S_j)$ satisfies (4)–(5) with $S = S_j$ and that each $C_n(1-\alpha, S_j, (j,k))$ with $(j,k) \in S_j$ is consistent in the sense that its length tends to zero as $n \to \infty$. If in addition all elements of $\theta(P)$ are distinct, then $R_{n,j} = r_j(P)$ with probability approaching one and, as a result, the coverage probability $P\{r_j(P) \in R_{n,j}\}$ converges to one. This feature follows from the fact that if $\theta(P_j) > \theta(P_k)$, then with probability tending to one, $C_n(1-\alpha, S_j, (j,k))$ lies entirely above zero. Similarly, if $\theta(P_j) < \theta(P_k)$, then with probability tending to one, $C_n(1-\alpha, S_j, (j,k))$ lies entirely below zero. ■

Remark 3.5. Since the coverage result in Theorem 3.1 only requires the confidence set $C_n(1-\alpha, S_j)$ to be rectangular and to satisfy (5), Remark 3.1 implies that there are no conceptual challenges in allowing for dependence in the estimators $\hat{\theta}_1, \ldots, \hat{\theta}_p$. ■

Remark 3.6. In the presence of ties, there is some ambiguity in the way in which we define the rank of a population. Let $\underline{r}_j(P) \equiv 1 + \sum_{k \in J} \mathbb{1}\{\theta(P_k) > \theta(P_j)\}$ and $\bar{r}_j(P) \equiv p - \sum_{k \in J} \mathbb{1}\{\theta(P_k) > \theta(P_j)\}$ be the smallest (i.e., best) and largest (i.e., worst) possible rank of population j. If population j is not tied with any other population, then $\underline{r}_j(P) = \bar{r}_j(P)$ and the rank is unique. On the other hand, when population j is tied with at least one other population, then $\underline{r}_j(P) < \bar{r}_j(P)$ and different definitions of the rank may select different values from the interval $R_j(P) \equiv [\underline{r}_j(P), \bar{r}_j(P)]$. Suppose $C_n(1-\alpha, S)$ satisfies (4) and (5) with $S = S_j$. An inspection of the proof of Theorem 3.1 reveals that the confidence set $R_{n,j}$ not only covers our definition of the rank, $r_j(P)$, in the sense of (1), but also any other "reasonable" definition of the rank in the sense that

$$\liminf_{n \to \infty} \inf_{P \in \mathbf{P}} P\left\{R_j(P) \subseteq R_{n,j}^{\text{cont}}\right\} \geq 1 - \alpha, \tag{14}$$

where $R_{n,j}^{\text{cont}} \equiv [\min(R_{n,j}), \max(R_{n,j})]$ is the interval from the smallest to the largest value in the confidence set $R_{n,j}$. In fact, one can show that there exist distributions P, namely those for which all $\theta(P_j)$ are equal, so that the inequality holds with equality:

$$\lim_{n \to \infty} P\left\{R_j(P) \subseteq R_{n,j}^{\text{cont}}\right\} = 1 - \alpha \ . \tag{15}$$

In this sense, our confidence set is not conservative. ■

Remark 3.7. In contrast to the confidence set $R_{n,j}$, those based on the bootstrap and Bayes approaches such as those in Goldstein and Spiegelhalter (1996) perform poorly when for some $k \neq j$ $\theta(P_k)$ is (close to) equal to $\theta(P_j)$. For concreteness, consider the following bootstrap procedure. For a population j, denote

by $\hat{\theta}_j^*$ the estimator of $\theta(P_j)$ computed on a bootstrap sample and let $\hat{r}_j^*$ be the rank computed using the bootstrap estimators $\hat{\theta}_1^*, \ldots, \hat{\theta}_p^*$. Confidence sets for r_j could then be constructed using upper and/or lower empirical quantiles of $\hat{r}_j^*$ conditional on the data. In Appendix A, we show that this intuitive approach fails to satisfy the uniform coverage requirement (1) unless $p = 2$. When there are ties with population j and $p > 2$, then the approach even fails the pointwise coverage requirement for a fixed P and, in fact, the coverage probability tends to zero as p grows. For further discussion, see Xie et al. (2009) and Hall and Miller (2009). Our approach, on the other hand, does not rely on a consistent estimator of the distribution of estimated ranks but rather on the availability of simultaneous confidence sets for the differences $\Delta_{S_j}(P)$ with asymptotically coverage no less than the desired level uniformly over $P \in \mathbf{P}$. Such simultaneous confidence sets are available under weak conditions and, in particular, do not restrict the configuration of the features $\theta(P_j)$. In comparison to Xie et al. (2009), our approach also circumvents smoothing of the indicator in the definition of the ranks and thus the need for choosing such a smoothing parameter. ■

Remark 3.8. Requiring the confidence sets for the differences $C_n(1-\alpha, S)$, to be rectangular in the sense of (4) simplifies the presentation of our approach and the results, but is not essential. We note, however, that using a non-rectangular confidence set would be equivalent to using the smallest rectangle that contains it. Since the latter set must be at least as large as the non-rectangular confidence set, using a non-rectangular confidence set would be overly conservative. The difference in size of the projected non-rectangular and the rectangular confidence sets may be particularly large in high-dimensions; see Appendix D for a simulated example. In this sense, it is undesirable to use non-rectangular confidence sets for the differences. ■

3.2.2 A Stepwise Improvement

In this section, we propose a stepwise method to improve the confidence set in Theorem 3.1. Our inference problem shares some similarities with Tukey's simultaneous comparisons of all pairwise means and Dunnet's comparisons of all means with a common control, which can be improved through the use of stepwise procedures; see Chapter 8 of Westfall et al. (1999) and Section 9 of Lehmann and Romano (2005). One key difference, however, is that the application of stepwise methods in our problem requires multiple tests that control not only the familywise error rate, but also directional errors. Unfortunately, little is known about control of directional errors in stepwise methods; Guo and Romano (2015) is one of only a few exceptions.

Consider the construction of a two-sided confidence set for the rank $r_j(P)$ by inverting tests of the family of two-sided hypotheses,

$$H_{j,k}\colon \Delta_{j,k}(P) = 0 \quad \text{versus} \quad K_{j,k}\colon \Delta_{j,k}(P) \neq 0 \tag{16}$$

for $(j,k) \in S_j$. A directional error occurs when the null hypothesis is rejected and $\Delta_{j,k}(P)$ is declared positive when in fact $\Delta_{j,k}(P)$ is negative; similarly, a directional error occurs if $\Delta_{j,k}(P)$ is declared negative when it is positive. By making directional claims to multiple tests of two-sided hypotheses, one is increasing the possibility of making errors and it is important to account for the possibility of such directional (or Type 3) errors. Define

$$S_j^-(P) \equiv \{(j,k) \in S_j : \Delta_{j,k}(P) \leq 0\},$$
$$S_j^+(P) \equiv \{(j,k) \in S_j : \Delta_{j,k}(P) \geq 0\},$$

which are the sets of pairs of populations whose differences are smaller/larger than or equal to zero, and

$$\begin{aligned}\text{Rej}_j^- &\equiv \{(j,k) \in S_j \colon \text{reject } H_{j,k} \text{ and claim } \Delta_{j,k}(P) < 0\},\\ \text{Rej}_j^+ &\equiv \{(j,k) \in S_j \colon \text{reject } H_{j,k} \text{ and claim } \Delta_{j,k}(P) > 0\},\end{aligned}$$

which are the sets of pairs for which a test rejects the difference being equal to zero in favor of it being, respectively, strictly smaller or larger than zero. The probability of making any mistake, either a false rejection or an incorrect determination of a sign, is

$$\text{mdFWER}_P \equiv P\left\{S_j^+(P) \cap \text{Rej}_j^- \neq \emptyset \text{ or } S_j^-(P) \cap \text{Rej}_j^+ \neq \emptyset\right\}, \tag{17}$$

which Guo and Romano (2015) and Grandhi et al. (2019) refer to as the *mixed directional familywise error rate*.

Our goal is to develop a multiple hypothesis testing procedure for (16) that controls the mdFWER and then obtain the desired two-sided confidence set for the rank $r_j(P)$ by replacing N_j^- and N_j^+ in (13) by Rej_j^- and Rej_j^+. We consider multiple hypotheses testing procedures that control the mdFWER in the sense that

$$\limsup_{n\to\infty} \sup_{P\in\mathbf{P}} \text{mdFWER}_P \leq \alpha \tag{18}$$

because the coverage probability of the resulting confidence set is bounded from below by one minus the mdFWER:

Theorem 3.2. *For any P,*

$$P\left\{|Rej_j^-| + 1 \leq r_j(P) \leq p - |Rej_j^+|\right\} \geq 1 - mdFWER_P.$$

Furthermore, if Rej_j^- and Rej_j^+ are computed by an algorithm for which (18) holds with $mdFWER_P$ as defined in (17), then the confidence set

$$R_{n,j} \equiv \left\{|Rej_j^-| + 1, \ldots, p - |Rej_j^+|\right\} \tag{19}$$

satisfies (1).

In order to implement the result in Theorem 3.2 we need to devise a procedure for testing (16) that controls the mdFWER. The only approach to controlling the mdFWER we are aware of is Bauer et al. (1986). We follow their idea and propose to test the family of one-sided hypotheses,

$$H'_{k,l} \colon \Delta_{k,l}(P) \leq 0 \quad \text{versus} \quad K'_{k,l} \colon \Delta_{k,l}(P) > 0 \tag{20}$$

for $(k,l) \in S'_j \equiv \{(k,l) \in J \times J \colon k \neq l \text{ and one of } k,l \text{ is equal to } j\}$. Note that this family of null hypotheses includes the hypotheses $\Delta_{j,k}(P) \leq 0$ and $\Delta_{k,j}(P) \leq 0$. With

$$\begin{aligned}\text{Rej}'^-_j &\equiv \{(j,k) \in S_j \colon \text{reject } H'_{k,j} \text{ and claim } \Delta_{j,k}(P) < 0\},\\ \text{Rej}'^+_j &\equiv \{(j,k) \in S_j \colon \text{reject } H'_{j,k} \text{ and claim } \Delta_{j,k}(P) > 0\},\end{aligned}$$

the *familywise error rate* for testing the family (20) can be written as

$$\begin{aligned}\text{FWER}'_P &\equiv P\left\{\text{reject at least one true hypothesis } H'_{k,l} \text{ with } (k,l) \in S'_j\right\} \\ &= P\left\{S_j^+(P) \cap \text{Rej}'^-_j \neq \emptyset \text{ or } S_j^-(P) \cap \text{Rej}'^+_j \neq \emptyset\right\}.\end{aligned}$$

Notice that the mdFWER for testing the family of two-sided hypotheses in (16) is equal to the FWER for testing the family of one-sided hypotheses in (20), i.e., $\text{mdFWER}_P = \text{FWER}'_P$. Therefore, instead of devising a procedure that satisfies (18) we could instead devise one that satisfies

$$\limsup_{n\to\infty} \sup_{P\in\mathbf{P}} \text{FWER}'_P \leq \alpha. \tag{21}$$

Consider the following simple one-step procedure. Let $C_n(1-\alpha, S'_j)$ be the one-sided confidence set in (9) with $S = S'_j$. We reject any hypothesis $H'_{k,l}$, $(k,l) \in S'_j$, for which $C_n(1-\alpha, S'_j, (k,l))$ does not contain zero and claim $\Delta_{k,l}(P) > 0$. Under suitable restrictions on $\mathbf{P}$, this approach satisfies (21), but it can be improved through a stepwise version similar to those in Romano and Wolf (2005):

Algorithm 3.1 (Stepdown Procedure)**.**

Step 0: Set $I_0 = S'_j$ and $s = 0$.

Step 1: Form the confidence set $C_n(1-\alpha, S)$ in (9) with $S = I_s$.

Step 2: Reject any $H'_{k,l}$ with $(k,l) \in I_s$ for which $0 \notin C_n(1-\alpha, I_s, (k,l))$ and claim $\Delta_{k,l}(P) > 0$.

(a) If no (further) null hypotheses are rejected, then stop.

(b) If any null hypotheses are rejected, then let $I_{s+1} \subset I_s$ denote the hypotheses that have not previously been rejected, set $s = s+1$, and return to Step 1.

Under suitable restrictions on $\mathbf{P}$, this stepwise procedure satisfies (21) when $C_n(1-\alpha, S)$ is, for example, one of the confidence sets described in Section 3.1; see Romano and Shaikh (2012). By Theorem 3.2, the confidence set

$$R_{n,j} \equiv \left\{|\text{Rej}'^-_j| + 1, \ldots, p - |\text{Rej}'^+_j|\right\},$$

where Rej'^-_j and Rej'^+_j are computed through Algorithm 3.1, therefore satisfies (1).

At this point it may be worth emphasizing again the special structure of the one-sided family of hypotheses in (20): it contains inequalities of the form $\Delta_{j,k}(P) \leq 0$ as well as the reverse $\Delta_{k,j}(P) \leq 0$. The procedure in Algorithm 3.1 treats all inequalities equally by ignoring this special structure. While this approach controls directional errors there may exist better procedures that exploit this structure in some way. How to improve upon our procedure is not obvious. For instance, a stepwise procedure that removes both hypotheses whenever one of the two is rejected, in general, does not control directional errors (Shaffer (1980)). The formulation of the inference problem in terms of controlling the mdFWER of the two-sided family clarifies that alternative procedures for controlling the mdFWER that are developed in the future may also be used for inference on ranks.

Remark 3.9. Consider the goal of constructing a one-sided confidence set for the rank $r_j(P)$ with lower endpoint equal to 1. In this case, it is not necessary to invert a two-sided family of hypotheses like (16), but

rather the one-sided family

$$H_{j,k}\colon \Delta_{j,k}(P) \leq 0 \quad \text{versus} \quad K_{j,k}\colon \Delta_{j,k}(P) > 0 \tag{22}$$

for all $(j,k) \in S_j$. In this case, there are no directional errors and the family of tests used must control the usual FWER.

A stepwise procedure can be devised through a small modification of Algorithm 3.1. Notice that the testing problem in (22) is identical to the one in (20) except that S'_j is replaced by S_j. Therefore, Algorithm 3.1, with S'_j replaced by S_j, yields $\text{Rej}'^{-}_{j} = \emptyset$ so the resulting confidence set $R_{n,j}$ in (19) has lower endpoint equal to 1 and satisfies (1). Analogously, we can construct a one-sided confidence set for the rank $r_j(P)$ with upper endpoint equal to p. ■

Remark 3.10. Similarly as in Remark 3.6 it is easy to see that the stepwise procedure in Algorithm 3.1 satisfies (14) and there exists a P such that (15) holds. Unlike for the single-step method, however, the stepwise procedure achieves the equality in (15) at distributions P for which there is at least one (not necessarily all) $k \neq j$ with $\theta(P_k) = \theta(P_j)$. ■

3.2.3 Discussion of Optimality

While the focus of this paper is the development of methods that are easily applied, are computationally viable, and possess proven coverage properties, the following discussion of optimality may be helpful (though a deeper analysis is left for future work). Our methodology allows one to apply any multiple testing procedure for jointly testing differences that controls the FWER or mdFWER, depending on whether the testing problem is one- or two-sided. Therefore, one certainly should apply multiple testing procedures that are ideally optimal in some sense. Although little is known about optimality in multiple testing, some known results support our choices.

To be concrete, consider the multiple testing problem

$$\Delta_{j,k}(P) \leq 0 \quad \text{versus} \quad \Delta_{j,k}(P) > 0$$

for a fixed population $j \in J$ and all $(j,k) \in S_j$. The asymptotic version of this testing problem corresponds to the situation where we observe independent draws Z_j with $Z_j \sim N(\theta(P_j), \sigma^2(P_j))$, with $\sigma^2(P_j)$ known. In certain parametric settings, Lehmann et al. (2005) derive optimal multiple testing procedures for such one-sided testing problems; the resulting procedure rejects based on the corresponding maximum statistic, analogous to our procedure. Moreover, more refined optimality criteria lead to stepdown improvements, of the form also considered here. The only caveat to note is that a monotonicity assumption is required to hold (such as in the normal limit problem), because for such (moment inequality) testing problems, considerations such as unbiasedness and invariance do not apply. Also, the results do not apply to the two-sided testing problem considered in (16). In this situation, there remain open questions concerning Type 1 error (or mdFWER) control, and optimality considerations remain even harder. On the other hand, if the goal is the construction of a two-sided $1-\alpha$ confidence interval, then it makes good sense that each endpoint should be a $1-\alpha/2$ confidence bound; that is, the chance that the true parameter lies above the upper endpoint or below the lower endpoint should be no bigger than $\alpha/2$; in other words, the interval should be equi-tailed. With this restriction, the methodology for each endpoint can be obtained through one-sided tests.

An alternative approach for both one-sided and two-sided multiple testing problems is developed in Spjøtvoll (1972), based on the criteria of expected number of rejections (and not the FWER). Note that if the expected number of false rejections is bounded by α, then so is the FWER (by Markov's inequality). In the normal problem, optimal procedures are derived; see Examples 1 and 3 in Section 3 of that paper. The important point is that the form of the optimal procedures rejects each hypothesis based on the normalized value of $Z_j - Z_k$; the asymptotic version is analogous to our bootstrap procedure. In other words, the test statistics for the individual tests are used. The only difference is that critical values are obtained under a slightly different criterion, leading to the conservative Bonferroni critical values. Our procedure uses less conservative critical values based on the bootstrap's estimated distribution of the maximum, even though the FWER is controlled if the expected number of false rejections is controlled. In any case, the critical values are asymptotically quite similar.

In summary, the above approaches to optimality are quite distinct, but lead to procedures that are indeed quite analogous to the ones recommended here. However, there is one remaining gap in the logic that deserves further attention. It may be possible to construct confidence intervals for ranks without having to explicitly make decisions about each of the differences $\Delta_{j,k}(P)$. While it may not seem natural, it leaves open the possibility of improved methods. On the other hand, consider a confidence interval for $r_j(P)$. The construction we propose yields a confidence interval for $r_j(P)$, as well as decisions about which populations have their corresponding $\theta(P_k)$ above (or below, or both) that of $\theta(P_j)$. Moreover, the additional claims can be made with no added contribution to Type 1 error rates. Such information can be quite useful. For example, if country j is ranked below specified countries in terms of academic achievement, country j could then use this information to consider why these countries have better performance. If we require that any procedure be able to provide such information as well, then any procedure must be derived by some multiple testing procedure, thereby closing the gap, and attention may then be restricted to constructions based on multiple testing procedures.

3.3 Joint Confidence Sets for the Ranks of All Populations

In this section, we show how arguments similar to those in Sections 3.2.1 and 3.2.2 can be used to construct simultaneous confidence sets for the ranks of all populations.

3.3.1 A Simple Construction

Define $S_{\text{all}} \equiv \{(j,k) \in J \times J : j \neq k\}$. Let $C_n(1-\alpha, S_{\text{all}})$ be a confidence region for $\Delta_{S_{\text{all}}}(P)$ that is rectangular in the sense of (4) with $\{C_n(1-\alpha, S_{\text{all}}, (j,k)) : (j,k) \in S_{\text{all}}\}$. Similarly to the definitions of N_j^- and N_j^+, for each $j \in J$, denote by

$$N_{j,\text{all}}^- \equiv \{k \in J \setminus \{j\} \colon C_n(1-\alpha, S_{\text{all}}, (j,k)) \subseteq \mathbf{R}_-\},$$
$$N_{j,\text{all}}^+ \equiv \{k \in J \setminus \{j\} \colon C_n(1-\alpha, S_{\text{all}}, (j,k)) \subseteq \mathbf{R}_+\}$$

the sets of confidence sets for the differences $\Delta_{S_{\text{all}}}(P)$ that lie entirely below and above zero. The set $N_{j,\text{all}}^-$ ($N_{j,\text{all}}^+$) therefore contains all populations k whose features $\theta(P_k)$ are significantly larger (smaller) than that of population j. The following result is analogous to Theorem 3.1:

Theorem 3.3. *If $C_n(1-\alpha, S)$ satisfies (4) with $S = S_{\text{all}}$, then, for any P,*

$$P\left\{\bigcap_{j\in J}\left\{|N^-_{j,\text{all}}|+1 \le r_j(P) \le p - |N^+_{j,\text{all}}|\right\}\right\} \ge P\{\Delta_{S_{\text{all}}}(P) \in C_n(1-\alpha, S_{\text{all}})\}.$$

If, in addition, $C_n(1-\alpha, S)$ also satisfies (5) with $S = S_{\text{all}}$, then the confidence set

$$R^{\text{joint}}_n \equiv \prod_{j\in J}\left\{|N^-_{j,\text{all}}|+1, \ldots, p - |N^+_{j,\text{all}}|\right\} \tag{23}$$

satisfies (2).

Remarks similar to those after Theorem 3.1 also apply to Theorem 3.3.

Remark 3.11. An alternative approach to constructing a confidence set that satisfies (2) is based on simultaneous confidence sets for the features $\theta(P)$ rather than for their pairwise differences $\Delta_{S_{\text{all}}}(P)$. The recent paper by Klein et al. (2020) is a special case of this approach. In Appendix B, we prove that, in some special cases, the resulting confidence set for the ranking is strictly larger than our proposal in (23). In addition, in our simulations in Section 4 and Appendix F, we find that their confidence set is always at least as large as ours, but in most cases substantially larger.

An intuitive explanation for these findings is that the rank of a population depends on the features $\theta(P)$ only through their pairwise differences. Although a $1-\alpha$ simultaneous confidence set for the features $\theta(P)$ can be transformed into a $1-\alpha$ simultaneous confidence set for the vector of pairwise differences $\Delta_{S_{\text{all}}}(P)$, we show in Appendix B that such a construction covers the true differences with probability strictly larger than $1-\alpha$. In this sense, the simultaneous confidence set for the differences is conservative. In contrast, our proposal is based on a nonconservative simultaneous confidence set for the vector of pairwise differences (in the sense that the coverage probability is equal to $1-\alpha$ for some data-generating processes), which is weakly shorter for all data-generating processes and strictly shorter for some. The improvement in size of the simultaneous confidence set for the pairwise differences then translates into an improvement in size of the confidence set for the rank. ■

3.3.2 A Stepwise Improvement

Consider the goal of constructing a two-sided confidence set for all ranks. In order to describe a way in which we can improve upon Theorem 3.3 consider the problem of testing (16) for all $(j,k) \in S_{\text{all}}$. Define

$$S^-_{\text{all}}(P) \equiv \{(j,k) \in S_{\text{all}} : \Delta_{j,k}(P) \le 0\}\ ,$$
$$S^+_{\text{all}}(P) \equiv \{(j,k) \in S_{\text{all}} : \Delta_{j,k}(P) \ge 0\}$$

and let $\text{Rej}^-_{j,\text{all}} \equiv \{k \in J\colon (j,k) \in \text{Rej}^-_{\text{all}}\}$ and $\text{Rej}^+_{j,\text{all}} \equiv \{k \in J\colon (j,k) \in \text{Rej}^+_{\text{all}}\}$ with

$$\text{Rej}^-_{\text{all}} \equiv \{(j,k) \in S_{\text{all}}\colon \text{reject } H_{j,k} \text{ and claim } \Delta_{j,k}(P) < 0\}\ ,$$
$$\text{Rej}^+_{\text{all}} \equiv \{(j,k) \in S_{\text{all}}\colon \text{reject } H_{j,k} \text{ and claim } \Delta_{j,k}(P) > 0\}\ .$$

The mixed directional familywise error rate for the problem of testing (16) for all $(j,k) \in S_{\text{all}}$ is then

$$\text{mdFWER}_P \equiv P\left\{S^+_{\text{all}}(P) \cap \text{Rej}^-_{\text{all}} \neq \emptyset \text{ or } S^-_{\text{all}}(P) \cap \text{Rej}^+_{\text{all}} \neq \emptyset\right\}. \tag{24}$$

The following result is analogous to Theorem 3.2.

Theorem 3.4. *For any P,*

$$P\left\{\bigcap_{j\in J}\left\{|Rej^-_{j,\text{all}}| + 1 \leq r_j(P) \leq p - |Rej^+_{j,\text{all}}|\right\}\right\} \geq 1 - mdFWER_P\ .$$

Furthermore, if $Rej^-_{j,\text{all}}$ and $Rej^+_{j,\text{all}}$ are computed by an algorithm for which (18) holds with $mdFWER_P$ as defined in (24), then the confidence set

$$R^{\text{joint}}_n \equiv \prod_{j\in J}\left\{|Rej^-_{j,\text{all}}| + 1, \ldots, p - |Rej^+_{j,\text{all}}|\right\} \tag{25}$$

satisfies (2).

In order to implement the result in Theorem 3.4 we need to devise a procedure that controls the mdFWER. As for the marginal confidence sets, we can control the mdFWER for the two-sided testing problem by controlling the FWER,

$$\text{FWER}'_P \equiv P\{\text{reject at least one true hypothesis } H'_{k,l} \text{ with } (k,l) \in S_{\text{all}}\}\ , \tag{26}$$

for the one-sided testing problem (20) with $(k,l) \in S_{\text{all}}$.

Consider the following simple one-step procedure. Let $C_n(1-\alpha, S_{\text{all}})$ be the one-sided confidence set in (9) with $S = S_{\text{all}}$. We reject any hypothesis $H'_{k,l}$, $(k,l) \in S_{\text{all}}$, for which $C_n(1-\alpha, S_{\text{all}}, (k,l))$ does not contain zero and claim $\Delta_{k,l}(P) > 0$. Under suitable restrictions on $\mathbf{P}$, this approach satisfies (21) with FWER'_P as defined in (26), but it can be improved through a stepwise version similar to those in Romano and Wolf (2005):

Algorithm 3.2 (Stepdown Procedure)**.**

Step 0: Set $I_0 = S_{\text{all}}$ and $s = 0$.

Step 1: Form the confidence set $C_n(1-\alpha, S)$ in (9) with $S = I_s$.

Step 2: Reject any $H'_{k,l}$ with $(k,l) \in I_s$ for which $0 \notin C_n(1-\alpha, I_s, (k,l))$ and claim $\Delta_{k,l}(P) > 0$.

(a) If no (further) null hypotheses are rejected, then stop.

(b) If any null hypotheses are rejected, then let $I_{s+1} \subset I_s$ denote the hypotheses that have not previously been rejected, set $s = s+1$, and return to Step 1.

Under suitable restrictions on $\mathbf{P}$, this stepwise procedure satisfies (21) with FWER'_P as defined in (26) when $C_n(1-\alpha, S)$ is, for example, one of the confidence sets described in Section 3.1; see Romano and

Shaikh (2012). By Theorem 3.4, the confidence set

$$R_n^{\text{joint}} \equiv \prod_{j \in J} \left\{ |\text{Rej}_{\text{all}}^-| + 1, \ldots, p - |\text{Rej}_{\text{all}}^+| \right\} \ ,$$

where $\text{Rej}_{\text{all}}^-$ and $\text{Rej}_{\text{all}}^+$ are computed through Algorithm 3.2, therefore satisfies (2).

3.4 Confidence Sets for the τ-Best Populations

The goal of this section is to construct confidence sets for the τ-best populations, i.e., for given values of $\tau \in J$ and $\alpha \in (0,1)$, we want to construct (random) sets $J_n^{\tau-\text{best}}$ that satisfy (3).

Given a confidence set $R_n^{\text{joint}} \equiv \prod_{j \in J} R_{n,j}^{\text{joint}}$ that satisfies (2), such as those in (23) and (25), it is straightforward to construct $J_n^{\tau-\text{best}}$ satisfying (3) by defining

$$J_n^{\tau-\text{best}} \equiv \left\{ j \in J : \tau \in R_{n,j}^{\text{joint}} \right\} \ . \tag{27}$$

In this section, however, we propose a more "direct" approach which, in simulations, we have found to perform better than the naive projection in (27). For a given value of $\tau \in J$ and some $j \in J$, consider the hypothesis

$$H_j \colon r_j(P) \leq \tau.$$

Let π be a permutation of J such that $\theta(P_{\pi(1)}) \geq \theta(P_{\pi(2)}) \geq \ldots \geq \theta(P_{\pi(p)})$ and define $\mathcal{K} \equiv \{K \subset J \colon |K| = \tau - 1\}$ to be the set of all subsets of J with cardinality $\tau - 1$ (i.e., $\mathcal{K} = \{\emptyset\}$ when $\tau = 1$). The null hypothesis H_j is equivalent to

$$\max_{k \in J \setminus \{\pi(1), \ldots, \pi(\tau-1)\}} \{\theta(P_k) - \theta(P_j)\} \leq 0$$

and implies

$$\min_{K \in \mathcal{K}} \max_{k \in J \setminus K} \{\theta(P_k) - \theta(P_j)\} \leq 0.$$

In order to form a test statistic for this inequality, we replace the features $\theta(P_j)$ by their estimators:

$$T_{n,j} \equiv \min_{K \in \mathcal{K}} \max_{k \in J \setminus K} \{\hat{\theta}_k - \hat{\theta}_j\}. \tag{28}$$

Further, for $I \subseteq J$ and $K \in \mathcal{K}$, let

$$T_{n,I,K} \equiv \max_{j \in I} \max_{k \in J \setminus K} \{\hat{\theta}_k - \hat{\theta}_j - \Delta_{k,j}(P)\}$$

and denote by $M_n(x, I, K, P) \equiv P\{T_{n,I,K} \leq x\}$ the cdf of $T_{n,I,K}$. Finally, define the critical value

$$\hat{c}_n(1-\alpha, I) \equiv \max_{K \in \mathcal{K}} M_n^{-1}(1-\alpha, I, K, \hat{P}_n) \tag{29}$$

for some estimate $\hat{P}_n$ of P. The following algorithm is a stepwise procedure for testing the family of null hypotheses H_j with $j \in J$.

Algorithm 3.3.

Step 0: Set $I_0 = J$ and $s = 0$.

Step 1: Reject any H_j with $j \in I_s$ for which $T_{n,j} > \hat{c}_n(1-\alpha, I_s)$.

(a) If no (further) null hypotheses are rejected, then stop.

(b) If any null hypotheses are rejected, then let $I_{s+1} \subset I_s$ denote the hypotheses that have not previously been rejected, set $s = s+1$, and repeat Step 1.

The confidence set for the τ-best populations can then be defined as all those $j \in J$ for which H_j is not rejected by Algorithm 3.3.

Theorem 3.5. *Assume that, for each* $K \in \mathcal{K}$,

$$\liminf_{n\to\infty} \inf_{P\in\mathbf{P}} P\left\{T_{n,J_0^{\tau-\text{best}}(P),K} \le M_n^{-1}(1-\alpha, J_0^{\tau-\text{best}}(P), K, \hat{P}_n)\right\} \ge 1-\alpha \ . \tag{30}$$

Then the confidence set

$$J_n^{\tau-\text{best}} \equiv \{j \in J\colon H_j \text{ is not rejected } \} \ ,$$

computed through Algorithm 3.3, satisfies (3).

Under a uniform integrability condition on $\mathbf{P}$, the uniform asymptotic coverage requirement (30) holds for various choices of $\hat{P}_n$; see Romano and Shaikh (2012).

Remark 3.12. One could replace $T_{n,j}$ by a studentized version of the statistic,

$$T_{n,j} \equiv \min_{K\in\mathcal{K}} \max_{k\in J\setminus K} \frac{\hat{\theta}_k - \hat{\theta}_j}{\hat{\sigma}_{k,j}} \ ,$$

where $\hat{\sigma}^2_{k,j}$ is an estimator of the variance of $\hat{\theta}_k - \hat{\theta}_j$, and modify $M_n(x, I, K, P)$ to be the distribution of

$$T_{n,I,K} \equiv \max_{j\in I} \max_{k\in J\setminus K} \frac{\hat{\theta}_k - \hat{\theta}_j - \Delta_{k,j}(P)}{\hat{\sigma}_{k,j}} \ .$$

Studentization may be especially desirable when the distributions of $\hat{\theta}_j$ vary considerably. ■

Remark 3.13. The computation of the critical value $\hat{c}_n(1-\alpha, I)$ involves the maximization of $M_n^{-1}(1-\alpha, I, K, \hat{P}_n)$ over $K \in \mathcal{K}$. For instance, when $\tau = 2$, then $\mathcal{K} = \{\{1\}, \{2\}, \ldots, \{p\}\}$. For $\tau > 1$, there are $\binom{p}{\tau-1}$ elements in $\mathcal{K}$, so the construction of the critical value becomes computationally more demanding the larger τ. There are, however, at least two special cases in which the optimization becomes trivial. First of all, to form a confidence set for the best population ($\tau = 1$), no optimization is necessary because in this case $\mathcal{K} = \{\emptyset\}$. Second, suppose $\hat{\theta}_1 - \theta(P_1), \ldots, \hat{\theta}_p - \theta(P_p)$ are exchangeable. In this case, one can show that $M_n(1-\alpha, I, K, \hat{P}_n)$ is independent of K, so the the computation of the critical value $\hat{c}_n(1-\alpha, I)$ does not require optimization over $K \in \mathcal{K}$ regardless of the value of τ. ■

Remark 3.14. An alternative approach to computing critical values for the test statistic in (28) is based on an estimate of the set $J \setminus \{\pi(1), \ldots, \pi(\tau-1)\}$. To see this first notice that (28) is bounded above by $\max_{k\in K}\{\hat{\theta}_k - \hat{\theta}_j - \Delta_{k,j}(P)\}$ with $K = J \setminus \{\pi(1), \ldots, \pi(\tau-1)\}$. Letting $S_n(\cdot, K, P)$ denote the cdf of $\max_{k\in K}\{\hat{\theta}_k - \hat{\theta}_j - \Delta_{k,j}(P)\}$ for some set $K \subseteq J$, an infeasible critical value for (28) is therefore given by

$S_n^{-1}(1-\alpha, J \setminus \{\pi(1), \ldots, \pi(\tau-1)\}, \hat{P}_n)$ with $\hat{P}_n$ an estimate of P. To obtain a feasible counterpart, one could replace the set $J \setminus \{\pi(1), \ldots, \pi(\tau-1)\}$ by an estimate as follows.

Suppose $\hat{\theta}$ is $\sqrt{n}$-consistent. Let $\hat{\pi}_n$ be a permutation of J such that $\hat{\theta}_{\hat{\pi}_n(1)} \geq \ldots \geq \hat{\theta}_{\hat{\pi}_n(p)}$ and define

$$\hat{K}_n \equiv \left\{ j \in J \colon \hat{\theta}_j \leq \hat{\theta}_{\hat{\pi}_n(\tau)} + \epsilon_n \right\},$$

where $\{\epsilon_n\}_{n \geq 1}$ is a sequence of positive constants so that $\epsilon_n \to 0$ and $\epsilon_n \sqrt{n} \to \infty$. Then, it is easy to see that the estimated set $\hat{K}_n$ contains $J \setminus \{\pi(1), \ldots, \pi(\tau-1)\}$ with probability approaching one. Since $S_n^{-1}(1-\alpha, K, \hat{P}_n)$ is monotone with respect to K, a feasible critical value is therefore given by $S_n^{-1}(1-\alpha, \hat{K}_n, \hat{P}_n)$. An obvious challenge in the implementation of this construction is the choice of ϵ_n (a problem that is similar to moment selection), but an advantage is that it remains computationally feasible even for large values of τ. Whether or not the alternative critical value is smaller or larger than that in (29) is not obvious to us and thus left for future research.

It may also be possible to employ a two-step method as in Romano et al. (2014) by finding $\hat{K}_n$ that contains $J \setminus \{\pi(1), \ldots, \pi(\tau-1)\}$ with probability approaching $1-\beta$ for $0 < \beta < \alpha$ and using the critical value $S_n^{-1}(1-\alpha+\beta, \hat{K}_n, \hat{P}_n)$ instead. ■

Remark 3.15. The τ-worst populations in terms of $\theta_1(P), \ldots, \theta_p(P)$ are also the τ-best populations in terms of $-\theta_1(P), \ldots, -\theta_p(P)$. Therefore, the procedure described above can be used for the construction of a confidence set for the τ-worst populations by simply changing the signs of the features $\theta(P_j)$ and their estimators. ■

Remark 3.16. Similarly to the reasoning in Remark 3.1 there are no conceptual challenges in satisfying (30) while allowing for dependence in the estimators $\hat{\theta}_1, \ldots, \hat{\theta}_p$. ■

Remark 3.17. The problem of finding a subset $J_n^{\tau\text{-best}}$ satisfying (3) is related to the subset selection problem in the PhD thesis by Gupta (1956). He assumed that $\hat{\theta}_j$ and $\theta(P_j)$ are the sample and population mean, respectively, and are such that $\hat{\theta}_1 - \theta(P_1), \ldots, \hat{\theta}_p - \theta(P_p)$ are i.i.d. from a normal distribution with known variance. For this case, he proposed a confidence set for the identity of the population with the largest mean. Many extensions of Gupta's idea have been proposed since then; see Gupta and Panchapakesan (1979, Chapters 11–19) for a review. Appendix C shows that this approach only guarantees coverage of one, but not necessarily all, of the best populations in case there are ties. For example, if there are only two populations ($p = 2$) and their means are equal, then the probability of Gupta's confidence set covering $J_0^{1\text{-best}}$ is strictly less than the desired level $1-\alpha$. In contrast, the confidence set proposed in this section asymptotically covers $J_0^{\tau\text{-best}}$ for any $\tau \in J$ with probability no less than the desired level. Importantly, unlike Gupta's confidence set, our proposal does not rely on his i.i.d. assumption. Allowing for heterogeneity in the populations' distributions is important in our empirical applications in which the populations' variances differ substantially across populations. ■

4 Examining the Finite-Sample Performance through Simulations

In this section, we examine the finite-sample performance of several procedures for constructing confidence sets for the rank of a single population, for the ranks of all populations, and for the set of τ-best populations with a simulation study.

Data generating process

The data generating process is calibrated to the data from the empirical application in Section 5.2 in which we rank neighborhoods in the U.S. by measures of intergenerational mobility. Let $\hat{\theta}_C \equiv (\hat{\theta}_{C,1}, \ldots, \hat{\theta}_{C,p})$ and $\hat{se}_C \equiv (\hat{se}_{C,1}, \ldots, \hat{se}_{C,p})$ denote the "correlational" estimates of mobility and their standard errors for the p most populous commuting zones (CZs) from Chetty et al. (2018). Similarly, denote by $\hat{\theta}_M$ and $\hat{se}_M$ the "movers" estimates of mobility and their standard errors for the p most populous CZs from Chetty and Hendren (2018). Let $\hat{n} \equiv (\hat{n}_1, \ldots, \hat{n}_p)$ be the vector with j-th element equal to the number of individuals either moving from or to CZ j in the movers dataset (i.e. it is an estimate of the number of observations used in estimation of the movers estimate for CZ j).

For each CZ $j = 1, \ldots, p$, we generate an i.i.d. sample $X_{j,1}, \ldots, X_{j,n_j}$ with sample size $n_j \equiv \kappa \hat{n}_j$ and

$$X_{j,i} \sim N(\theta_j, \sigma_j^2), \text{ independently across } j,$$

where, for $t \in \{C, M\}$ (i.e. depending on which of the two datasets we base the simulations on),

$$\theta_j \equiv \begin{cases} \hat{\theta}_{t,1}, & j = 1 \\ \hat{\theta}_{t,1} + \delta \sum_{k=1}^{j-1} (\hat{\theta}_{t,k+1} - \hat{\theta}_{t,k}), & j > 1 \end{cases} \tag{31}$$

and $\sigma_j^2 \equiv \hat{n}_j \hat{se}_{t,j}^2$. There are two main parameters that govern the design, δ and κ.

The parameter δ governs how close the mobility measures θ_j are to each other. When $\delta = 1$, then they are equal to those from the data. For smaller (larger) values, the measures are closer to (further away from) each other than in the data and, in the extreme case of $\delta = 0$, they are all identical.

The parameter κ governs the sizes of the samples drawn for each CZ. $\kappa = 1$ corresponds to the case in which the CZ-specific sample sizes are the same as in the data. Smaller (larger) κ means we draw a smaller (larger) sample than in the data or, equivalently, the standard errors for the resulting mobility estimates are larger (smaller).

Figures 4 and 5 show the data generating processes (more precisely, the configuration of mobility measures and standard errors) for different parameter values of κ and δ. In each figure, the panel in the center depicts the data generating process calibrated to the data. The panels to its left and right show how changing δ affects the slope of the vector of mobility measures. The panels above and below show how changing κ changes the standard errors of the mobility estimates. All simulations are based on $2,000$ Monte Carlo samples and nominal coverage of 95%.

Simulation exercises

We consider three inference problems: inference on the rank of a single CZ, simultaneous inference on the ranks of all CZs, and inference on the set of τ-best CZs with $\tau = 2$. In this section, we only present and discuss the results for inference on a single rank, the other results can be found in Appendix F. We perform inference on the CZ j with the largest mobility estimate $\hat{\theta}_{t,j}$ among the p most populous CZs.

We consider different procedures for constructing the confidence sets. Critical values for the different methods are based on the parametric bootstrap so as to mimic the empirical analysis in Section 5.2, in which

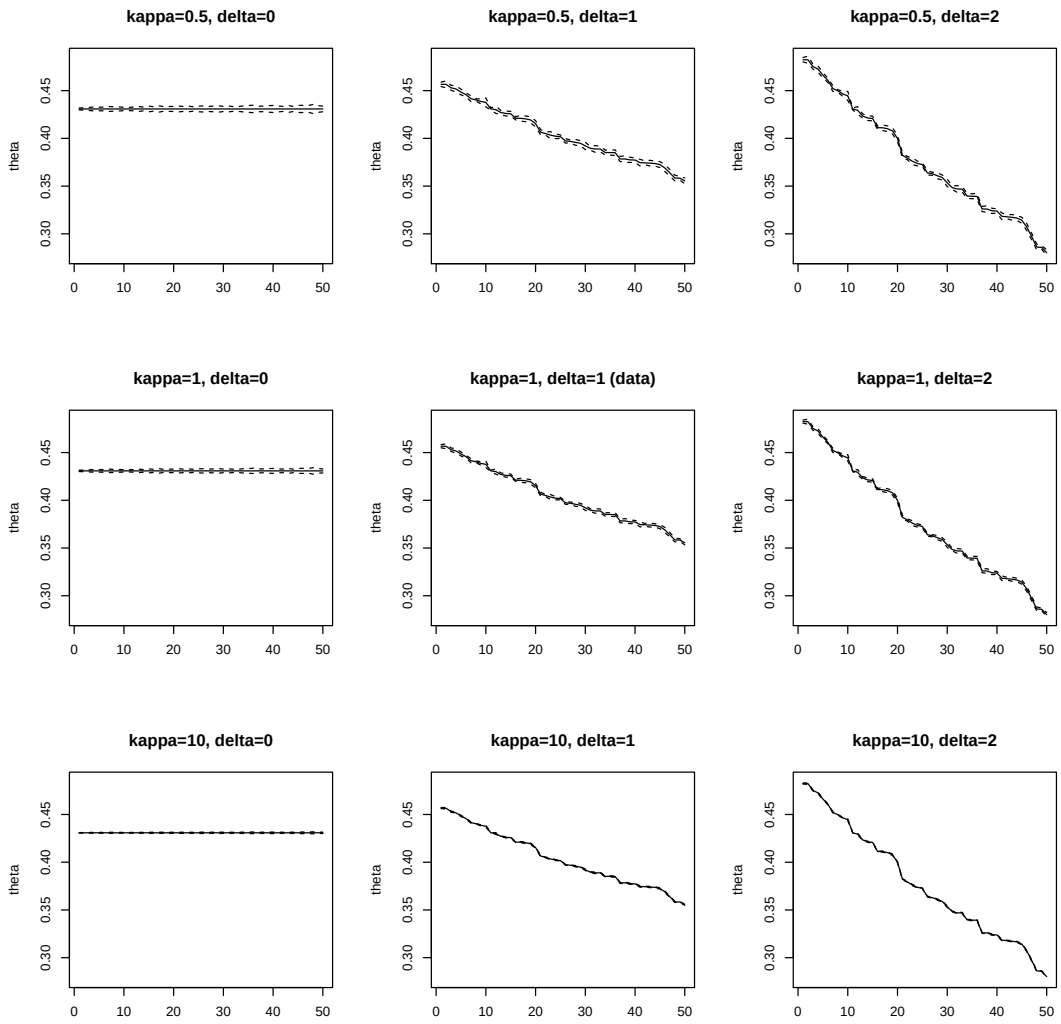

Figure 4: **Correlational simulation design:** for $j = 1, \ldots, 50$, the panels show θ_j (solid lines) and $\theta_j \pm 2\sigma_j/\sqrt{n_j}$ (dashed lines), varying κ and δ. The panel denoted by "(data)" shows the design calibrated to the data.

we only observe point estimates and standard errors so a nonparametric bootstrap cannot be implemented.[2] Specifically, for a given Monte Carlo sample $X_{j,1}, \ldots, X_{j,n_j}$, we compute the sample average $\hat{\theta}_j$, the sample variance $\hat{\sigma}_j^2$, the standard error $\hat{se}_j \equiv \hat{\sigma}_j/\sqrt{n_j}$, and $\hat{\sigma}_{j,k}^2 \equiv \hat{\sigma}_j^2 + \hat{\sigma}_k^2$. Then, we generate $1,000$ draws of normal random vectors $Z \equiv (Z_1, \ldots, Z_p)' \sim N(0, \text{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2))$. We compare the following methods, setting $S = S_j$:

"DM": the simple construction in (13), based on symmetric confidence sets for the differences in means as in (11), where $L_{\text{symm,n}}^{-1}(1-\alpha, S, \hat{P}_n)$ is the empirical $(1-\alpha)$-quantile of the $1,000$ draws of $\max_{(j,k)\in S} |Z_j - Z_k|/\hat{\sigma}_{j,k}$.

"DM.step": the stepwise constructions (19) computed through Algorithm 3.1, based on confidence sets

[2]Simulation results for the nonparametric bootstrap, which are not reported here, are very similar to the results for the parametric bootstrap.

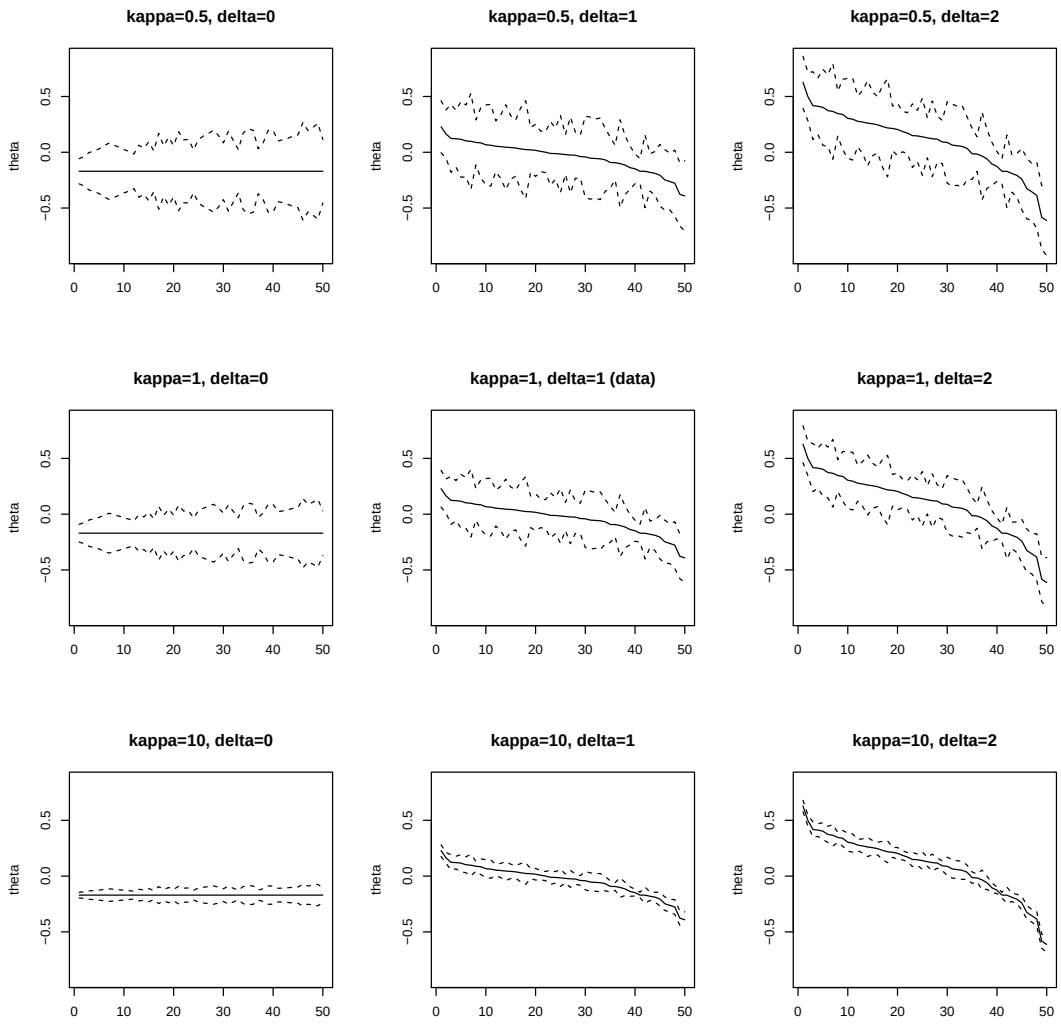

Figure 5: **Movers simulation design:** for $j = 1, \ldots, 50$, the panels show θ_j (solid lines) and $\theta_j \pm 2\sigma_j/\sqrt{n_j}$ (dashed lines), varying κ and δ. The panel denoted by "(data)" shows the design calibrated to the data.

for the differences in means as in (9), where, in the sth step, $L^{-1}_{\text{lower,n}}(1-\alpha, I_s, \hat{P}_n)$ is the empirical $(1-\alpha)$-quantile of the 1,000 draws of $\max_{(j,k)\in I_s}(Z_j - Z_k)/\hat{\sigma}_{j,k}$.

"M": the alternative confidence set described in Appendix B, based on symmetric confidence sets for the means as in (39), where $\widetilde{q}_{1-\alpha}$ is the empirical $(1-\alpha)$-quantile of the 1,000 draws of $\max_{j\in J}|Z_j|/\hat{\sigma}_j$.[3]

Results

Table 2 shows coverage frequencies, where coverage is computed as in Remark 3.6, i.e. of the set of ranks. Figures 6–7 plot the "relative" length of the marginal confidence sets, which is computed as the length of the

[3]Using $\widetilde{q}_{1-\alpha}$ as defined in (40) or (41) yields almost identical results, but we choose to use bootstrap quantiles here to make the method more similar to our proposals "DM" and "DM.step", which also use bootstrap quantiles.

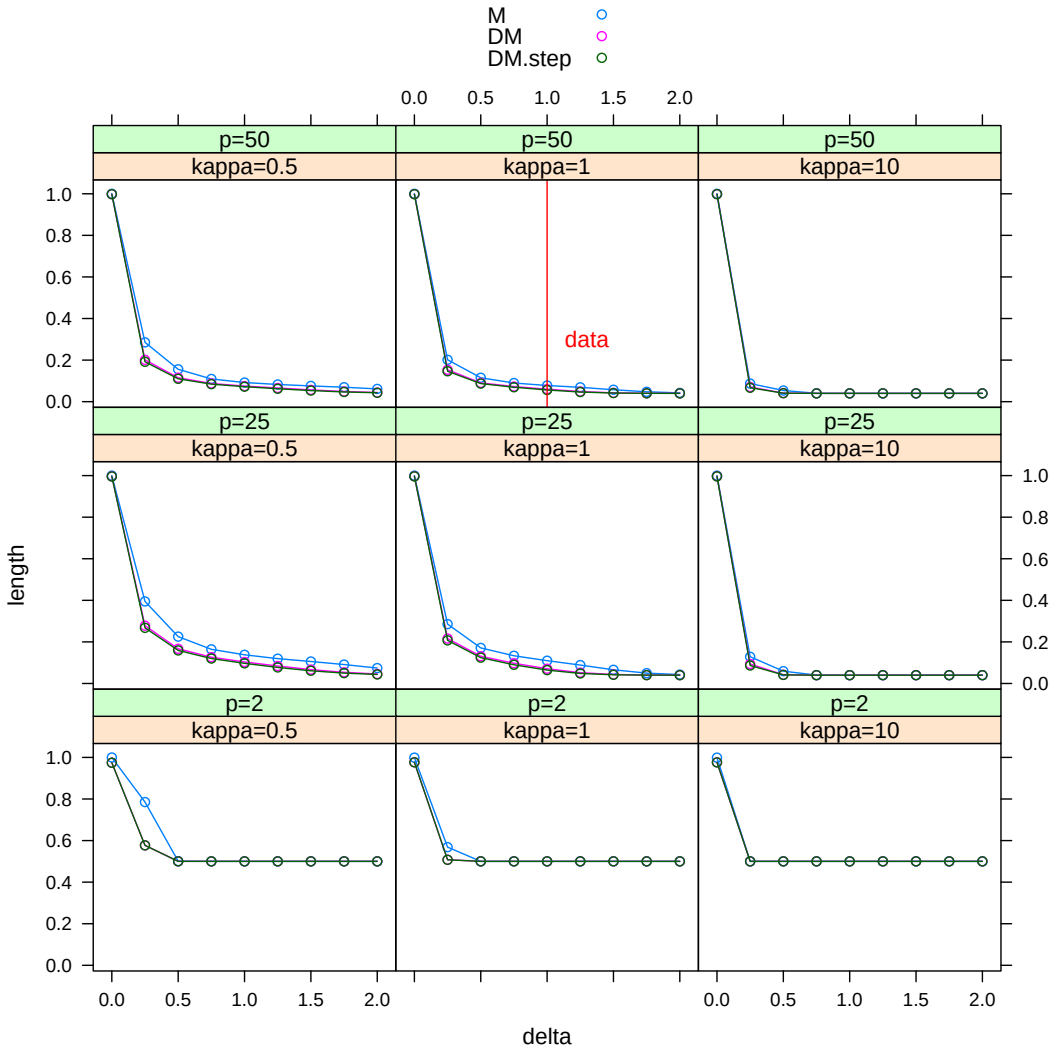

Figure 6: Marginal confidence set for the rank of a single CZ: relative length for the correlational design

confidence set averaged over the Monte Carlo samples and divided by the number of CZs p. For example, a relative length equal to one (or 0.4) means that all (or 40% of the) CZs are included in the confidence set.

We obtain five insights from the simulation results for inference on a single rank. First, Table 2 shows that all methods control the coverage frequency at the desired nominal level for small and large sample sizes, regardless of whether mobility measures are well separated (δ large), nearly tied ($\delta \approx 0$), or tied ($\delta = 0$), and regardless of whether there are few or many CZs to be ranked (p small or large). The only instances of a small amount of undercoverage occur when the mobility measures are all equal ($\delta = 0$) and, at the same time, either the sample size is small or the number of CZs is large.

Second, the coverage frequency of "M" is approximately equal to one in all scenarios whereas our methods "DM" and "DM.step" tend to have coverage frequency closer to the desired level. In consequence, our methods tend to lead to confidence sets for the ranks that are not larger than those of "M" and substantially smaller in many scenarios. For instance, Figure 7 with $p = 25$ and $\kappa = 10$ shows that the confidence set of

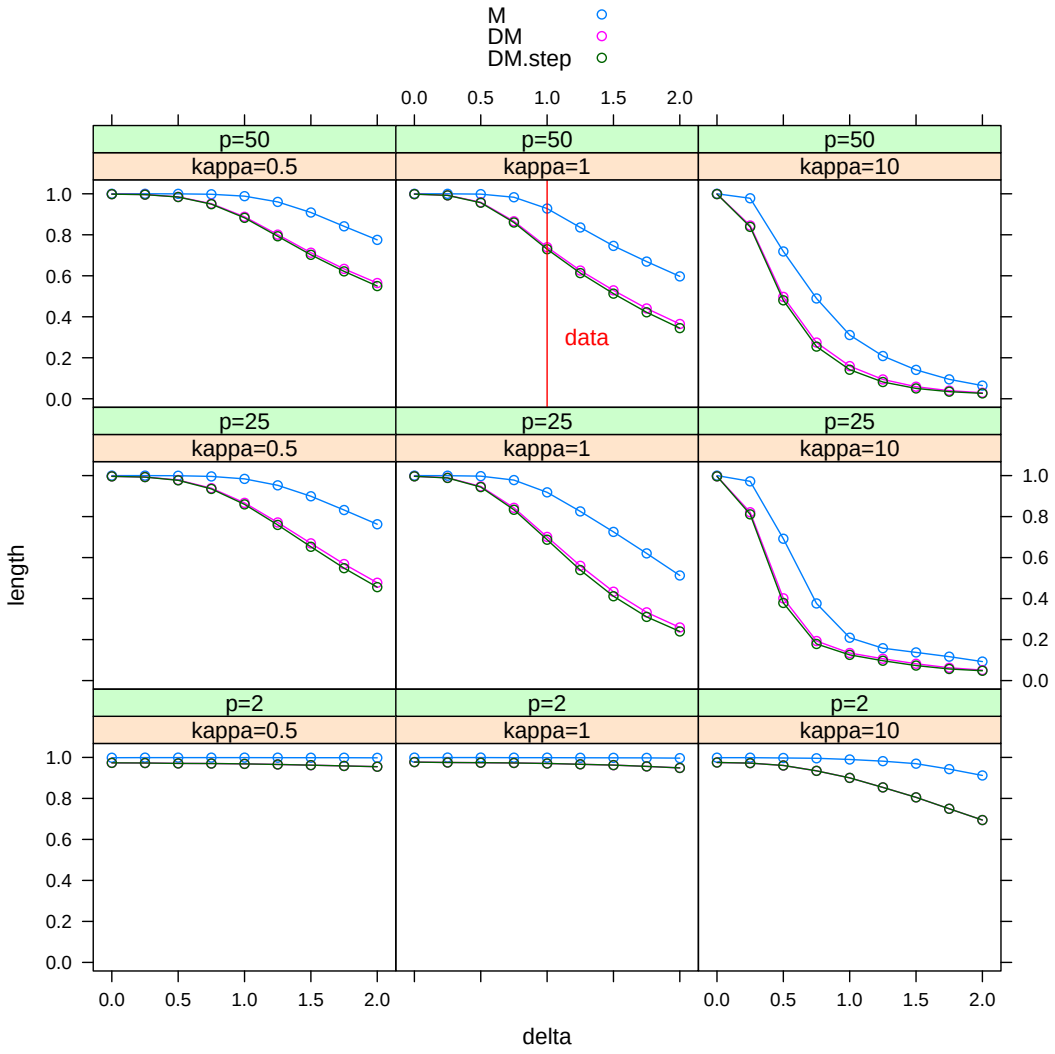

Figure 7: Marginal confidence set for the rank of a single CZ: relative length for the movers design

"M" maybe almost twice as large as those of "DM" and "DM.step". Improvements in the size of confidence sets by using the stepwise method ("DM.step") instead of the single-step method ("DM") are small, but in Appendix F we provide a different simulation design in which the stepwise method leads to significant reductions of length.

Third, Figures 6 and 7 show that all methods produce confidence sets that decrease in size as the mobility measures become more separated from each other (δ increases). When $\delta = 0$, then all CZs have equal mobility measures and the confidence sets for a the rank of a single CZ include all CZs, i.e. relative length is equal to one. This is expected because, in this case, all CZs are tied at rank equal to one. As δ increases the mobility measure of the CZ of interest becomes well separated from all other CZs. The length of the confidence set decreases towards $1/p$ and the coverage frequency increases towards 1. Provided the sample size is not too small, this behavior is a consequence of the confidence sets for the differences being consistent (see Remark 3.4).

Fourth, comparing left and right columns of Figures 6 and 7 shows that, as the sample size increases and as long as there are no ties, the length of the confidence sets decreases and the coverage frequency increases. This behavior is a consequence of the fact that a large sample size leads to small variances of the estimated mobility measures so that, by the consistency of the confidence sets for the differences, true differences in mobility measures are easier to detect (see Remark 3.4).

Finally, the differences between the correlational and movers designs have substantial impact on the length of the confidence sets. Consider the second panels in the top rows of Figures 6 and 7. The red vertical lines at $\delta = 1$ mark the data-generating process that is calibrated to the data. In the correlational design, our confidence sets "DM" and "DM.step" are relatively small (both containing on average 6% of the 50 CZs) whereas in the movers design the confidence sets are very large (containing on average more than 70% of the 50 CZs, respectively). This finding is not surprising because, in the movers design, standard errors are much larger and the mobility measures less well-separated than in the correlational design (compare the middle panels in Figures 4 and 5). To illustrate the magnitude of the statistical noise in the movers design notice that even a 10-fold increase in the sample size ($\kappa = 10$) and its implied $\sqrt{10} \approx 3.16$-fold decrease in the standard errors reduces the size of the confidence sets only to about 16% ("DM") and 14% ("DM.step") of the 50 CZs (i.e. 8 and 7 of the 50 CZs). In contrast, in the correlational design, the same increase in sample size reduces the size of both confidence sets from about 6% to 4% of the 50 CZs (i.e. 2 of the 50 CZs). Therefore, even after a 10-fold increase in the sample size, there still is considerable uncertainty in the rank in the movers design whereas the rank in the correlational design is almost certain.

			correlational design						movers design					
κ	p	method	$\delta = 0$	$\delta = 0.25$	$\delta = 0.5$	$\delta = 0.75$	$\delta = 1$	$\delta = 2$	$\delta = 0$	$\delta = 0.25$	$\delta = 0.5$	$\delta = 0.75$	$\delta = 1$	$\delta = 2$
0.5	2	M	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.999	0.999	1.000	1.000	1.000
		DM	0.949	1.000	1.000	1.000	1.000	1.000	0.947	0.979	0.980	0.984	0.986	0.991
		DM.step	0.949	1.000	1.000	1.000	1.000	1.000	0.947	0.979	0.980	0.984	0.986	0.991
	25	M	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000
		DM	0.946	1.000	1.000	1.000	1.000	1.000	0.948	0.989	0.997	0.999	0.999	1.000
		DM.step	0.946	1.000	1.000	1.000	1.000	1.000	0.948	0.989	0.997	0.999	0.999	1.000
	50	M	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		DM	0.939	0.999	1.000	1.000	1.000	1.000	0.949	0.990	0.997	1.000	1.000	1.000
		DM.step	0.939	0.998	0.998	0.998	0.998	0.998	0.949	0.990	0.997	1.000	1.000	1.000
1	2	M	0.998	1.000	1.000	1.000	1.000	1.000	0.998	1.000	1.000	1.000	1.000	1.000
		DM	0.953	1.000	1.000	1.000	1.000	1.000	0.955	0.981	0.984	0.986	0.988	0.995
		DM.step	0.953	1.000	1.000	1.000	1.000	1.000	0.955	0.981	0.984	0.986	0.988	0.995
	25	M	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		DM	0.942	1.000	1.000	1.000	1.000	1.000	0.949	0.992	0.998	1.000	1.000	1.000
		DM.step	0.942	1.000	1.000	1.000	1.000	1.000	0.949	0.992	0.998	1.000	1.000	1.000
	50	M	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		DM	0.939	0.999	0.999	0.999	0.999	0.999	0.951	0.995	0.998	0.999	1.000	1.000
		DM.step	0.939	0.998	0.998	0.998	0.998	0.998	0.951	0.995	0.998	0.999	1.000	1.000
10	2	M	0.998	1.000	1.000	1.000	1.000	1.000	0.998	1.000	1.000	1.000	1.000	1.000
		DM	0.953	1.000	1.000	1.000	1.000	1.000	0.951	0.987	0.994	0.998	0.999	1.000
		DM.step	0.953	1.000	1.000	1.000	1.000	1.000	0.951	0.987	0.994	0.998	0.999	1.000
	25	M	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		DM	0.949	1.000	1.000	1.000	1.000	1.000	0.946	1.000	1.000	1.000	1.000	1.000
		DM.step	0.949	1.000	1.000	1.000	1.000	1.000	0.946	1.000	1.000	1.000	1.000	1.000
	50	M	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		DM	0.946	1.000	1.000	1.000	1.000	1.000	0.953	1.000	1.000	1.000	1.000	1.000
		DM.step	0.946	1.000	1.000	1.000	1.000	1.000	0.953	1.000	1.000	1.000	1.000	1.000

Table 2: Marginal confidence set for the rank of a single CZ: coverage

5 Empirical Applications

5.1 Ranking of Developed Countries by Student Performance in PISA

We now apply our inference procedures from Section 3 to re-examine the question that motivates the PISA test: Which countries do best and worst at reading, math, and science?

What is PISA and why does it matter?

Over the past two decades, the Organisation for Economic Co-operation and Development (OECD) have conducted the PISA test. The goal of this test is to evaluate and compare educational systems across countries by measuring 15-year-old school students' scholastic performance on math, science, and reading. The PISA test was first performed in 2000 and then repeated every three years. Each country that participates in a given year has to draw a sample of at least 5,000 students to be tested. The results from the PISA test are reported on a scale constructed using a generalized form of the Rasch model (OECD, 2017). For each domain (reading, math, and science), the scale is constructed with a mean score of 500 and standard deviation of 100. The scores are then tabulated by country in what has become known as PISA's international league tables.

Every three years, the release of these league tables stimulates a global discussion about education systems and school reform in both international media and at the national level across many OECD countries. Indeed, several governments have set national performance targets based on how well the country ranks in the league tables (Breakspear, 2012). A low ranking in the PISA league table is known to cause media attention and political discussion. In Germany, for example, the poor results in the first PISA test triggered a heated debate about the country's education system, which ultimately resulted in wideranging reforms (Hubert, 2006).

How much should we trust the ranking in PISA's league tables?

In order to examine which countries do best and worst at reading, math, and science, we use publicly available data from the 2018 PISA test. We restrict attention to the OECD countries. Since PISA never combines math, science, and reading scores into an overall score, we perform our analyses separately for each domain. For brevity, we focus on the league table for reading, but we report a complete set of results for each domain in Appendix G.1.

We begin by presenting the point estimates and marginal confidence intervals (estimates plus or minus twice the standard errors) for the expected reading test score in each OECD country.[4] These results are reported in Figure 8. There is considerable variation in the point estimates across countries. Estonia ranks first with an average test score of around 523. The runner up is Canada, followed by Finland in third place. At the bottom of the league table, one finds Chile, Mexico, and Columbia. These countries have reading scores that are more than 20% lower than the countries at the top of the league table.

By applying our procedures from Section 3 to the point estimates and standard errors in Figure 8, we can compute (i) the marginal confidence set for the rank of a given country, (ii) the simultaneous confidence

[4]The only exception is Spain, for which there is no data available.

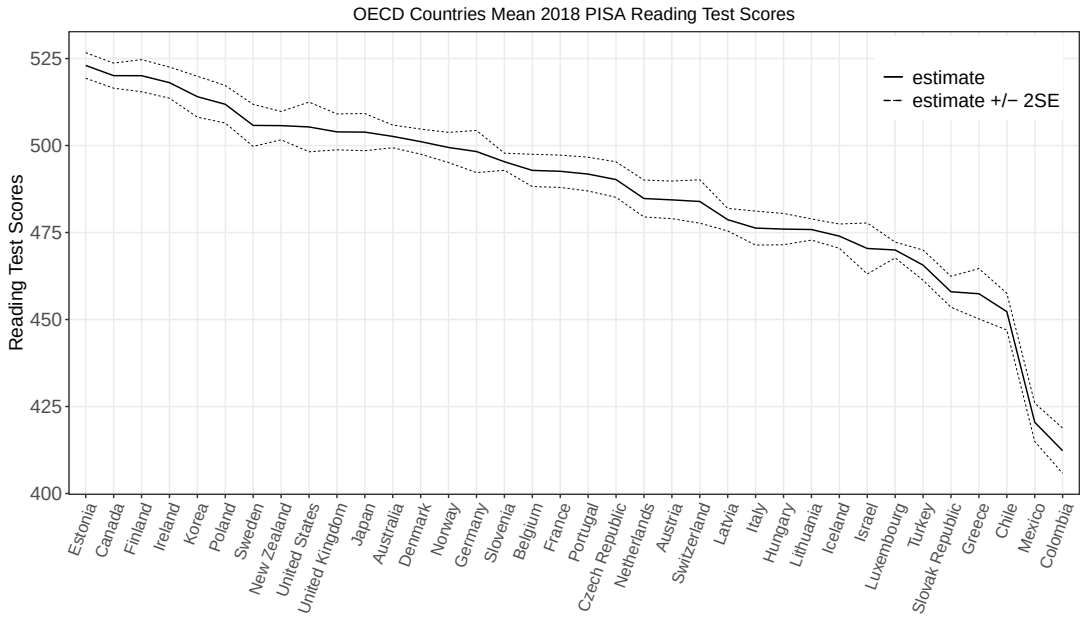

Figure 8: Point estimates and marginal confidence intervals (estimates plus or minus twice the standard errors) of the expected reading score on the PISA test for each OECD country (except for Spain for which there is no data)

set for the ranks of all countries, and (iii) confidence sets for the τ-best (or the τ-worst) countries. Marginal confidence sets answer the question of whether a given country is performing relatively well on the reading test as compared to the other countries. Thus, (i) is relevant if one is interested in whether a particular country is among the worst or the best countries in terms of its scholastic performance on reading. Simultaneous confidence sets allow such inferences to be drawn simultaneously for all countries. Thus, (ii) is relevant if one is interested in broader geographic patterns of scholastic performance in reading across OECD countries. By comparison, confidence sets for the τ-best (or τ-worst) answer the more specific question of which OECD countries cannot be ruled out as being among the countries with the best (worst) scholastic performance in reading. In other words, (iii) is relevant if one is interested in only the top (or bottom) of the international league table.

The confidence sets are implemented as for the simulations in Section 4 and Appendix F, using the stepwise procedures ("DM.step") for the confidence sets for ranks and the projection method ("DM.step") for the τ-best and τ-worst problems. All confidence sets are computed at the 95% nominal level.

Figure 9 presents the ranking of the OECD countries according to the point estimates of the expected reading scores. Panel A displays the marginal confidence sets while Panel B reports the simultaneous confidence sets. Table 3 reports additional results for the top five countries (Panel A) and the bottom five countries (Panel B). Each panel of this table presents results for math, reading, and science. For each domain, we report the point estimates, the standard errors, the 95% marginal confidence sets for the ranks, and the number of countries that cannot be ruled out (with 95% confidence) as being among the set of countries with the τ-highest (top panel) or the τ-lowest (bottom panel) expected PISA test scores.

As evident from Panel A of Figure 9, the marginal confidence sets are relatively narrow, especially for the countries at the top and the bottom of the ranking. This finding suggests that citizens of these countries

can be quite confident in the reading performance of their pupils. For instance, the upper endpoint of the confidence set for Estonia's rank suggests it is (with 95% confidence) the country with at least the fifth-highest expected test score. By comparison, the lower endpoint of the confidence set for Columbia's rank suggests (with 95% confidence) that it is among the bottom two OECD countries in terms of scholastic performance on reading.

A natural question is whether the ranking of the OECD countries according to the expected reading score remains informative if one allows inferences to be drawn simultaneous across all countries. The results in Panel B of Figure 9 suggest the ranking remains fairly informative, especially at the top and at the bottom of the PISA league table. Therefore, we can be fairly certain about which countries are at the top and bottom of the ranking. In addition, the columns denoted by "τ-best" and "τ-worst" in Table 3 show the number of countries in the 95% confidence set for the τ-best and τ-worst. Only eight (five) countries cannot be ruled out as being among the top (bottom) three countries in terms of scholastic performance on reading.

Remark 5.1. As discussed in the introduction, our methods and those developed by Andrews et al. (2018) share some technical similarities, but answer distinct economic questions, and should thus be viewed as complements, not substitutes. To illustrate this empirically, we apply the methods of Andrews et al. (2018) to the PISA data and construct 95% confidence sets for the expected PISA scores of the countries with the highest and the lowest estimated scores.[5] For instance, their 95% "conditional" confidence set for the expected reading score of the sample "winner", Estonia, is (517.9, 526.6). This is a fairly narrow confidence set for the expected value of the reading score and thus one can be confident that the sample "winner" truly has a high reading score.

However, just like Estonia's marginal confidence interval displayed in Figure 8, Andrews et al. (2018)'s confidence set does not allow us to draw any conclusions about what is the true rank of Estonia nor which country has true rank one. On the other hand, our marginal confidence sets for the rank of Estonia tell us that (with 95% probability) its true rank lies between 1 and 5. In addition, our τ-best confidence set for $\tau = 1$ shows that (with 95% probability) there are 6 countries in total that could be the best. See Appendix G.2 for more results and details. ■

[5] We are grateful to the authors for sharing their code, allowing us to easily apply their methods to our applications.

Panel A: Top 5

		Math					Reading					Science				
Rank	τ	Country	Score	SE	95% CS	τ-best	Country	Score	SE	95% CS	τ-best	Country	Score	SE	95% CS	τ-best
1	1	Japan	526.97	2.47	[1, 6]	6	Estonia	523.02	1.84	[1, 5]	6	Estonia	530.11	1.88	[1, 4]	4
2	2	Korea	525.93	3.12	[1, 6]	7	Canada	520.09	1.80	[1, 6]	7	Japan	529.14	2.59	[1, 4]	5
3	3	Estonia	523.41	1.74	[1, 6]	7	Finland	520.08	2.31	[1, 6]	8	Finland	521.88	2.51	[1, 6]	6
4	4	Netherlands	519.23	2.63	[1, 8]	11	Ireland	518.08	2.24	[1, 7]	8	Korea	519.01	2.80	[2, 7]	7
5	5	Poland	515.65	2.60	[1, 11]	13	Korea	514.05	2.94	[1, 11]	14	Canada	518.00	2.15	[3, 7]	11

Panel B: Bottom 5

		Math					Reading					Science				
Rank	τ	Country	Score	SE	95% CS	τ-worst	Country	Score	SE	95% CS	τ-worst	Country	Score	SE	95% CS	τ-worst
33	5	Turkey	453.51	2.26	[32, 34]	6	Slovakia	457.98	2.23	[30, 34]	8	Israel	462.20	3.62	[30, 34]	9
34	4	Greece	451.37	3.09	[32, 34]	6	Greece	457.41	3.62	[30, 34]	7	Greece	451.63	3.14	[33, 35]	6
35	3	Chile	417.41	2.42	[35, 36]	3	Chile	452.27	2.64	[32, 34]	5	Chile	443.58	2.42	[34, 35]	4
36	2	Mexico	408.80	2.49	[35, 36]	3	Mexico	420.47	2.75	[35, 36]	2	Mexico	419.20	2.58	[36, 37]	2
37	1	Colombia	390.93	2.99	[37, 37]	1	Colombia	412.30	3.25	[35, 36]	2	Colombia	413.32	3.05	[36, 37]	2

Table 3: **Panel A:** Top 5 among the OECD countries ranked by PISA test scores in Math, Reading, and Science with the marginal 95% confidence sets ("CS") for their ranks and the size of the 95% confidence set for the τ-best. **Panel B:** Bottom 5 among the OECD countries ranked by PISA test scores in Math, Reading, and Science with marginal 95% confidence sets ("CS") for their ranks and the size of the 95% confidence set for the τ-worst. *Note:* Spain is absent in the Reading test score results so the lowest possible rank for Reading is 36.

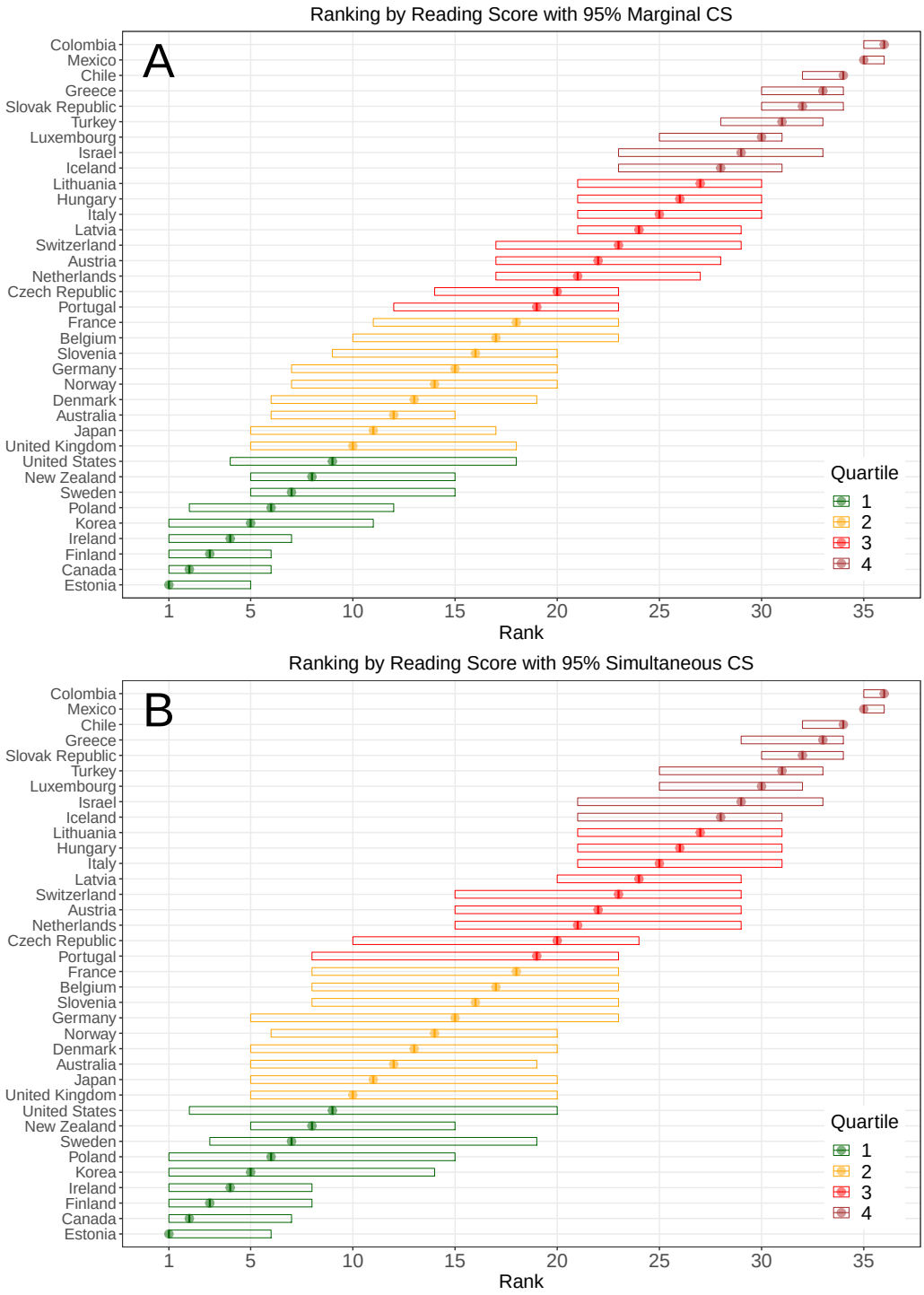

Figure 9: **Panel A:** for each OECD country, we plot its rank by reading score and the 95% marginal confidence set ("CS"). **Panel B:** for each OECD country, we plot its rank by reading score and the 95% simultaneous confidence set ("CS"). Different quartiles of the rankings are indicated with different colors.

5.2 Ranking Neighborhoods by Intergenerational Mobility

We now apply our inference procedures from Section 3 to re-examine the question that motivates the work of Chetty et al. (2014, 2018) and Chetty and Hendren (2018): Where in the United States is the land of opportunity?

Data and background

The empirical analysis in this section is based on publicly available estimates of intergenerational income mobility across areas in the United States. These estimates come from two studies that both use tax records covering the U.S. population. The first is Chetty et al. (2018).[6] They document how children's expected incomes conditional on their parents' incomes vary according to the area (commuting zone (CZ), county, or Census tract) in which they grew up. The second is Chetty and Hendren (2018).[7] The goal of this paper is to examine the degree to which the differences in income mobility across areas reflect causal effects of place. Both studies present the empirical results through league tables and heat maps which rank places according to point estimates of income mobility.

In the baseline analysis, Chetty et al. (2018) define the following measure of intergenerational mobility:

$$\bar{y}_{cp} \equiv E[y_i | c(i) = c, p(i) = p] \ , \tag{32}$$

where y_i is child i's percentile rank in the national distribution of incomes relative to all others in her birth cohort; child i's income is measured as her average income in the years 2014–2015 (aged 31–37 depending on cohort); $p(i)$ denotes the child's parental income percentile in the national distribution of parental income in child i's birth cohort; and $c(i)$ is the area in which the child i grew up. As in Chetty et al. (2018), we focus on $\bar{y}_{c25}$, the expected income rank of children who grew up in area c with parents at the 25th percentile of the national income distribution of parental income.[8] Following Chetty et al. (2018), we refer to the estimates of $\bar{y}_{c25}$ as *correlational estimates* of upward mobility.

In Figure 10, we present the estimates of $\bar{y}_{c25}$ with marginal confidence intervals (estimates plus or minus twice the standard errors) from Chetty et al. (2018). These correlational estimates of upward mobility cover all the 741 commuting zones and 3208 of the 3219 counties.[9] We first sort the places by the values of $\hat{\bar{y}}_{c25}$, and then report these point estimates and their marginal confidence intervals for each CZ (top graph) and county (bottom graph). There is considerable variation in $\hat{\bar{y}}_{c25}$ across areas. Since CZs typically comprise several counties, it is not surprising that the standard errors tend to be a lot larger when a neighborhood is defined as a county rather than as a CZ.

In Chetty and Hendren (2018), the parameters of interest are the exposure effects of spending an additional year of one's childhood in a given area. Consider a child i from a set of one-time movers from an

[6] The data files could be accessed following these links: commuting zones; counties and tracts. The variables of interest in all three files are *kfr_pooled_pooled_p25* and *kfr_pooled_pooled_p25_se*.

[7] The data files could be accessed following these links: commuting zones and counties. The variables of interest are *causal_p25_czkr26* and *causal_p25_czkr26_se* for commuting zones; *causal_p25_cty_kr26* and *causal_p25_cty_kr26_se* for counties.

[8] Chetty et al. (2018) take several steps to simplify the estimation problem of the $\bar{y}_{c25}$ across areas. We use the main estimates they report and refer to their paper for estimation details.

[9] Following Chetty et al. (2018) we use 1990 Commuting Zones classification and 2000 counties classification. For 11 counties data is not available.

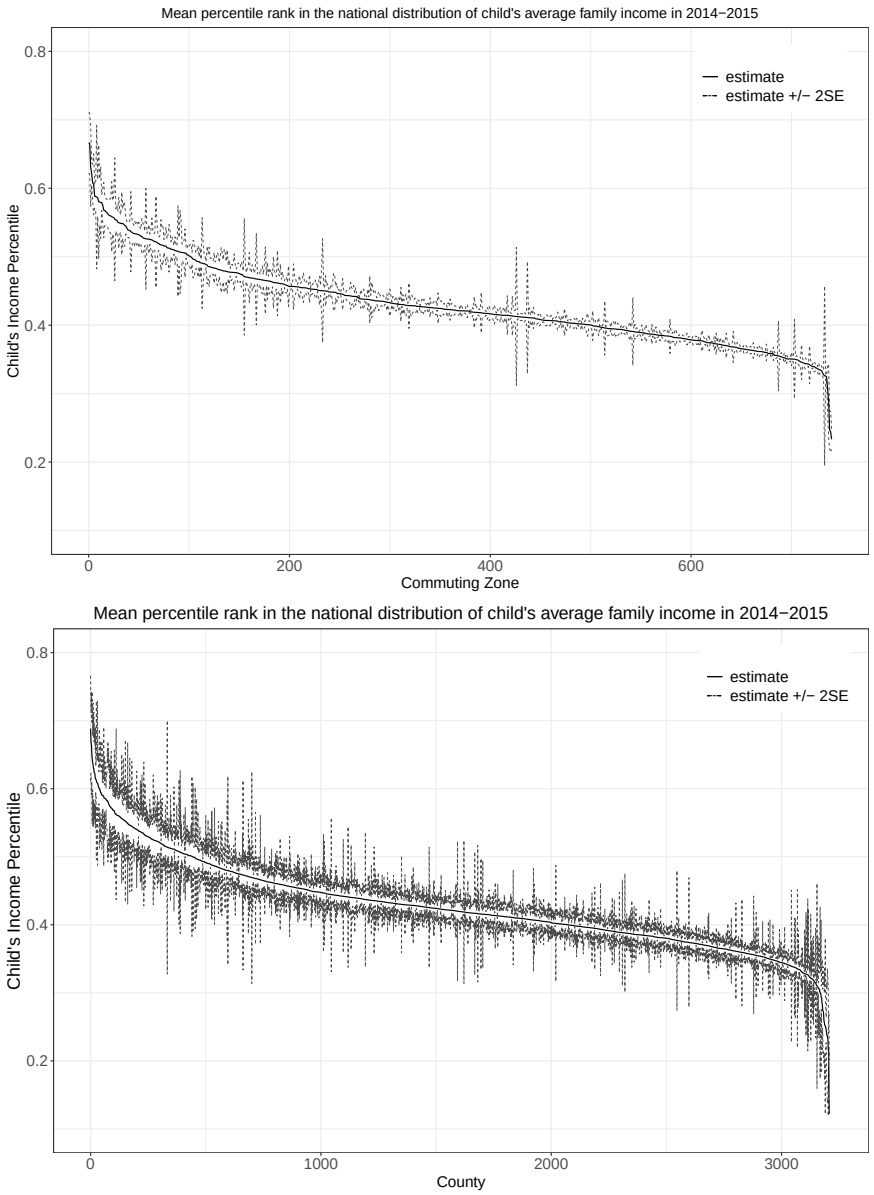

Figure 10: Estimates of $\bar{y}_{c25}$, the expected percentile rank of a child's average household income for 2014-2015 in the national distribution of her cohort, with marginal confidence intervals (estimates plus or minus twice the standard errors). The estimates cover all 741 commuting zones (Top) and 3208 of the 3219 counties (Bottom).

origin $o(i)$ to a destination $d(i)$. She moves at the age $m(i)$ and spends $A - m(i)$ time in the destination. The (vector of the) amount of time spent in a given area is denoted by:

$$e_{ic} \equiv \begin{cases} A - m(i) & \text{if } c = d(i) \\ m_i & \text{if } c = o(i) \\ 0 & \text{otherwise} \end{cases} \tag{33}$$

The exposure effects can be estimated by the regression model:

$$y_i = \alpha_{od} + \vec{\boldsymbol{e}}_i \cdot \vec{\boldsymbol{\mu}} + \varepsilon_i, \tag{34}$$

where α_{od} is an origin-by-destination fixed effect, $\vec{\boldsymbol{e}}_i \equiv (e_{ic} \colon c = 1, 2, \ldots)$ is a vector of explanatory variables for the number of years that child i lived in place c during her childhood, and the exposure effects are given by the parameters $\vec{\boldsymbol{\mu}} \equiv (\mu_{cp} \colon c = 1, 2, \ldots) \equiv (\mu_c^0 + \mu_c^1 p \colon c = 1, 2, \ldots)$, where p is the parental income percentile. The estimates are normalized to be mean zero across places, so that μ_{cp} measures the exposure effect relative to the average place. As in Chetty and Hendren (2018) we focus on μ_{c25}, the effect of spending an additional year of childhood in area c for children with parents at the 25th percentile of the national income distribution of parental income.[10] Following Chetty and Hendren (2018), we refer to the estimates of μ_{c25} as *movers estimates* of exposure effects.

In Figure 11, we present the point estimates of μ_{c25} with marginal confidence intervals (estimates plus or minus twice the standard errors) from Chetty and Hendren (2018). These results cover 595 of the 741 CZs and 2367 of the 3219 counties.[11] The point estimates suggest considerable variation in exposure effects across areas. The standard errors are, however, sizable, indicating that it can be difficult to draw firm conclusions about which areas produce more or less upward mobility.

Given the relatively large standard errors, in a subset of the analyses, we restrict attention to the most populous CZs and counties. The motivation for this sample restriction is to examine if one can achieve a more informative ranking by restricting attention to larger areas. This sample restriction is also imposed in a subset of the analyses of Chetty et al. (2014, 2018) and Chetty and Hendren (2018). In Appendix G.3, we present point estimates of $\bar{y}_{c25}$ and μ_{c25} with marginal confidence intervals for the 50 most populous CZs and counties. As expected, the estimates are more precise for this restricted set of areas as compared to the population of CZs and counties at large. The gains in precision are particularly salient for the correlational estimates. By way of comparison, the standard errors of the movers estimates remain relatively large even if one restricts attention to the most populous areas.

Before we present the confidence sets for the ranks, there are three remarks worth making. First, Chetty and Hendren (2018) report both the raw estimates of the exposure effect of place c, μ_{c25}, as well as forecasts that minimize the mean-squared-error (MSE) of the predicted impact of growing up in place c. We focus on the raw estimates. This choice is, in part, because Chetty and Hendren (2018) do not report the confidence intervals on the forecasts, but also because the forecasts are very similar to the correlational estimates in most CZs. The reason is that the forecasts are constructed as weighted averages of the correlational estimates (based on stayers) and the mover estimates, with greater weight on the mover estimates when they are more precisely estimated. Given that most estimates of μ_{c25} are very noisy, the forecast estimates are very similar to the correlational estimates. Indeed, we calculate that in a majority of the CZs, the forecasts assign at least 90 percent of the weight to the correlational estimates.

Second, the movers estimators may not necessarily be independent across CZs. While our inference procedures accommodate dependence in a straightforward fashion (see Remarks 3.5 and 3.16), doing so would require not only standard errors for each mobility estimate, but an estimate of the whole covariance matrix of the estimators. Such information is unfortunately not available to us. Thus, we are unable to

[10] Chetty and Hendren (2018) take several steps to simplify the estimation problem of μ_{c25} across areas. We use the main estimates they report and refer to their paper for estimation details.

[11] Chetty and Hendren (2018) do not report results for the other counties and CZs due to limited data in these areas.

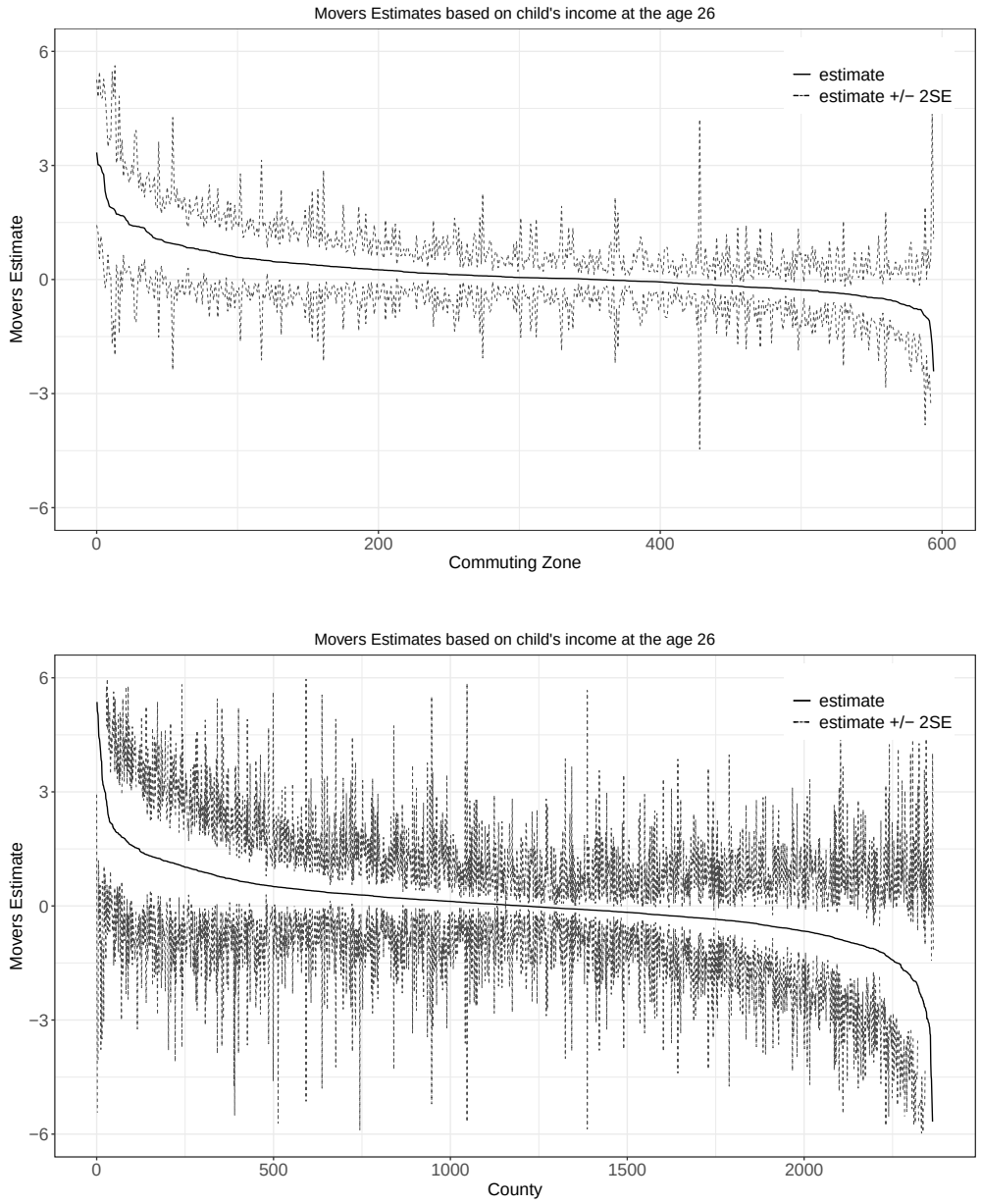

Figure 11: Movers estimates of exposure effects μ_{c25} with marginal confidence intervals (estimates plus or minus twice the standard errors). The estimates cover 595 of the 741 commuting zones (Top) and 2367 of the 3219 counties (Bottom).

examine if the movers estimators are dependent or incorporate such dependence in the construction of the confidence sets for the ranks. Furthermore, ignoring potential dependence among the estimators most likely understates the uncertainty in the estimates we use, so we conjecture our very wide confidence sets for the ranks would widen even further when accounting for dependence.

Third, we follow closely the analyses of Chetty et al. (2014, 2018) and Chetty and Hendren (2018), and do not try to change their definitions of neighborhoods to improve precision. If the commuting zones were

aggregated up to larger geographic regions such as states or Census regions, it is likely that one obtains a more informative ranking. However, this would change the target parameters and the analyses might no longer be suitable for the stated purposes, namely to draw the attention of policymakers to neighborhoods that need improvements and to help families move to high-mobility neighborhoods. For these purposes, local statistics are arguably needed. Indeed, in recent work, Chetty et al. (2018) define a neighborhood to be a Census-tract, which encompasses a population of between 2,500 and 7,500 people, and is even more granular than counties. They then construct heat maps (referred to as the Opportunity Atlas) by dividing the Census-tracts into deciles based on their estimated value of $\bar{y}_{c25}$. A stated goal of these maps are to identify local high-opportunity neighborhoods that are affordable to low-income families and providing an input into the design of affordable housing policies. The estimates and standard errors of $\bar{y}_{c25}$ for each Census-tract level is available here. When using this data, we find that both the marginal and simultaneous confidence sets are far too wide to draw conclusions about the ranks of the neighborhoods at such a granular level. For brevity, we chose not to report these results in the paper, focusing instead on CZs and counties for which informative rankings are more likely.

Confidence sets used to rank places by intergenerational mobility

By applying our procedures from Section 3 to the point estimates and standard errors in Figures 10 and 11, we can compute (i) the marginal confidence sets for the rank of a given place, (ii) the simultaneous confidence sets for the ranks of all places, and (iii) the confidence sets for the τ-best (or the τ-worst) ranked places.

Before presenting the results, we again emphasize that (i)–(iii) answer distinct economic questions. Marginal confidence sets answer the question of whether a given place has relatively high or low income mobility compared to other places. Thus, (i) is relevant if one is interested in whether a particular place is among the worst or the best places to grow up in terms of income mobility. Simultaneous confidence sets allow such inferences to be drawn simultaneously across all places. Thus, (ii) is relevant if one is interested in broader geographic patterns of income mobility across the United States. By comparison, confidence sets for the τ-best (or τ-worst) answer the more specific question of which places cannot be ruled out as being among the areas with the most (least) income mobility. In other words, (iii) is relevant if one is interested in only the top (or bottom) of a league table of neighborhoods by income mobility.

The confidence sets are implemented as for the simulations in Section 4 and Appendix F, using the stepwise procedures ("DM.step") for the confidence sets for ranks and the projection method ("DM.step") for the τ-best and τ-worst problems. All confidence sets are computed at the 95% nominal level.

Ranking of the most populous places

We begin the empirical analysis by considering the 50 largest CZs by population size. Figure 12 presents the ranking of these CZs according to the point estimates of $\bar{y}_{c25}$. Panel A displays the marginal confidence sets while Panel B reports the the simultaneous confidence sets. Table 4 reports additional results for the top five CZs (Panel A) and the bottom five CZs (Panel B). Each panel of this table presents two sets of results: Columns 3–7 are based on the correlational estimates of upward mobility $\bar{y}_{c25}$, while columns 8–12 are based on the movers estimates of exposure effects μ_{c25}. For each set of results, we report the point estimates, the

standard errors, the 95% marginal confidence sets, and the number of places in the 95% confidence sets for the τ-best (top panel) or the τ-worst values of $\bar{y}_{c25}$ or μ_{c25}.

Among the 50 largest CZs by population size, the point estimates of $\bar{y}_{c25}$ range from 0.457 in San Francisco to 0.355 in Charlotte. As evident from Panel A of Figure 12, the marginal confidence sets based on the correlational estimates are relatively narrow, especially for the CZs at the top and the bottom of the ranking. This finding suggests that citizens of these CZs can be quite confident in the mobility ranking of their hometown. For instance, with 95% confidence, San Francisco is among the top two of these 50 CZs in terms of income mobility. By comparison, with 95% confidence, Charlotte is among the bottom three of these 50 CZs in terms of income mobility.

A natural question is whether the ranking of the CZs according to the correlational estimates remains informative if one allows inferences to be drawn simultaneously across all places. The results in Panel B of Figure 12 suggest this is indeed the case and we can have high confidence about which CZs are at the top and bottom of the correlational ranking. The sizes of the 95% confidence sets for the τ-best and τ-worst CZs confirm this finding. For example, only four (three) places cannot be ruled out as being among the top (bottom) two CZs in terms of income mobility. Furthermore, there are only six places that cannot be ruled out as being among the top five CZs, while ten CZs cannot be ruled out as being among the bottom five places.

Taken together, the results based on the correlational estimates $\bar{y}_{c25}$ suggest it is possible to achieve a quite informative ranking of the 50 largest CZs according to upward mobility. By contrast, the exposure effects μ_{c25} are too imprecisely estimated to draw firm conclusions about which CZs produce more or less upward mobility. As evident from the marginal confidence sets for μ_{c25} in column 11 of Table 4, it is difficult to learn much about whether a particular CZ has relatively high or low exposure effects. For example, the citizens of Seattle cannot rule out with 95% confidence that the majority of other CZs have higher income mobility. Drawing inferences simultaneously across all CZs is even more challenging, as evident by the τ-best and τ-worst results for μ_{c25}. Consider, for example, column 12 of Panel A in Table 4. As these results show, none of the 50 CZs can be ruled out with 95% confidence as being among the top five places in terms of exposure effects.

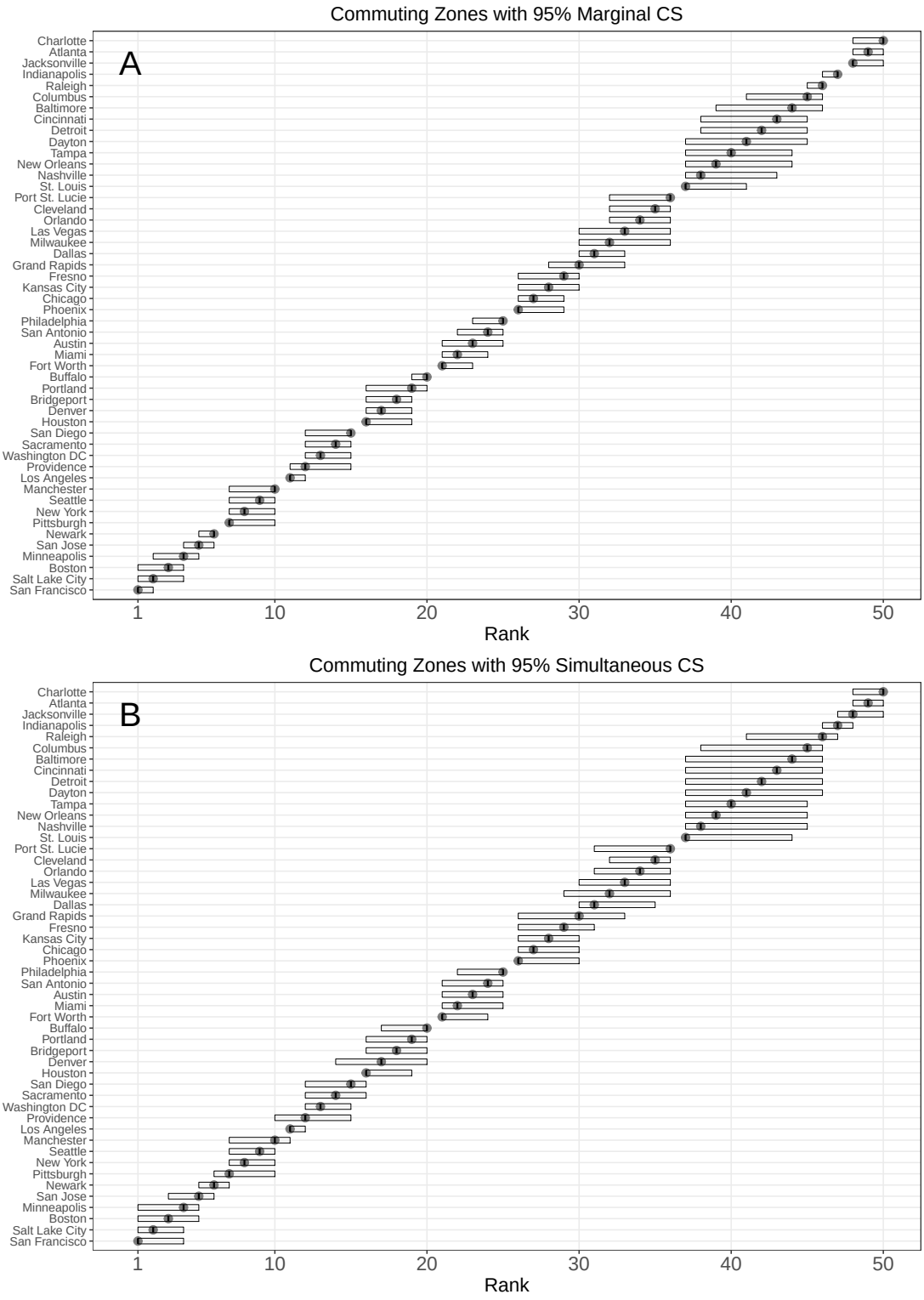

Figure 12: **Panel A:** point estimates and the 95% marginal confidence sets ("CS") for the ranking of the 50 most populous CZs by $\bar{y}_{c25}$. **Panel B:** point estimates and the 95% simultaneous confidence sets ("CS") for the ranking of the 50 most populous CZs by $\bar{y}_{c25}$.

Panel A: Top 5

		Correlational					Movers				
Rank	τ	CZ	$\hat{\bar{y}}_{c25}$	SE	95% CS	τ-best	CZ	$\hat{\mu}_{c25}$	SE	95% CS	τ-best
1	1	San Francisco	0.457	0.001	[1, 2]	4	Seattle	0.229	0.082	[1, 38]	44
2	2	Salt Lake City	0.457	0.001	[1, 4]	4	Washington DC	0.163	0.077	[1, 41]	48
3	3	Boston	0.453	0.001	[1, 4]	5	Cleveland	0.124	0.107	[1, 48]	50
4	4	Minneapolis	0.452	0.001	[2, 5]	5	Fort Worth	0.121	0.090	[1, 48]	50
5	5	San Jose	0.449	0.001	[4, 6]	6	Minneapolis	0.116	0.120	[1, 48]	50

Panel B: Bottom 5

		Correlational					Movers				
Rank	τ	CZ	$\hat{\bar{y}}_{c25}$	SE	95% CS	τ-worst	CZ	$\hat{\mu}_{c25}$	SE	95% CS	τ-worst
46	5	Raleigh	0.369	0.001	[45, 46]	10	Charlotte	-0.248	0.096	[3, 50]	49
47	4	Indianapolis	0.364	0.001	[46, 47]	5	Port St. Lucie	-0.263	0.090	[3, 50]	49
48	3	Jacksonville	0.358	0.001	[48, 50]	4	Raleigh	-0.278	0.105	[3, 50]	49
49	2	Atlanta	0.358	0.001	[48, 50]	3	Fresno	-0.377	0.100	[13, 50]	48
50	1	Charlotte	0.355	0.001	[48, 50]	3	New Orleans	-0.391	0.111	[14, 50]	48

Table 4: **Panel A:** Top 5 among the 50 most populous commuting zones ranked by the correlational estimates on the left and by the movers estimates on the right. **Panel B:** Bottom 5 among the 50 most populous commuting zones ranked by the correlational estimates on the left and by the movers estimates on the right. "95% CS" refers to the 95% marginal confidence set for the rank, and "τ-best" and "τ-worst" refer to the size of the 95% confidence sets for the "τ-best" and "τ-worst" commuting zones.

As shown in Appendix G.3, the above conclusions do not materially change if we instead consider the 50 largest counties by population size. On the one hand, it is possible to achieve a quite informative ranking of these counties according to $\bar{y}_{c25}$. Both the marginal and the simultaneous confidence sets are fairly narrow, and relatively few counties are included in the confidence sets for the τ-best or the τ-worst places. On the other hand, the exposure effects μ_{c25} are too imprecisely estimated to obtain an informative ranking of counties according to income mobility. First of all, the marginal confidence sets for μ_{c25} are generally too wide to draw conclusions about whether a particular county has among the highest or the lowest exposure effect, as evident from column 11 of Table 12.[12] Furthermore, the τ-best and τ-worst results for μ_{c25} show that the ranking of counties by exposure effects is largely uninformative when inferences are drawn simultaneously across all places. Consider, for example, column 12 of Table 12. These results show that none of these counties can be ruled out with 95% confidence as being among the top two places when it comes to exposure effects, and only one county can be ruled out with 95% confidence as being at the very bottom of this ranking.

So far, we have presented the statistical uncertainty through 95% confidence sets for the rank of a given place. An alternative way to present the statistical uncertainty is to compute the number of places whose confidence sets for the ranks have upper endpoint equal to the lower endpoint for a given confidence level. If the two endpoints are equal, then we know with 95% confidence the true rank of the place. In Appendix Table 13, we perform such computations to summarize the statistical uncertainty. We consider the simultaneous confidence sets for the ranks for both the 50 most populous CZs and the 50 most populous counties, and we use estimates of both the correlational measures $\bar{y}_{c25}$ and of the exposure effects μ_{c25}.

The results echo our previous conclusions about the uncertainty in the ranking of places by upward mobility. The results based on the correlational estimates suggest it is possible to achieve somewhat informative

[12] An exception is DuPage for which the marginal confidence set suggests that its exposure effect is relatively high compared most of the other counties.

conclusions about the ranking of the 50 largest CZs or counties. For example, at a 90% percent confidence level, there are 1 CZ and 7 counties for which the endpoints of the confidence sets for the rank are equal. By comparison, at a 10% percent confidence level, there are 4 CZs and 8 counties with that property. In contrast, the movers estimates for CZs and counties are too imprecise to obtain confidence sets for the ranks with equal endpoints, even at confidence levels as low as 5%.

National ranking of places by income mobility

So far, we have focused on the 50 largest CZs and counties by population size. We now shift attention to all CZs and counties, revisiting the key question of Chetty et al. (2014), Chetty et al. (2018) and Chetty and Hendren (2018): Where in the United States is the land of opportunity?

In order to analyze this question, the authors present heat maps based on estimates of upward mobility. They construct these maps by dividing the CZs (or the counties) into deciles based on their estimated value of $\bar{y}_{c25}$. Panel A of Figure 13 presents the heat map for the CZs. This map is the same as presented in Chetty et al. (2014). Lighter colors represent deciles with higher values of $\bar{y}_{c25}$. Equivalently, one can interpret the heatmap as showing the ranks of CZs by assigning the same color to ranks in a decile to easy readability (rather than a unique color to each rank). The point estimates of income mobility vary significantly across areas. For example, CZs in the top decile have an $\hat{\bar{y}}_{c25} > 0.517$, while those in the bottom decile have $\hat{\bar{y}}_{c25} < 0.362$. Note that the 36th percentile of the family income distribution for children at age 31–37 is \$26,800, while the 52nd percentile is \$44,800; hence, the differences in upward mobility across these areas correspond to substantial differences in children's incomes.

The stated purpose of heat maps such as the one in Panel A of Figure 13 is to draw the attention of policymakers to low-mobility neighborhoods that need improvement and to help low-income families move to high-mobility neighborhoods. A natural question is how informative the local statistics reported in these maps are about a given neighborhood having relatively high or low income mobility compared to other neighborhoods. To answer this question, we construct two new heat maps. These maps show the upper (Panel B of Figure 13) and lower (Panel C of Figure 13) endpoints of the simultaneous confidence sets for the CZs' ranks. These confidence sets allow inferences to be drawn simultaneously across all CZs. Thus, the new results in Figure 13 make precise what conclusions one can draw about which individual CZs have relatively high and low mobility.

In order to interpret the results, it is useful to observe that if the simultaneous confidence sets were sufficiently narrow, then the heat map in Panel B would be identical to the heat map in Panel C. It is only in this case the point estimates of $\bar{y}_{c25}$ and, thus, the heat map in Panel A (or, equivalently, in Chetty et al. (2014, p. 1591)), would give a reliable answer to the question of where in the United States is the land of opportunity. More generally, how much we can learn about this question depends on how similar the heat map in Panel B is to the heat map in Panel C. If the CZs that have lighter colors in Panel B also have lighter colors in Panel C, then we can with 95% confidence conclude that these areas have high mobility. Conversely, if the CZs that have darker colors in Panel C also have darker colors in Panel B, then we can with 95% confidence conclude that these areas have low mobility.

A visual inspection of the heat maps in Panels B and C of Figure 13 indicates that the uncertainty tends to be too large to draw firm conclusions about which CZs have high or low income mobility compared to other places in the United States. In other words, it is not possible to tell apart with 95% confidence the

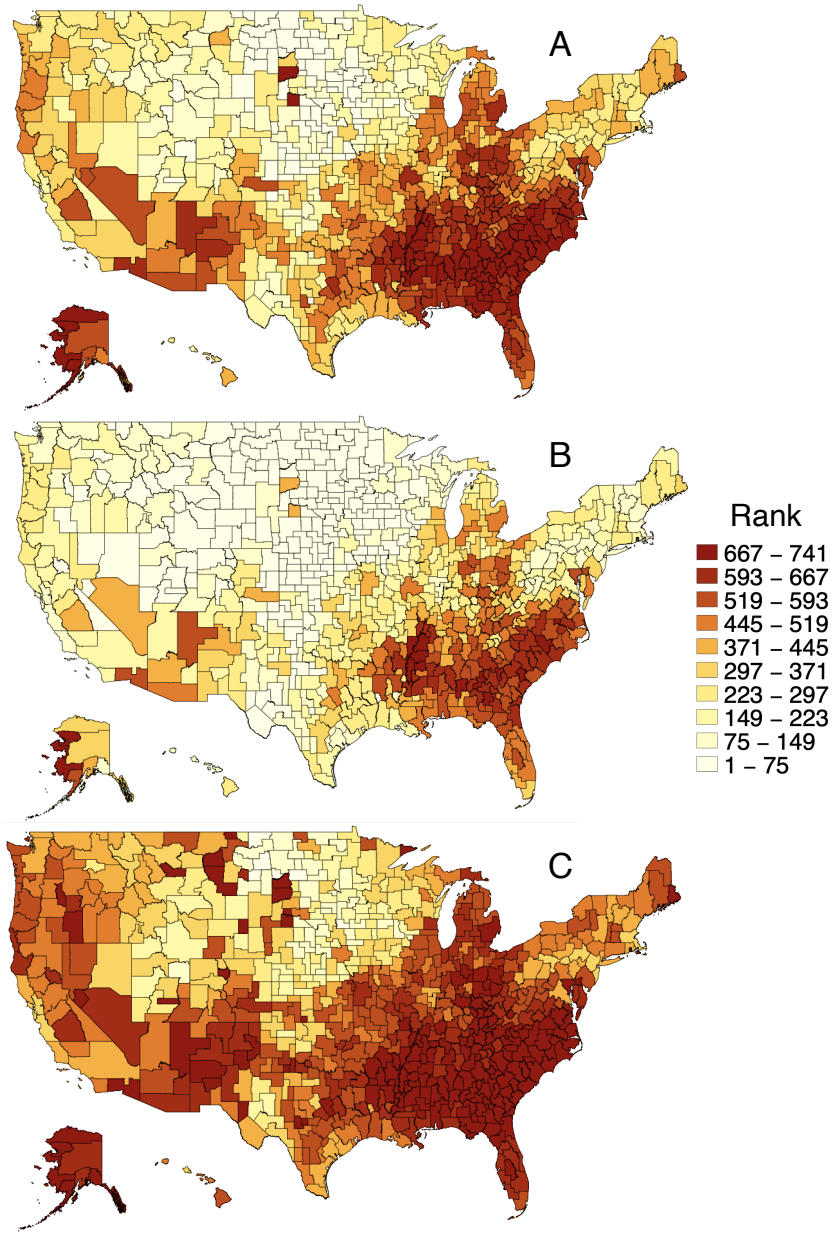

Figure 13: Ranking of Commuting Zones by point estimates and lower and upper endpoints of simultaneous confidence sets. The heat maps are based on estimates of $\bar{y}_{c25}$, the mean percentile rank of child's average household income for 2014-2015, for the full set of CZs. **Panel A:** the map is constructed by dividing the CZs into deciles based on the estimated values of $\bar{y}_{c25}$, and shading the areas so that lighter colors correspond to higher absolute mobility or, equivalently, lower ("better") rank. **Panel B** (**Panel C**) shows the lower (upper) endpoint of the 95% simultaneous confidence sets for the ranks of all CZs, using the same color coding as for the estimated ranks in Panel A.

CZs where children have opportunities to succeed from those without such opportunities. Notable exceptions include many of the individual CZs in the Southeast and in the Great Plains, where mobility is relatively low and high, respectively. In these regions, it is often possible to determine with 95% confidence whether a particular CZ has relatively high or low mobility compared to other CZs in the United States.

We investigate these tentative conclusions in greater depth in Figure 14. For each CZ, we compute the difference between the lower and the upper endpoint of the 95% simultaneous confidence set. Next, we plot these differences against the estimated ranks of the CZs. The larger the difference, the less we know about the ranking of a CZ. To ease interpretation, we normalize the differences by the number of CZs. Thus, a difference of 1 means one cannot determine with 95% confidence whether a CZ has the highest or the lowest income mobility in the United States. By comparison, a difference of 0 means we can be 95% confident in the exact rank of the CZ.

As evident from Figure 14, the results tend to be much more informative in the upper and the lower parts of the ranking. In other words, we can be most confident in conclusions about which CZs that have the highest or the lowest income mobility. One possible explanation of this finding is that $\bar{y}_{c25}$ is more precisely estimated among the CZs that rank at the top and at the bottom. As shown in Appendix G.5, this explanation is at odds with the data. The standard errors of $\bar{y}_{c25}$ are not particularly small for these CZs. Instead, the explanation is that the point estimates of $\bar{y}_{c25}$ differ more across the CZs in the upper and the lower parts of the ranking as compared to the CZs in the middle of the ranking.

A limitation of Figure 14 is that it only shows where in the ranking the results are most informative, not where in the United States. Thus, Figure 15 is useful because it highlights which of the spatial patterns of income mobility are robust to accounting for uncertainty in the estimates of $\bar{y}_{c25}$. The heat map in Panel A is constructed by assigning the CZs to groups depending on the lower and upper endpoints of the simultaneous confidence sets.[13] A CZ is assigned to a high mobility group if the upper endpoint of its simultaneous confidence set is in the top half of the national ranking of CZs, i.e., when the confidence set lies entirely in the top half of the ranking, indicating high mobility. A CZ is assigned to a low mobility group if the lower endpoint of its simultaneous confidence set is in the bottom half of the national ranking of CZs, i.e., when the confidence set lies entirely in the bottom half of the ranking, indicating low mobility. Grey colors represent the CZs with simultaneous confidence sets such that the places cannot be assigned to either the high or the low mobility group. The heat map in Panel B is constructed in the same way, except the high (low) mobility group is now defined as the top (bottom) quartile in the national ranking of the CZs.

The results in Figure 15 calls for caution on concluding whether an individual CZ has high or low income mobility compared to other CZs in the United States. In the national ranking of places by income mobility, it is rarely possible to tell with 95% confidence if a given CZ has relatively high or low income mobility compared to other CZs. There are, however, two main exceptions. With 95% confidence, it is often possible to identify individual CZs with relatively low mobility in the Southeast and individual CZs with relatively high mobility in the Great Plains.

As shown in Appendices G.5 and G.6, the national ranking becomes largely uninformative if one defines a neighborhood to be a county or if one uses the movers estimates. In other words, it is not possible to draw firm conclusions about which counties in the United States have relatively high or low values of $\bar{y}_{c25}$. Nor is

[13] In Appendix Figure 38, we show that this heat map does not materially change if we assign CZs to groups based on marginal confidence sets.

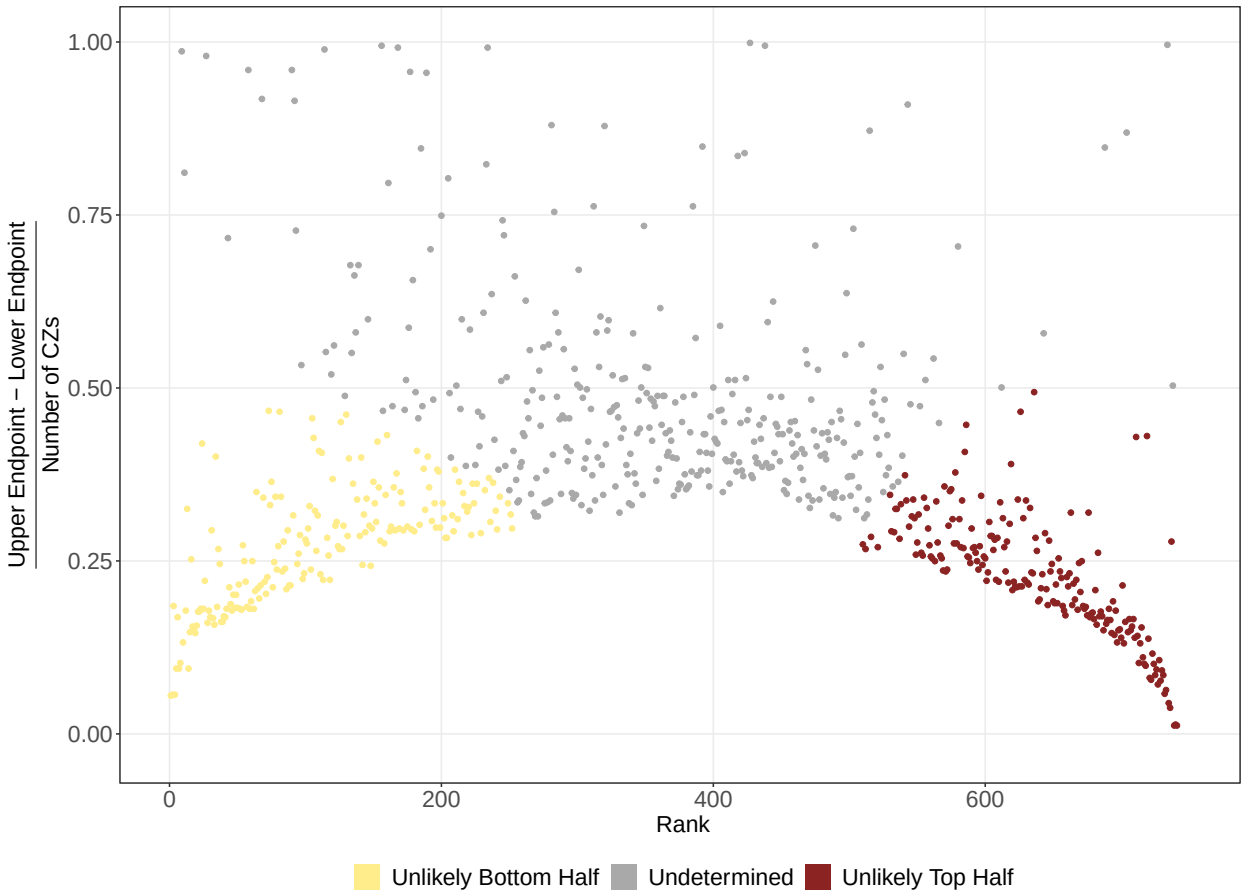

Figure 14: For each CZ, we compute the difference between the upper and the lower endpoint of the 95% simultaneous confidence set. Next, we plot these differences against the estimated ranks of the CZs. To ease interpretation, we normalize the differences by the number of CZs. Thus, a difference of 1 means one cannot tell whether a CZ has the highest or the lowest income mobility in the United States. By comparison, a difference of 0 means we can be confident in the exact rank of the CZ. Each dot in the graph represents a CZ. The CZ is assigned to a high mobility group (light color) if the upper endpoint of its simultaneous confidence set is in the top half of the national ranking. The CZ is assigned to a low mobility group (red color) if the lower endpoint of its simultaneous confidence set is in the bottom half of the national ranking. Grey colors represent the CZs with simultaneous confidence sets such that the places cannot be assigned to either the high or the low mobility group.

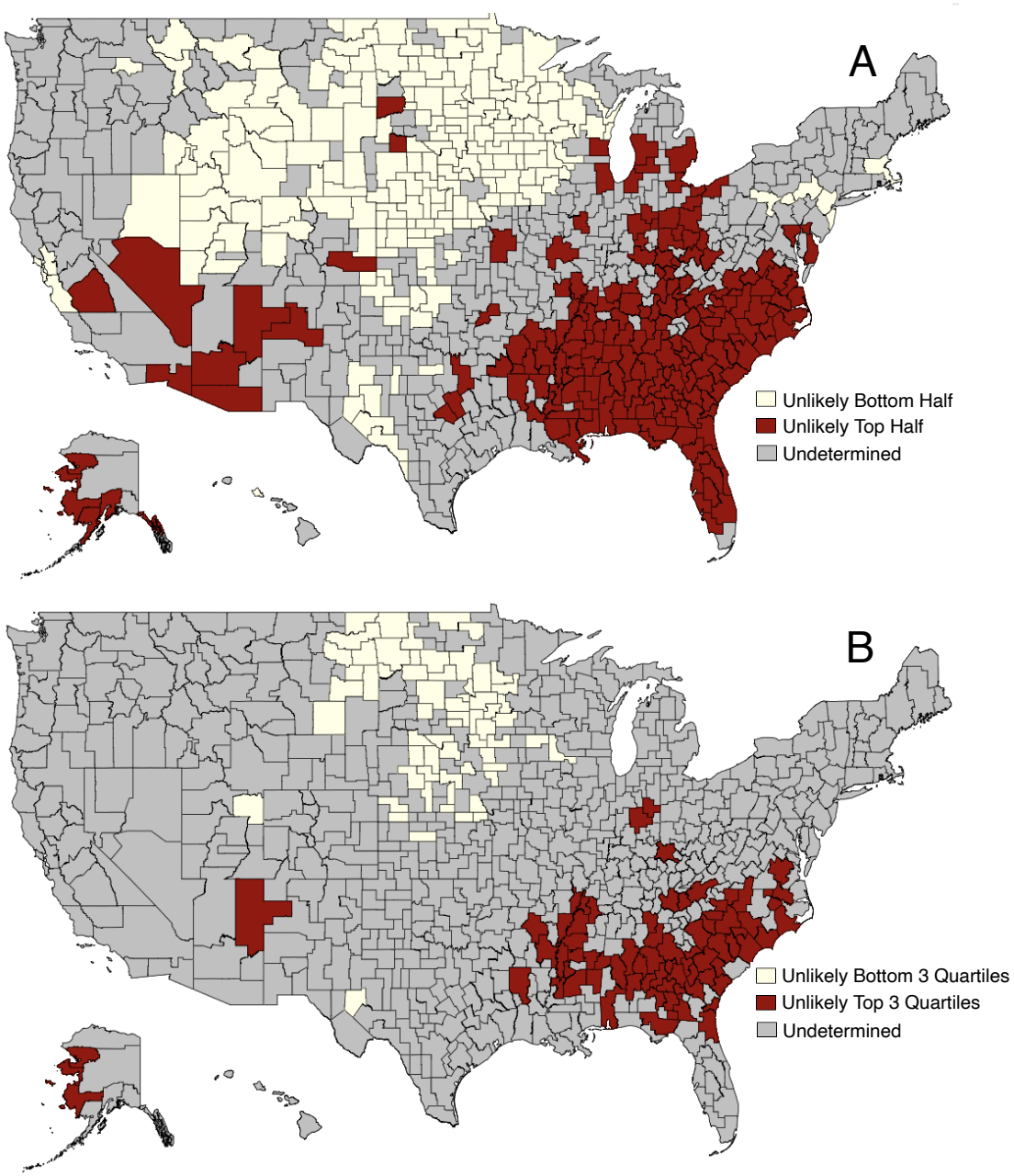

Figure 15: The heat map in **Panel A** is constructed by assigning the CZs to groups depending on the lower and upper endpoints of the simultaneous confidence sets. A CZ is assigned to a high mobility group, **Unlikely Bottom Half**, if the upper endpoint of its simultaneous confidence set is in the top half of the national ranking of CZs, i.e., when the confidence set lies entirely in the top half of the ranking, indicating high mobility. A CZ is assigned to a low mobility group, **Unlikely Top Half**, if the lower endpoint of its simultaneous confidence set is in the bottom half of the national ranking of CZs, i.e., when the confidence set lies entirely in the bottom half of the ranking, indicating low mobility. Grey colors represent the CZs with simultaneous confidence sets such that the places cannot be assigned to either the high or the low mobility group, i.e., the **Undetermined** CZs. The heat map in **Panel B** is constructed in the same way, except the high and low mobility groups are now defined in terms of top and bottom quartiles in the national ranking of the CZs. Thus, we refer to these groups as **Unlikely Bottom 3 Quartiles** and **Unlikely Top 3 Quartiles**.

it possible to say much about which CZs or counties produce more or less upward mobility as measured by the exposure effects μ_{c25}.

Remark 5.2. Since both the correlational and the movers estimates are constructed from tax records covering most of the U.S. population, one may wonder why we assess uncertainty in the ranks stemming from uncertainty in such estimates. We follow the original papers Chetty et al. (2018) and Chetty and Hendren (2018) by using their reported standard errors, for which the authors must have had in mind some underlying source of randomness that leads to uncertainty in their estimates. Potential such sources could be: (i) the U.S. population is a sample from a superpopulation, or (ii) some of the explanatory variables of outcomes (child's income rank) are not fixed and could have experienced different realizations. The description of how standard errors are computed in Chetty et al. (2018) and Chetty and Hendren (2018) suggests that the authors view (i) as the main source of randomness. However, (ii) also seems relevant in the present context. For instance, for any given parent, the income rank could take on various values, but only one is observed, which means only one of the many potential outcomes (child's income rank) is observed. The randomness in parents' income rank therefore induces randomness in estimates of a neighborhood's mobility. Similarly, any given family could, in principle, move or not move from one neighborhood to another and it may do so at different points in time. Since we observe only one realization, we also only observe one of the many potential outcomes (child's income rank). The randomness in the decision to move and its timing therefore induce randomness in the movers estimates of a neighborhood's mobility. Note, however, that assessing the uncertainty from (ii) rather than (i) requires different methods for computing standard errors (Abadie et al. (2020)) and access to the microdata. ■

Remark 5.3. As discussed in the introduction and in Remark 5.1, our methods and those developed by Andrews et al. (2018) share some technical similarities, but answer distinct economic questions, and should thus be viewed as complements, not substitutes. To illustrate this empirically, we apply the methods of Andrews et al. (2018) to the correlational estimates and construct 95% confidence sets for the true mobility of the CZ with the highest estimated mobility.[14] For instance, their 95% "conditional" confidence set for the true mobility of the "winning" CZ among all CZs is (-2.70, 0.66). Since the confidence set includes zero (the smallest possible value of the mobility measure), one cannot be confident that the sample "winner" truly has high mobility. The corresponding confidence set for the "winning" CZ among only the 50 most populous CZs, San Francisco, is (0.389, 0.457). While this confidence set excludes zero, comparing it to the range of estimates for the 50 most populous CZs, it is still fairly wide. Taken together, the results using Andrews et al. (2018)'s methods suggest there is considerable statistical uncertainty about the true value of upward mobility at the top of the estimated ranking of CZs, even if one restricts the study to the 50 most populous CZs.

However, just like the marginal confidence interval for the "winner" displayed in Figure 10, Andrews et al. (2018)'s confidence sets do not allow us to draw any conclusions about what is the true rank of the sample "winner" nor which CZ has true rank one. In contrast, our confidence set for the rank of San Francisco among the 50 most populous CZs tells us that (with 95% probability) its true rank lies between 1 and 2. In addition, our τ-best confidence set for $\tau = 1$ shows that (with 95% probability) there are only 4 CZs that could be the best. It is interesting to note that our confidence sets for the ranks are very narrow even though Andrews et al. (2018)'s confidence sets for expected mobility of the "winner" are fairly wide. This finding illustrates that it can be possible to achieve a statistically informative ranking even if one cannot draw firm

[14] We are grateful to the authors for sharing their code, allowing us to easily apply their methods to our applications.

conclusions about the true value of the sample "winner". See Appendix G.4 for more results and details. ■

5.3 Illustrating the Policy Implications of the Uncertainty in the Rankings

The estimates of Chetty et al. (2014, 2018) and Chetty and Hendren (2018) have been highly influential both among policymakers and researchers. For example, the rankings of neighborhoods by (point estimates of) intergenerational mobility play a key role in Chetty's 2014 Testimony for the United States Senate Committee on the Budget (Chetty, April 1, 2014). In this testimony (pages 6 and 7), he emphasizes that policy should target areas that are ranked at the bottom of the league tables based on their estimates of upward mobility:

"Since rates of upward mobility vary widely across cities, place-based policies that focus on specific cities such as Charlotte or Milwaukee may be more effective than addressing the problem at a national level."

and, moreover, that it is key to disseminate information about which areas have relatively high and low estimates of upward mobility:

"Perhaps the most cost-effective way to improve mobility may be to publicize local statistics on economic mobility and other related outcomes. Simply drawing attention to the areas that need improvement can motivate local policy makers to take action. Moreover, without such information, it is difficult to determine which programs work and which do not. The federal government is well positioned to construct such statistics at minimal cost with existing data. The government could go further by offering awards or grants to areas that have substantially improved their rates of upward mobility. Shining a spotlight on the communities where children have opportunities to succeed can enable others to learn from their example and increase opportunities for economic mobility throughout America."

In light of the large degree of uncertainty, however, one may be concerned that such local statistics (e.g., the league tables and heat maps) do not necessarily contain valuable information about upward mobility. As a consequences of this uncertainty, it can also be problematic to use such statistics to disseminate information or target interventions. The spotlight might be shining on noise, not signal.

In order to illustrate the issues that may arise if one chooses to design interventions based on the local mobility statistics, we next re-visit the recent Creating Moves to Opportunity (CMTO) experiment of Bergman et al. (2019). This experiment is a collaboration between researchers and public housing authorities to introduce and evaluate interventions to "create moves to opportunity" for low-income families.

The CMTO experiment

The motivation for the CMTO experiment is the argument that low-income families tend to live in neighborhoods with low upward mobility. In order to understand how policy may be designed to help low-income families move to neighborhoods with higher mobility rates, the authors perform a randomized controlled trial with housing voucher recipients in Seattle and King County. A treatment group of low-income families were offered assistance and financial support to find and lease units in areas that were classified as high upward-mobility neighborhoods.

The authors "define high upward-mobility neighborhoods as Census tracts that have point estimates of upward income mobility in approximately the top one-third among tracts in the Seattle and King County

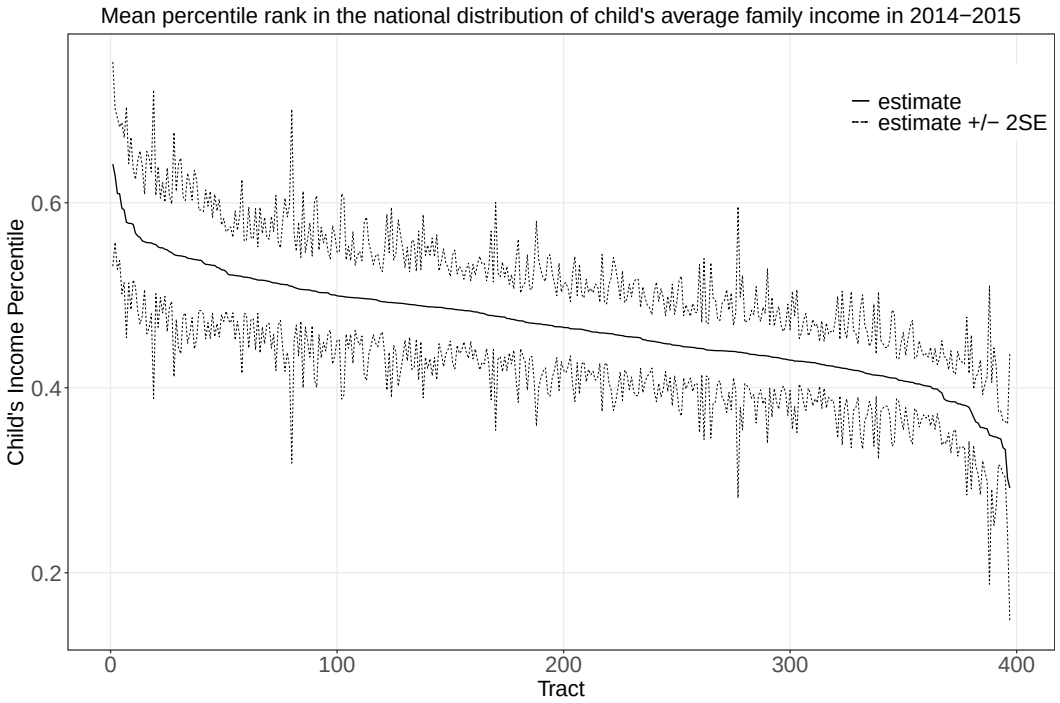

Figure 16: Estimates of $\bar{y}_{c25}$, the expected percentile rank of child's average household income for 2014-2015 in the national distribution of her cohort, with marginal confidence intervals (estimates plus or minus twice the standard errors) for all 397 Census tracts in Seattle and King County.

area" (Bergman et al., 2019, p. 10, Section III.B Defining Opportunity Areas).[15] Following this definition, we use the estimates of $\bar{y}_{c25}$ for the 397 tracts in the Seattle and King County area to classify areas as upward-mobility neighborhoods.[16] Figure 16 plots these estimates alongside marginal confidence intervals (estimates plus or minus twice the standard errors). The point estimates vary considerably, but the standard errors are relatively large. Figure 17 presents a map of Seattle and King County which shows the location of the 132 tracts that have point estimates of upward income mobility in the top one-third, and, as a result, are classified by us as high upward-mobility neighborhoods.

Remark 5.4. Chetty et al. (2020) show that mobility estimates across tracts are (significantly) spatially correlated even though the estimates are computed from samples that are independent across tracts. They argue that, therefore, the estimates in Figure 16 are more precise than their own standard errors suggest.

We do not agree with this conclusion for the following reason. For simplicity consider two tracts $c = A, B$. From each tract we draw an i.i.d. sample from the distribution F_c in tract c, independently of the data in the

[15] In practice, the authors make a few adjustments to this definition. We refer to Appendix A of Bergman et al. (2019) for a discussion of these adjustments.

[16] These estimates are available at https://www.opportunityatlas.org/. To address concerns about noise in the correlational estimates Bergman et al. (2019) also construct forecasts of upward mobility at the tract level. The forecasts are weighted averages of the correlational estimates of $\bar{y}_{c25}$ and predictions based on observable characteristics of the tracts, with greater weight on the correlational estimates when they are more precisely estimated. Given that most correlational estimates are very noisy, the forecast estimates are very similar to the predictions based on covariates. For two reasons, we do not consider the forecasts. First of all, Bergman et al. (2019) do not report standard errors for the forecasts, which we would need for the implementation of our method. Second, forecasts are based on covariates such as the estimated mobility rate of the tract based on information on children's income around age 22. We conjecture that this estimate of lagged upward mobility suffers from similar amounts of noise as the estimates of mobility when children are in their 30s. Therefore, we expect the forecasts of upward mobility to be similarly noisy as the estimates of $\bar{y}_{c25}$ that we use.

Figure 17: This map of Seattle and King County shows the location of the 132 tracts that were classified as high upward-mobility neighborhoods. High-upward-mobility neighborhood consists the Census tracts with estimates of $\bar{y}_{c25}$ among the top one-third of the tracts in the Seattle and King County area.

other tract. We then construct estimates $\hat{\bar{y}}_{c25}$ of the mobility measure $\bar{y}_{c25}$, the latter being a feature of the distribution F_c. Therefore, the estimates $\hat{\bar{y}}_{A25}$ and $\hat{\bar{y}}_{B25}$ being similar must either come from finite-sample bias in the estimates or from true mobility $\bar{y}_{A25}$ and $\bar{y}_{B25}$ being similar. The latter might occur, for instance, when the distributions F_A and F_B are similar. Therefore, similarity of mobility estimates across tracts is unrelated to the precision with which they are estimated.

Chetty et al. (2020) also argue that there is significant variation in true mobility. This argument involves decomposing the variance of estimates across tracts, $\hat{\bar{y}}_{c25}$, $c = 1, 2, \ldots$, into the variance of true mobility $\bar{y}_{c25}$, $c = 1, 2, \ldots$, (the "signal") and into the variance of "noise". Since "noise" turns out to be low in this decomposition, they argue that therefore there is large variation in the signal, i.e. true mobility. However, a small variance of the "noise" across tracts does not imply that the precision of any particular tract-level mobility estimate is small.[17] ■

[17] In fact, Chetty et al. (2018) estimate the variance of "noise" as the average squared standard error of the tract-level estimates (after accounting for additional noise that was introduced to guarantee privacy of individuals in the sample). Therefore, a small "noise" variance implies that the average of the standard errors is small, but not necessarily any individual standard error.

Panel A: Tracts

	all tracts				50 most populous tracts			
κ	top group	τ-best	τ-worst	bottom group	top group	τ-best	τ-worst	bottom group
1/3	1 . . . 132	1 . . . 395	1 . . . 397	266 . . . 397	1 . . . 17	1 . . . 50	1 . . . 50	34 . . . 50
1/4	1 . . . 99	1 . . . 392	1 . . . 397	299 . . . 397	1 . . . 12	1 . . . 47	3 . . . 50	39 . . . 50
1/10	1 . . . 40	1 . . . 390	3 . . . 397	358 . . . 397	1 . . . 5	1 . . . 42	12 . . . 50	46 . . . 50
1/100	1 . . . 4	1 . . . 360	116 . . . 397	394 . . . 397				

Panel B: CZs

	all CZs				50 most populous CZs			
κ	top group	τ-best	τ-worst	bottom group	top group	τ-best	τ-worst	bottom group
1/3	1 . . . 247	1 . . . 446	289 . . . 741	495 . . . 741	1 . . . 17	1 . . . 20	31 . . . 50	34 . . . 50
1/4	1 . . . 185	1 . . . 367	396 . . . 741	557 . . . 741	1 . . . 12	1 . . . 15	37 . . . 50	39 . . . 50

Table 5: **Panel A:** "top group" shows the indices of the κ fraction of tracts with the largest estimated mobility (among the 50 most populous tracts and among all tracts, respectively). "τ-best" shows the indices of the tracts that are in the 95% confidence set for the τ-best tracts, where τ is set to κ times the number of tracts. Similarly, "bottom group" and "τ-worst" show the indices of the κ fraction of tracts with the smallest estimated mobility and the indices of the tracts in 95% confidence set for the τ-worst. **Panel B:** shows the corresponding results of Panel A for CZs instead of tracts. The confidence sets are implemented as described in Appendix F, using the stepwise procedure ("DM.step").

Does the classification of an opportunity neighborhood reflect noise or signal?

In light of our previous findings, one might worry that a tract defined as a high upward-mobility neighborhood does not have statistically higher mobility rates as compared to the other tracts. To examine this, we compute a 95% confidence set for the τ-best tracts in the Seattle and King County, where τ is set equal to 132 (approximately one-third). The confidence set is implemented as described in Appendix F, using the stepwise procedure ("DM.step"). The result is shown in the first row of Panel A of Table 5. We find that all but 2 out of 397 tracts could be among the top one-third, and, as a result, be classified as high upward-mobility neighborhood according to our definition. In addition, all 397 tracts lie in the confidence set for the τ-worst. Thus, we conclude the classification of a given tract as a high upward-mobility neighborhood may simply reflect statistical uncertainty ("noise") rather than particularly high mobility ("signal").

A natural question is whether an alternative definition could have reduced the noise in the classification. We investigate this in the second to fourth rows of Table 5, showing the 95% confidence sets for the τ-best and the τ-worst when τ is set to 1/4, 1/10, or 1/100 of the tracts. We see that 360 tracts could even be among the top-4 (approximately, 1/100 of the tracts) tracts and that 394 could be among the worst-4. The results are very similar when we try to determine the top 1/3, 1/4, or 1/10 only among the 50 most populous tracts. Taken together, these results suggest that changes in the definition of upward-mobility neighborhoods at the tract-level is unlikely to reduce the noise in the classification.

To show how, in principle, the τ-best and τ-worst confidence sets could be used to select treatment and control groups consider Panel B of Table 5. It shows the analogous results to those in Panel A for CZs. Among the 50 most populous CZs, the confidence sets for the τ-best and the τ-worst when $\tau = 17$ (approximately one-third of the CZs) do not overlap. The CZs 1...20 could be among the top one-third while CZs 31...50 could be among the bottom one-third. Therefore, we can be confident that none of the tracts classified as top one-third are in fact among the bottom one-third and vice versa, which means we

could define the top/bottom one-third of the 50 most populous CZs as treatment/control group.

Among all CZs, the confidence sets for the τ-best and τ-worst when $\tau = 247$ (one-third of the CZs) do overlap. However, the same confidence sets for $\tau = 185$ (approximately one-fourth) do not overlap, which means we could define the top-/bottom-fourth of all CZs as treatment/control group and be confident that none of the treatment CZs is in fact among the bottom one-third and none of the control CZs is in fact among the top one-third.

Implications of the noise in the classification of an opportunity neighborhood

This noise in the classification raises the question of whether, and under what conditions, one could be confident that CMTO would actually help families move to high upward-mobility neighborhoods, prior to the experiment taking effect. In other words, assuming that the correlational estimates of tract-level upward mobility can be given a causal interpretation, would one then expect a positive average treatment effect of CMTO?

The answer to this question depends on what assumptions, if any, one is willing to make about the location choices of the treatment group in response to CMTO.[18] Without assumptions on these location choices, it is insufficient to test whether average upward mobility among tracts in the top one-third (the treatment group) is higher than in the bottom two-thirds (the control group). Rather, the key question becomes whether some of the tracts in the bottom two-thirds can have mobility higher than the tracts in the top one-third. To see why, consider the following two examples.

Suppose average mobility among tracts in the top one-third is higher because of one tract having high mobility while all other tracts in the top one-third have mobility lower than those in the remaining two-thirds. In that case, only families moving to that single high-mobility tract are treated with higher mobility while all other families are treated with lower mobility. On average, the families that moved to top one-third tracts because of the experiment may therefore have moved to neighborhoods with lower mobility. Of course, whether or not this is the case will depend on the distribution of mobility across neighborhoods to which families actually moved. Without making assumptions about the individual tracts the families would move from and to as a result of the experiment, the average treatment effect can be zero, positive or negative.

Suppose instead the average in the top one-third is higher because all the tracts in the top one-third have high mobility except one tract that has mobility lower than those in the remaining two-thirds. In that case, families moving to that single low-mobility tract are treated with lower mobility while all other families are treated with higher mobility. On average, families that moved to the top one-third tracts because of the experiment may again have moved to neighborhoods with lower mobility. As before, whether or not this is the case will depend on the distribution of mobility across neighborhoods to which the families would move from and to as a result of the experiment. Without making assumptions about this, the average treatment effect can again be zero, positive or negative.

Of course, these are extreme cases but they help make an important point: Whether the tracts in the top one-third have higher rates of upward mobility on average is neither sufficient nor necessary for low

[18] Given the realized location choices of the families in CMTO, one could in principle check or modify any assumption made. We do not do this, however, as data on location choice are not available to us.

income families to be moving to neighborhoods with higher upward mobility as a result of the experiment.[19] Without making assumptions about the individual tracts that the treated families would move from and to in response to the experiment, the relevant condition to test is that none of the tracts in the bottom two-thirds can have mobility higher than the tracts in the top one-third. Above, we examined this, finding that if one uses a confidence level of 95%, all but 2 out of 397 tracts could be among the top one-third. This finding suggests that assumptions on location choices indeed are needed to be statistically confident at conventional levels, prior to the experiment, that CMTO would have a positive impact.

One possible assumption to make is that the treated families move from randomly selected tracts in the bottom two-thirds to randomly selected tracts in the top one-third. Under this assumption, the relevant condition to test is whether the average upward mobility among tracts in the top one-third is higher than in the bottom two-thirds. However, when performing this test, it is necessary to address the concern that the classification is based on the estimated levels of upward mobility, not the true ones. To do so, one may use the methods of Andrews et al. (2018). When we apply their method to the CMTO data, we estimate a 95% conditional confidence set of (-0.0305; 0.0911) for the difference between the average upward mobility of the top one-third and the bottom two-thirds of tracts. This means that conditional on the identities of tracts in the top one-third we cannot rule out with 95% confidence that they have lower average upward mobility than tracts in the bottom two-thirds.[20]

Interestingly, we reach a different conclusion if we rely on the unconditional confidence sets instead of the conditional ones. We estimate the unconditional hybrid confidence set to be (0.0342, 0.0908) for the difference between the average upward mobility of the top one-third and the bottom two-thirds of tracts. As argued by Andrews et al. (2018), the choice between conditional and unconditional inference methods is necessarily context-specific, as it depends on the extent to which we care about validity conditional on selecting a given target. In the context of CMTO, conditional inference is relevant to answer the question of the expected effect of moving to the tracts that were actually classified as high upward-mobility neighborhoods. Unconditional inference, on the other hand, is relevant if the primary goal is to assess the efficacy of hypothetical experiments targeting the top one-third of Census tracts, with less focus on the collection of tracts that were actually classified as high upward-mobility neighborhoods in CMTO.

6 Concluding remarks

In this paper we show how to account for uncertainty in the ranking of different populations according to the value of some feature of each population. We consider both the problem of constructing marginal confidence sets for the rank of a particular population as well as simultaneous confidence sets for the ranks of all populations. We show how to construct such confidence sets under weak assumptions.

We also provide two empirical examples in which our method produces highly informative confidence sets for ranks. One is the ranking of countries according to the results on the PISA test. The other is the ranking

[19] This argument does not rely on heterogeneous effects of place or non-random mobility. Even if the effect of a given tract is the same for all families or low-income families move randomly to tracts in the top one-third, the average effect of the experiment may very well be negative

[20] Andrews et al. (2018) also analyze the CMTO data. Their analysis differs in two ways. First, they consider all tracts in the Seattle CZ, while we use the same data as in Bergman et al. (2019), which consist of only the tracts inside the Seattle city boundary and King County. Second, they compare the average upward mobility of the top one-third to the average upward mobility of all tracts. We focus instead on the the top one-third versus the bottom two-third, since this contrast maps directly to the assumption of location choices that we consider.

of the most populous commuting zones or counties in the United States according to upward mobility. Such rankings by upward mobility, however, become much less informative if one includes all commuting zones of counties, if one defines neighborhoods with even more granularity (e.g., by considering Census tracts), or if one uses movers across areas to address concerns about selection.

A natural question is why it is difficult to achieve an informative ranking in certain cases. Based on our simulations, in which our confidence sets in some cases cover the true ranking with probability close to one, one may be concerned that this phenomenon stems from a lack of power of our procedures. We emphasize, however, that these situations may arise precisely when the ranking is most informative. To see this, consider the case in which standard errors of the mobility estimates, say, are nearly zero for all neighborhoods (relative to the differences in mobility estimates across neighborhoods). Due to the discreteness of the ranks, the ranking has essentially no uncertainty and any "reasonable" confidence set should cover the true ranking with probability (close to) one. In other situations when there is more uncertainty in the ranking, our method achieves coverage closer to the nominal level. These features are borne out in our simulations.

We therefore argue that a more appropriate explanation for why it may not be possible to achieve an informative ranking is that researchers may simply be demanding too much from the data. This explanation is most plausible when estimates vary substantially across populations but standard errors are large, when standard errors are small but the estimates do not vary much across populations, or when both standard errors are large and estimates do not vary much across populations. To think about when our (or any) approach will deliver an informative ranking, a useful starting point is the naive pairwise comparisons that ignore the multiple testing issue. Take, for example, the 397 tracts in the CMTO experiment in Seattle. To obtain a complete ranking of these tracts by upward mobility, it is necessary to compare 78,606 unique pairs. The problem is that at most 30.2% of these pairs consist of tracts that differ significantly at the 95% significance levels. Importantly, the conclusion that 30.2% of the pairs are significantly different ignores that one has performed 78,606 comparisons, so even by chance, many of these comparisons will show up as significant when in fact they are not. Indeed, when taking the multiple testing into account, it is clear the uncertainty is too large to achieve an informative ranking of the tracts in Seattle according to upward mobility.

Appendix A The "Naive" Bootstrap Undercovers

In this section, we show that the "naive" bootstrap as described in Remark 3.7 does not satisfy the uniform coverage requirement unless $p = 2$. Furthermore, we show that when there are ties and $p > 2$, then the approach even fails the pointwise coverage requirement for a fixed P and, in fact, the coverage probability tends to zero as p grows.

To simplify the subsequent discussion, we focus on the case in which the estimators of the features $\theta(P_j)$ are independent and normally distributed, $\hat{\theta}_j \sim N\left(\theta(P_j), 1\right)$ for all $j \in J$. In this case, we obtain finite-sample results, but they easily extend to the asymptotic case when $n \to \infty$ and variances are unknown. Suppose $\theta(P_j) = 0$ for all $j \in J$. Consider the parametric bootstrap in which we draw $\hat{\theta}_j^* \sim N(\hat{\theta}_j, 1)$ independently, conditional on the data $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)'$.

Suppose we want to construct a one-sided confidence set for the rank of population j, for which the upper endpoint is equal to p:

$$R_{n,j}^* \equiv \left\{\hat{L}_j^*, \ldots, p\right\} ,$$

where $\hat{L}_j^*$ is the α-quantile of $\hat{r}_j^*$, conditional on the data $\hat{\theta}$. Further suppose all populations are tied with $\theta(P_j) = 0$ for all $j \in J$, so that all ranks are equal to one, $r_j(P) = 1$ for all $j \in J$. For $R_{n,1}^*$ to cover the rank $r_1(P)$, it must be the case that the event

$$E \equiv \left\{\hat{L}_1^* = 1\right\} = \left\{P\{\hat{\theta}_1^* > \max\{\hat{\theta}_2^*, \ldots, \hat{\theta}_p^*\} | \hat{\theta}\} \geq \alpha\right\}$$

holds. Consider first $p = 2$. Then,

$$\begin{aligned} P\left\{\hat{\theta}_1^* > \max\{\hat{\theta}_2^*, \ldots, \hat{\theta}_p^*\} \middle| \hat{\theta}\right\} &= P\left\{\hat{\theta}_1^* > \hat{\theta}_2^* \middle| \hat{\theta}\right\} \\ &= P\left\{\frac{\hat{\theta}_1^* - \hat{\theta}_1 - (\hat{\theta}_2^* - \hat{\theta}_2)}{\sqrt{2}} > \frac{\hat{\theta}_2 - \hat{\theta}_1}{\sqrt{2}} \middle| \hat{\theta}\right\} \\ &= 1 - \Phi\left(\frac{\hat{\theta}_2 - \hat{\theta}_1}{\sqrt{2}}\right) \end{aligned}$$

so that

$$P\{E\} = P\left\{1 - \Phi\left(\frac{\hat{\theta}_2 - \hat{\theta}_1}{\sqrt{2}}\right) \geq \alpha\right\} = 1 - \alpha .$$

Therefore, the bootstrap confidence interval $R_{n,1}^*$ covers $r_1(P)$ with the desired probability.

Now consider $p > 2$. Let M be the index such that $\hat{\theta}_M = \max\{\hat{\theta}_2, \ldots, \hat{\theta}_p\}$. First note that

$$\begin{aligned} P\left\{\hat{\theta}_1^* > \max\{\hat{\theta}_2^*, \ldots, \hat{\theta}_p^*\} \middle| \hat{\theta}\right\} &< P\left\{\hat{\theta}_1^* > \hat{\theta}_M^* \middle| \hat{\theta}\right\} \\ &= P\left\{\frac{\hat{\theta}_1^* - \hat{\theta}_1 - (\hat{\theta}_M^* - \hat{\theta}_M)}{\sqrt{2}} > \frac{\hat{\theta}_M - \hat{\theta}_1}{\sqrt{2}} \middle| \hat{\theta}\right\} \\ &= 1 - \Phi\left(\frac{\hat{\theta}_M - \hat{\theta}_1}{\sqrt{2}}\right) \end{aligned} \tag{35}$$

Let F be the event

$$F \equiv \left\{1 - \Phi\left(\frac{\hat{\theta}_M - \hat{\theta}_1}{\sqrt{2}}\right) \geq \alpha\right\} .$$

Clearly, by the strict inequality in (35), $E \subset F$ and $P\{E\} < P\{F\}$. Letting $z_{1-\alpha}$ be the $(1-\alpha)$-quantile of the standard normal distribution, we have

$$P\{F\} = P\left\{\frac{\hat{\theta}_M - \hat{\theta}_1}{\sqrt{2}} \leq z_{1-\alpha}\right\} = P\left\{\frac{\max\{\hat{\theta}_2, \ldots, \hat{\theta}_p\} - \hat{\theta}_1}{\sqrt{2}} \leq z_{1-\alpha}\right\} ,$$

which, for example, is strictly less than $P\{(\hat{\theta}_2 - \hat{\theta}_1)/\sqrt{2} \leq z_{1-\alpha}\} = 1 - \alpha$. Therefore, $P\{E\} < 1 - \alpha$ and the confidence set $R_{n,1}^*$ does not cover the rank $r_1(P)$ with the desired probability. Moreover, as $p \to \infty$, $\max\{\hat{\theta}_2, \ldots, \hat{\theta}_p\} \to \infty$ in probability, so the coverage probability tends to zero.

Appendix B An Alternative Construction of Confidence Sets for Ranks

Let $\widetilde{C}_n(1-\alpha)$ be a confidence set that is rectangular in the sense that

$$\widetilde{C}_n(1-\alpha) = \prod_{j\in J} \widetilde{C}_n(1-\alpha, j)$$

for suitable sets $\{\widetilde{C}_n(1-\alpha, j)\colon j \in J\}$, and that simultaneously covers the vector of features $\theta(P)$ with limiting probability $1-\alpha$:

$$\liminf_{n\to\infty} \inf_{P\in\mathbf{P}} P\{\theta(P) \in \widetilde{C}_n(1-\alpha)\} \geq 1-\alpha \ .$$

For instance, the confidence set could be constructed as in Example 3.7 of Romano and Shaikh (2012). Define

$$\widetilde{N}_j^- \equiv \left\{k \in J\setminus\{j\}\colon \widetilde{C}_n(1-\alpha, j) \text{ lies entirely below } \widetilde{C}_n(1-\alpha, k)\right\}$$
$$\widetilde{N}_j^+ \equiv \left\{k \in J\setminus\{j\}\colon \widetilde{C}_n(1-\alpha, j) \text{ lies entirely above } \widetilde{C}_n(1-\alpha, k)\right\} \ .$$

Then, it is easy to see that

$$\liminf_{n\to\infty} \inf_{P\in\mathbf{P}} P\left\{|\widetilde{N}_j^-| + 1 \leq r_j(P) \leq p - |\widetilde{N}_j^+|\right\} \geq 1-\alpha$$

and

$$\liminf_{n\to\infty} \inf_{P\in\mathbf{P}} P\left\{\bigcap_{j\in J}\left\{|\widetilde{N}_j^-| + 1 \leq r_j(P) \leq p - |\widetilde{N}_j^+|\right\}\right\} \geq 1-\alpha \ .$$

Therefore,

$$\widetilde{R}_{n,j} \equiv \left\{|\widetilde{N}_j^-| + 1, \ldots, p - |\widetilde{N}_j^+|\right\} \tag{36}$$

is a confidence set that covers the rank $r_j(P)$ with limiting probability at least $1-\alpha$ and

$$\widetilde{R}_n^{\text{joint}} \equiv \prod_{j\in J}\left\{|\widetilde{N}_j^-| + 1, \ldots, p - |\widetilde{N}_j^+|\right\} \tag{37}$$

is a confidence set that covers the vector of ranks $r(P)$ with limiting probability at least $1-\alpha$.

To formally compare the approach proposed in the main text with the one of this section, we focus on the case in which the estimators of the features $\theta(P_j)$ are independent and normally distributed, i.e.,

$$\begin{pmatrix}\hat{\theta}_1\\ \vdots \\ \hat{\theta}_p\end{pmatrix} \sim N\left(\theta(P), diag\left(\frac{\sigma^2(P_1)}{n_1}, \ldots, \frac{\sigma^2(P_p)}{n_p}\right)\right) \ , \tag{38}$$

where $\sigma^2(P_j) > 0, j \in J$ are known. In this case, we obtain finite-sample comparisons between the two methods, but the results easily extend to asymptotic (as $n\to\infty$) comparisons and to the case of unknown variances.

We consider the confidence set for the entire ranking, $\widetilde{R}_n^{\text{joint}}$, based on $\widetilde{C}_n(1-\alpha)$, where $\widetilde{C}_n(1-\alpha) = \prod_{j\in J}\widetilde{C}_n(1-\alpha, j)$ is such that

$$\widetilde{C}_n(1-\alpha, j) = \left[\hat{\theta}_j \pm \frac{\sigma(P_j)}{\sqrt{n_j}}\widetilde{q}_{1-\alpha}\right] \ , \quad j \in J \ , \tag{39}$$

and $\widetilde{q}_{1-\alpha}$ is either the

$$\frac{1+(1-\alpha)^{1/p}}{2} - \text{quantile of the } N(0,1) \text{ distribution} \tag{40}$$

or the

$$\left(1-\frac{\alpha}{2p}\right) - \text{quantile of the } N(0,1) \text{ distribution} \ . \tag{41}$$

The quantile in (40) imposes independence of the estimators and (41) is the quantile used in the Bonferroni method. We compare $\widetilde{R}_n^{\text{joint}}$ to our confidence set R_n^{joint} based on $C_n(1-\alpha, S_{\text{all}})$ with $C_n(1-\alpha, S_{\text{all}}) = \prod_{(j,k)\in S_{\text{all}}} C_n(1-\alpha, S_{\text{all}}, (j,k))$ such that

$$C_n(1-\alpha, S_{\text{all}}, (j,k)) = \left[\hat{\theta}_j - \hat{\theta}_k \pm \sqrt{\frac{\sigma^2(P_j)}{n_j} + \frac{\sigma^2(P_k)}{n_k}}q_{1-\alpha}\right] \ , \tag{42}$$

where $q_{1-\alpha}$ is the

$$(1-\alpha)-\text{quantile of } \max_{(j,k)\in S_{\text{all}}} \frac{|\hat{\theta}_j - \hat{\theta}_k - \Delta_{j,k}(P)|}{\sqrt{\frac{\sigma^2(P_j)}{n_j} + \frac{\sigma^2(P_k)}{n_k}}} .$$

This quantile is similar to $L^{-1}_{\text{symm,n}}(1-\alpha, S_{\text{all}}, P)$ as defined in (8) except that the estimated variances in (8) are replaced by the population variances.

For the following lemma, let $R(P) \equiv \prod_{j\in J}\{\underline{r}_j(P), \ldots, \bar{r}_j(P)\}$, where $\underline{r}_j(P)$ and $\bar{r}_j(P)$ are the smallest and largest ranks defined in Remark 3.6.

Lemma B.1. *Suppose (38) holds. Let $\widetilde{R}_n^{\text{joint}}$ be based on $\widetilde{C}_n(1-\alpha) = \prod_{j\in J}\widetilde{C}_n(1-\alpha, j)$ satisfying (39) with $\widetilde{q}_{1-\alpha}$ as defined in either (40) or (41). Let R_n^{joint} be based on $C_n(1-\alpha, S_{\text{all}}) = \prod_{(j,k)\in S_{\text{all}}} C_n(1-\alpha, S_{\text{all}}, (j,k))$ satisfying (42). Then the following statements hold:*

(i) For any $\alpha \in (0,1)$, $\widetilde{R}_n^{\text{joint}}$ satisfies $P(r(P) \in \widetilde{R}_n^{\text{joint}}) \geq P(R(P) \subseteq \widetilde{R}_n^{\text{joint}}) > 1-\alpha$.

(ii) For any $\alpha \in (0,1)$, R_n^{joint} satisfies $P(r(P) \in R_n^{\text{joint}}) \geq P(R(P) \subseteq R_n^{\text{joint}}) \geq 1-\alpha$, where the second inequality is satisfied with equality when all elements of $\theta(P)$ are equal.

(iii) If $p = 2$, then R_n^{joint} is a subset of $\widetilde{R}_n^{\text{joint}}$, and a strict subset with positive probability.

(iv) If $\sigma(P_j) = \sigma(P_k)$ for all $j,k \in J$ and $n_j = n_k$ for all $j,k \in J$, then R_n^{joint} is a subset of $\widetilde{R}_n^{\text{joint}}$, and a strict subset with positive probability.

This lemma shows, first, that the alternative confidence set $\widetilde{R}_n^{\text{joint}}$ covers the ranking $r(P)$ and the set of ranks $R(P)$ each with probability strictly larger than $1-\alpha$, independently of the configuration of features $\theta(P_1), \ldots, \theta(P_p)$. On the other hand, our proposed confidence set for the set of ranks achieves coverage probability equal to $1-\alpha$ in the case when all features $\theta(P_j)$ are equal. In addition, there are two special cases in which our approach leads to bounds on the ranks that are not wider and strictly narrower than the alternatives proposed in this section: when there are only two populations or when the variances and sample sizes of all populations are equal.

In the case when $p > 2$, not all variances, and not all sample sizes are equal, then our confidence set and the alternative proposed in this section cover the ranking with probability strictly larger than $1-\alpha$. In this case, we do not know whether our method leads to smaller confidence sets. However, we can compare the endpoints of the two confidence sets as follows. In the proof of the lemma, we show that

$$P\left\{r(P) \in \widetilde{R}_n^{\text{joint}}\right\} \geq P\left\{\theta(P_j) - \theta(P_k) \in \left[\hat{\theta}_j - \hat{\theta}_k \pm (\tau_j + \tau_k)\widetilde{q}_{1-\alpha}\right] \text{ for all } (j,k) \in S_{\text{all}}\right\} > 1-\alpha ,$$

where $\tau_j \equiv \sigma(P_j)/\sqrt{n_j}$. Similarly, our confidence set satisfies

$$P\left\{r(P) \in R_n^{\text{joint}}\right\} \geq P\left\{\theta(P_j) - \theta(P_k) \in \left[\hat{\theta}_j - \hat{\theta}_k \pm \sqrt{\tau_j^2 + \tau_k^2}q_{1-\alpha}\right] \text{ for all } (j,k) \in S_{\text{all}}\right\} \geq 1-\alpha .$$

Comparing the two expressions, it is clear that we cannot have

$$(\tau_j + \tau_k)\widetilde{q}_{1-\alpha} < \sqrt{\tau_j^2 + \tau_k^2}q_{1-\alpha}$$

for all $(j,k) \in S_{\text{all}}$. In particular, if there is one pair (j,k) such that this strict inequality holds, then there must be at least one other pair (j',k') such that the inequality is reversed:

$$(\tau_{j'} + \tau_{k'})\widetilde{q}_{1-\alpha} > \sqrt{\tau_{j'}^2 + \tau_{k'}^2}q_{1-\alpha} .$$

Therefore, the confidence set $\widetilde{R}_n^{\text{joint}}$ cannot be contained in ours. Whether our confidence set is contained in $\widetilde{R}_n^{\text{joint}}$ in general, we leave open for future research.

Remark B.1. Consider the case in which $\theta(P)$ is a vector of expectations and $\hat{\theta}$ the corresponding vector of sample means. Then, the two confidence sets $\widetilde{R}_n^{\text{joint}}$ in Lemma B.1, one using the critical value in (40) and the other the Bonferroni critical value in (41), coincide with the two proposals in Klein et al. (2020). The simulations in Section 4 confirm the results in Lemma B.1 by showing that our confidence sets for ranks are either of similar or strictly smaller size than those by Klein et al. (2020). ■

Appendix C Comparison With Gupta (1956)

Gupta (1956) proposes a confidence set that contains the identity of the population with the largest mean, based on observations from independent, normally distributed populations with equal and known variances:

$$\bar{X}_n \sim N\left(\mu(P), \frac{\sigma^2(P)}{n}\right) , \tag{43}$$

where $\sigma^2(P) > 0$ is known. His confidence set J_n^{Gupta} contains all $j \in J$ such that

$$\max_{k \in J} \bar{X}_{n,k} - \bar{X}_{n,j} \leq d\frac{\sigma(P)}{\sqrt{n}} ,$$

where d solves

$$\int \Phi(u+d)^{p-1}\phi(u)du = 1-\alpha , \tag{44}$$

Φ and ϕ denote the cdf and pdf of the standard normal distribution, and $\alpha \in (0, 1/2)$. Let π be an arbitrary permutation of J such that $\mu(P_{\pi(1)}) \geq \mu(P_{\pi(2)}) \geq \ldots \geq \mu(P_{\pi(p)})$, where $\mu(P_j)$ is the jth element of $\mu(P)$. Gupta shows that the best population, $\pi(1)$, is contained in his confidence set with probability no less than $1-\alpha$:

$$\begin{aligned}
P\left\{\pi(1) \in J_n^{\text{Gupta}}\right\} &= P\left\{\max_{k \in J} \bar{X}_{n,k} - \bar{X}_{n,\pi(1)} \leq d\frac{\sigma(P)}{\sqrt{n}}\right\} \\
&= \int \prod_{k=2}^{p} \Phi_{\pi(k)}\left(x + d\frac{\sigma(P)}{\sqrt{n}}\right) \phi_{\pi(1)}(x)dx \\
&\geq \int \Phi_{\pi(1)}\left(x + d\frac{\sigma(P)}{\sqrt{n}}\right)^{p-1} \phi_{\pi(1)}(x)dx \\
&= \int \Phi\left(u+d\right)^{p-1} \phi(u)du
\end{aligned}$$

where Φ_j and ϕ_j denote the cdf and pdf of the normal distribution with mean $\mu(P_j)$ and variance $\sigma^2(P)/n$. The first equality uses normality and independence of the means. The inequality above follows because, since the populations have equal variances their distributions are stochastically ordered by their means. The final equality is due to a change of variables. Since d is chosen so that the last expression is equal to $1-\alpha$, the coverage probability $P\left\{\pi(1) \in J_n^{\text{Gupta}}\right\}$ is no smaller than $1-\alpha$. The inequality becomes an equality when all means are equal to each other. In this sense, Gupta's approach selects the critical value from the least-favorable configuration of means.

The requirement of covering $\pi(1)$ with probability no less than a prespecified level is not the same as covering the set of 1-best populations, $J_0^{1-\text{best}}$, as defined in Section 3.4. In fact, Gupta's confidence set may cover $J_0^{1-\text{best}}$ with probability strictly less than $1-\alpha$ when the largest mean $\mu(P_{\pi(1)})$ is tied with at least one other mean. To see this consider the case $p = 2$ and suppose $\mu(P_1) = \mu(P_2)$. Then, by the distributional assumption (43), we have

$$\begin{aligned}
P\left\{J_0^{1-\text{best}} \subseteq J_n^{\text{Gupta}}\right\} &= P\left\{\{1,2\} \subseteq J_n^{\text{Gupta}}\right\} \\
&= P\left\{\max_{j \in J} \max_{k \in J}\{\bar{X}_{n,k} - \bar{X}_{n,j}\} \leq d\frac{\sigma(P)}{\sqrt{n}}\right\} \\
&= P\left\{\max_{k \in J} \frac{\bar{X}_{n,k} - \mu_k(P)}{\sigma(P)/\sqrt{n}} - \min_{j \in J} \frac{\bar{X}_{n,j} - \mu_j(P)}{\sigma(P)/\sqrt{n}} \leq d\right\} \\
&= 2\int \left[\Phi(u+d) - \Phi(u)\right] \phi(u)du
\end{aligned}$$

The last equality uses the expression of the distribution of the range statistic for two i.i.d. standard normal random variables. Since, for all $d > 0$,

$$2\int \left[\Phi(u+d) - \Phi(u)\right] \phi(u)du = \int \Phi(u+d)\phi(u)du + \left[\int \Phi(u+d)\phi(u)du - 1\right] < \int \Phi\left(u+d\right) \phi(u)du$$

and since $\alpha > 1/2$ implies that d solving (44) for $p = 2$ must be positive, we have

$$P\left\{J_0^{1-\text{best}} \subseteq J_n^{\text{Gupta}}\right\} < 1-\alpha .$$

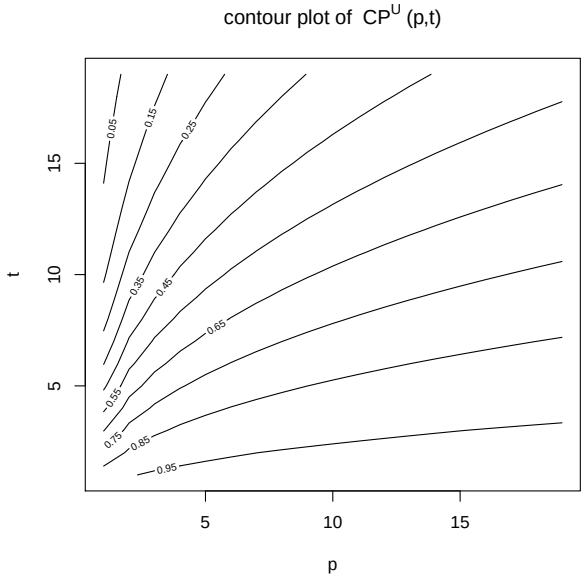

Figure 18: Contour plot of $CP^U(p,t)$ with $1-\alpha = 0.95$.

Consider now the case when $p > 2$ and t of the means are tied as the best ($2 \leq t \leq p$), say

$$\mu(P_1) = \ldots = \mu(P_t) > \mu(P_{t+1}) \geq \ldots \geq \mu(P_p) \ .$$

Then

$$\begin{aligned} P\left\{J_0^{1-\text{best}} \subseteq J_n^{\text{Gupta}}\right\} &= P\left\{\{1,\ldots,t\} \subseteq J_n^{\text{Gupta}}\right\} \\ &= P\left\{\max_{j\in\{1,\ldots,t\}} \max_{k\in J}\{\bar{X}_{n,k} - \bar{X}_{n,j}\} \leq d\frac{\sigma(P)}{\sqrt{n}}\right\} \\ &\leq P\left\{\max_{k\in\{1,\ldots,t\}} \frac{\bar{X}_{n,k} - \mu(P_k)}{\sigma(P)/\sqrt{n}} - \min_{j\in\{1,\ldots,t\}} \frac{\bar{X}_{n,j} - \mu(P_j)}{\sigma(P)/\sqrt{n}} \leq d\right\} \\ &= \underbrace{t\int \left[\Phi(u+d) - \Phi(u)\right]^{t-1}\phi(u)du}_{\equiv \widetilde{CP}^U(d,t)} \end{aligned}$$

Denote by $d(p)$ the solution to (44), i.e. Gupta's choice of critical value for a given number of populations p. Figure 18 plots the contours of $CP^U(p,t) \equiv \widetilde{CP}^U(d(p),t)$ (with $1-\alpha = 0.95$) as a function of the number of populations p and the number of ties at the largest mean t. This is an upper bound on the probability with which Gupta's confidence set covers the set of 1-best populations, $J_0^{1-\text{best}}$. Only in the lower right corner of the plot, i.e. for large p and small t, is the upper bound of the coverage probability larger than the desired level 0.95, otherwise it is strictly smaller.

Therefore, for most (t,p) combinations, Gupta's confidence set does not cover $J_0^{1-\text{best}}$ with the desired probability whereas our proposals in Section 3.4 asymptotically cover $J_0^{\tau-\text{best}}$ with probability no less than $1-\alpha$ for $\tau = 1$, but also for any other $\tau > 1$.

Appendix D Rectangular Versus Non-Rectangular Confidence Sets

In this section, we compare the size of two different simultaneous confidence sets $C_n(1-\alpha, S)$ for the differences: rectangular ones and ellipses. Suppose we want to construct a two-sided confidence set for the rank of the first population, i.e. we require a simultaneous confidence set $C_n(1-\alpha, S)$ with $S = \{(j,k)\colon j = 1 \text{ and } k \neq 1\}$.

Suppose for simplicity that $\hat{\theta} \sim N(\theta(P), I)$ and let $Z \equiv (Z_1, \ldots, Z_p) \sim N(0, I)$ denote a random vector that is independent of the data. Then

$$C_{\text{rect},n}(1-\alpha, S) \equiv \prod_{k\neq 1}\Big[\hat{\theta}_1 - \hat{\theta}_k \pm c_{\text{rect},n}(1-\alpha)\Big],$$

where $c_{\text{rect},n}(1-\alpha)$ is the $(1-\alpha)$-quantile of

$$\max_{k\neq 1} |Z_1 - Z_k|,$$

is a rectangular simultaneous confidence set for the differences. Another possible confidence set for the differences is the ellipse

$$\widetilde{C}_{\text{ellipse},n}(1-\alpha, S) \equiv \left\{\Delta_S = (\Delta_{1,k})_{k\neq 1}\colon \sum_{k\neq 1}(\hat{\theta}_1 - \hat{\theta}_k - \Delta_{1,k})^2 \leq c^2_{\text{ellipse},n}(1-\alpha)\right\},$$

where $c_{\text{ellipse},n}(1-\alpha)$ is the $(1-\alpha)$-quantile of

$$\sqrt{\sum_{k\neq 1}(Z_1 - Z_k)^2}.$$

To use this non-rectangular confidence set for the construction of a confidence set for the rank, one needs to check whether components of the vector of differences are significantly above or below zero. To this end, we need to project the confidence ellipse onto the axes, leading to the smallest rectangle in $\mathbb{R}^{p-1}$ that contains the confidence ellipse:

$$C_{\text{ellipse},n}(1-\alpha, S) \equiv \prod_{k\neq 1}\Big[\hat{\theta}_1 - \hat{\theta}_k \pm c_{\text{ellipse},n}(1-\alpha)\Big].$$

Since $\max_{k\neq 1}|Z_1 - Z_k| \leq \sqrt{\sum_{k\neq 1}(Z_1 - Z_k)^2}$, it follows that $c_{\text{rect},n}(1-\alpha) \leq c_{\text{ellipse},n}(1-\alpha)$ and, thus, that $C_{\text{ellipse},n}(1-\alpha, S)$ must be at least as large as $C_{\text{rect},n}(1-\alpha, S)$. To analyze the magnitude of the difference in sizes, we simulate the quantiles $c_{\text{rect},n}(1-\alpha)$ and $c_{\text{ellipse},n}(1-\alpha)$ with $1,000$ draws of Z and set $\alpha = 0.05$. Figure 19 shows the two quantiles for various values of p.

As expected the quantile for the rectangular confidence set for the differences is smaller than that for the non-rectangular one for all p. For $p = 3$, the two quantiles are close, but the difference between them grows with the dimension p.

Appendix E Proofs

E.1 Proofs of Results in the Main Text

Proof of Theorem 3.1. Suppose the event $\Delta_{S_j}(P) \in C_n(1-\alpha, S_j)$ holds. Then, any $k \neq j$ such that $C_n(1-\alpha, S_j, (j,k)) \subseteq \mathbf{R}_-$ satisfies $\theta(P_j) < \theta(P_k)$. Therefore, the rank $r_j(P)$ is strictly larger than the number of $k \neq j$ for which $C_n(1-\alpha, S_j, (j,k)) \subseteq \mathbf{R}_-$. Similarly, any $k \neq j$ such that $C_n(1-\alpha, S_j, (j,k)) \subseteq \mathbf{R}_+$ satisfies $\theta(P_j) > \theta(P_k)$. Therefore, the rank $r_j(P)$ is bounded above by the number of elements in J minus the number of $k \neq j$ for which $C_n(1-\alpha, S_j, (j,k)) \subseteq \mathbf{R}_+$. This establishes the first inequality of the theorem and the coverage statement follows immediately. ■

Proof of Theorem 3.2. Suppose $S_j^+(P) \cap \text{Rej}_j^- = \emptyset$ and $S_j^-(P) \cap \text{Rej}_j^+ = \emptyset$. Then, $\theta(P_j) < \theta(P_k)$ for $(j,k) \in \text{Rej}_j^-$ and $\theta(P_j) > \theta(P_k)$ for $(j,k) \in \text{Rej}_j^+$, so the bounds on the rank follow just as in the proof of Theorem 3.1. This establishes the first inequality of the theorem and the coverage statement follows immediately. ■

Proof of Theorem 3.3. Analogous to the proof of Theorem 3.1. ■

Proof of Theorem 3.4. Analogous to the proof of Theorem 3.2. ■

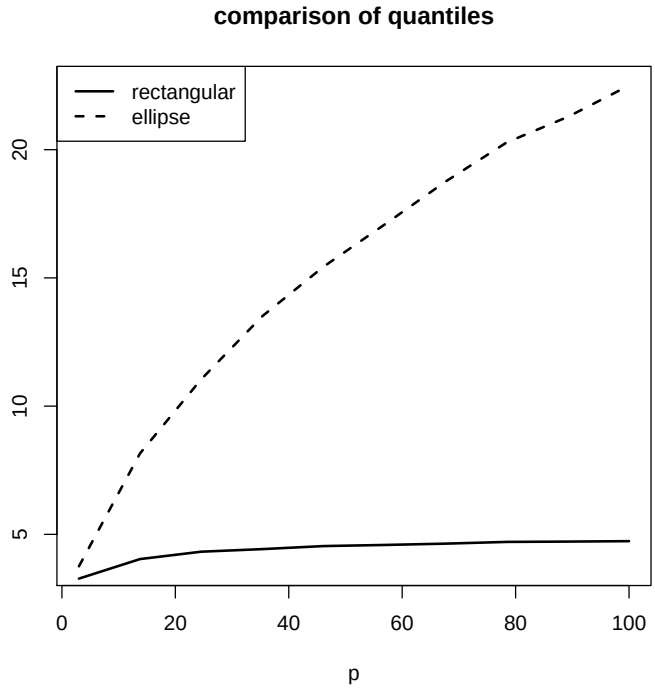

Figure 19: Graph of $c_{\mathrm{rect},n}(1-\alpha)$ ("rectangular") and $c_{\mathrm{ellipse},n}(1-\alpha)$ ("ellipse") for $\alpha = 0.05$ and various values of the dimension p.

Proof of Theorem 3.5. Let $\Pi \equiv \{\pi \text{ permutation of } J\colon \theta(P_{\pi(1)}) \geq \ldots \geq \theta(P_{\pi(p)})\}$ be the set of permutations of J that preserve the ranking of the elements of $\theta(P)$. Notice that $r_j(P) = \min_{\pi\in\Pi} \pi^{-1}(j)$ and denote by $\pi_j^* \in \Pi$ a permutation that achieves the minimum, i.e. $(\pi_j^*)^{-1}(j) = \min_{\pi\in\Pi} \pi^{-1}(j)$. The permutation π_j^* may vary with j, but two different permutations π_j^* and π_k^* can differ only on elements for which the corresponding elements in $\theta(P)$ are equal (i.e. $\theta(P_{\pi_j^*(t)}) = \theta(P_{\pi_k^*(t)})$ for all $j,k,t \in J$). Pick an arbitrary $\pi^* \in \{\pi_1^*, \ldots, \pi_p^*\}$ and define

$$J^* \equiv \{\pi^*(1), \ldots, \pi^*(\tau-1)\}\,.$$

Then:

$$\begin{aligned}
H_j \quad &\Leftrightarrow \quad r_j(P) \leq \tau \\
&\Leftrightarrow \quad (\pi_j^*)^{-1}(j) \leq \tau \\
&\Leftrightarrow \quad \theta(P_j) \geq \theta(P_{\pi_j^*(\tau)}) \\
&\Leftrightarrow \quad \theta(P_j) \geq \theta(P_{\pi^*(\tau)}) \\
&\Leftrightarrow \quad \theta(P_j) \geq \theta(P_k) \quad \forall k \in J \setminus J^* \\
&\Leftrightarrow \quad \max_{k\in J\setminus J^*} \{\theta(P_k) - \theta(P_j)\} \leq 0
\end{aligned} \tag{45}$$

The statement of the theorem obviously holds when all hypotheses are false. Therefore, assume that at least one of the hypotheses is true. Let $\hat{s}$ be the smallest integer such that there is a false rejection at Step $\hat{s}$, i.e. there is a $\hat{j} \in I_{\hat{s}} \cap J_0^{\tau-\mathrm{best}}(P)$ such that $T_{n,\hat{j}} > \hat{c}_n(1-\alpha, I_{\hat{s}})$. By definition, $J_0^{\tau-\mathrm{best}}(P) \subseteq I_{\hat{s}}$ and therefore $\hat{c}_n(1-\alpha, J_0^{\tau-\mathrm{best}}(P)) \leq \hat{c}_n(1-\alpha, I_{\hat{s}})$. Thus,

$$\max_{j\in J_0^{\tau-\mathrm{best}}(P)} T_{n,j} \geq T_{n,\hat{j}} > \hat{c}_n(1-\alpha, J_0^{\tau-\mathrm{best}}(P))$$

and

$$\begin{aligned}\text{FWER}_P &\equiv P\left\{\text{reject at least one } H_j, j \in J_0^{\tau-\text{best}}(P)\right\} \\ &\leq P\left\{\max_{j\in J_0^{\tau-\text{best}}(P)} T_{n,j} > \hat{c}_n(1-\alpha, J_0^{\tau-\text{best}}(P))\right\}. \end{aligned} \tag{46}$$

To compute this probability consider:

$$\begin{aligned} P\left\{\max_{j\in J_0^{\tau-\text{best}}(P)} T_{n,j} \leq x\right\} &= P\left\{\max_{j\in J_0^{\tau-\text{best}}(P)} \min_{K\in\mathcal{K}} \max_{k\in J\setminus K}\left\{\hat{\theta}_k - \hat{\theta}_j\right\} \leq x\right\} \\ &\geq P\left\{\max_{j\in J_0^{\tau-\text{best}}(P)} \max_{k\in J\setminus J^*}\left\{\hat{\theta}_k - \hat{\theta}_j\right\} \leq x\right\} \\ &\geq P\left\{\max_{j\in J_0^{\tau-\text{best}}(P)} \max_{k\in J\setminus J^*}\left\{\hat{\theta}_k - \hat{\theta}_j - \Delta_{k,j}(P)\right\} \leq x\right\} \\ &\geq \min_{K\in\mathcal{K}} P\left\{\max_{j\in J_0^{\tau-\text{best}}(P)} \max_{k\in J\setminus K}\left\{\hat{\theta}_k - \hat{\theta}_j - \Delta_{k,j}(P)\right\} \leq x\right\} \\ &= \min_{K\in\mathcal{K}} P\left\{T_{n,J_0^{\tau-\text{best}}(P),K} \leq x\right\} \end{aligned} \tag{47}$$

where the second inequality follows from (45).

Then there exists a set $K^* = K_n^*(P) \in \mathcal{K}$ such that, by combining (46), (47), and the definition of $\hat{c}_n(1-\alpha, J_0^{\tau-\text{best}}(P))$, we have

$$\begin{aligned}\text{FWER}_P &\leq 1 - \min_{K\in\mathcal{K}} P\left\{T_{n,J_0^{\tau-\text{best}}(P),K} \leq \hat{c}_n(1-\alpha, J_0^{\tau-\text{best}}(P))\right\} \\ &= 1 - P\left\{T_{n,J_0^{\tau-\text{best}}(P),K^*} \leq \max_{K\in\mathcal{K}} M_n^{-1}(1-\alpha, J_0^{\tau-\text{best}}(P), K, \hat{P}_n)\right\} \\ &\leq 1 - P\left\{T_{n,J_0^{\tau-\text{best}}(P),K^*} \leq M_n^{-1}(1-\alpha, J_0^{\tau-\text{best}}(P), K^*, \hat{P}_n)\right\}. \end{aligned}$$

Therefore,

$$\begin{aligned}\limsup_{n\to\infty} \sup_{P\in\mathbf{P}} \text{FWER}_P &\leq 1 - \liminf_{n\to\infty} \inf_{P\in\mathbf{P}} P\left\{T_{n,J_0^{\tau-\text{best}}(P),K^*} \leq M_n^{-1}(1-\alpha, J_0^{\tau-\text{best}}(P), K^*, \hat{P}_n)\right\} \\ &\leq 1 - \liminf_{n\to\infty} \inf_{P\in\mathbf{P}} \min_{K\in\mathcal{K}} P\left\{T_{n,J_0^{\tau-\text{best}}(P),K} \leq M_n^{-1}(1-\alpha, J_0^{\tau-\text{best}}(P), K, \hat{P}_n)\right\} \\ &\leq \alpha\ , \end{aligned}$$

where the last inequality follows from (30) and the fact that $\mathcal{K}$ is a finite set. The desired result now follows because

$$\liminf_{n\to\infty} \inf_{P\in\mathbf{P}} P\{J_0^{\tau-\text{best}}(P) \subseteq J_n^{\tau-\text{best}}\} = 1 - \limsup_{n\to\infty} \sup_{P\in\mathbf{P}} \text{FWER}_P \geq 1-\alpha.$$

■

E.2 Proofs of Results in the Appendix

Proof of Lemma B.1. Since the quantile in (40) is smaller than that in (41), we show the results for (40) and the analogous results for (41) then follow immediately. Define $\tau_j \equiv \sigma(P_j)/\sqrt{n_j}$.

Consider the claim (i). Suppose $\Delta_{j,k}(P) \in \widetilde{C}_n(1-\alpha,(j,k)) \equiv [\hat{\theta}_j - \hat{\theta}_k \pm (\tau_j + \tau_k)\widetilde{q}_{1-\alpha}]$ for all $(j,k) \in S_{\text{all}}$. Then, for all $k \in \widetilde{N}_j^-$, the interval $\widetilde{C}_n(1-\alpha,(j,k))$ lies entirely below zero, so that $\theta(P_j) < \theta(P_k)$. Similarly, for all $k \in \widetilde{N}_j^+$, we have $\theta(P_j) > \theta(P_k)$. Therefore, the smallest possible value of the rank of j cannot be smaller than the number of elements in $\widetilde{N}_j^-$, i.e. $\underline{r}_j(P) \geq |\widetilde{N}_j^-|$, and the largest value of the rank of j cannot be larger than p minus the number of elements in $\widetilde{N}_j^+$, i.e. $\bar{r}_j(P) \leq p - |\widetilde{N}_j^+|$. Therefore,

$$P\left\{R(P) \subseteq \widetilde{R}_n^{\text{joint}}\right\} \geq P\left\{\Delta_{j,k}(P) \in \widetilde{C}_n(1-\alpha,(j,k)) \text{ for all } (j,k) \in S_{\text{all}}\right\}.$$

Letting $\alpha_{j,k} \equiv \tau_j/(\tau_j+\tau_k)$, we have

$$\begin{aligned}
&P\left\{\Delta_{j,k}(P) \in \widetilde{C}_n(1-\alpha,(j,k)) \text{ for all } (j,k) \in S_{\text{all}}\right\} \\
&= P\left\{\theta(P_j)-\theta(P_k) \in \left[\hat{\theta}_j-\hat{\theta}_k \pm (\tau_j+\tau_k)\widetilde{q}_{1-\alpha}\right] \text{ for all } (j,k) \in S_{\text{all}}\right\} \\
&= P\left\{\max_{(j,k)\in S_{\text{all}}} \left|\frac{\hat{\theta}_j-\theta(P_j)}{\tau_j+\tau_k} - \frac{\hat{\theta}_k-\theta(P_k)}{\tau_j+\tau_k}\right| \le \widetilde{q}_{1-\alpha}\right\} \\
&= P\left\{\max_{(j,k)\in S_{\text{all}}} \left|\alpha_{j,k}\frac{\hat{\theta}_j-\theta(P_j)}{\tau_j} - (1-\alpha_{j,k})\frac{\hat{\theta}_k-\theta(P_k)}{\tau_k}\right| \le \widetilde{q}_{1-\alpha}\right\} \\
&> P\left\{\max_{(j,k)\in S_{\text{all}}} \max\left\{\left|\frac{\hat{\theta}_j-\theta(P_j)}{\tau_j}\right|, \left|\frac{\hat{\theta}_k-\theta(P_k)}{\tau_k}\right|\right\} \le \widetilde{q}_{1-\alpha}\right\} \\
&= P\left\{\max_{j\in J}\left|\frac{\hat{\theta}_j-\theta(P_j)}{\tau_j}\right| \le \widetilde{q}_{1-\alpha}\right\} \\
&= 1-\alpha \ ,
\end{aligned}$$

The strict inequality follows from the fact that, for any numbers $\omega \in (0,1)$ and $A \neq -B$, we have $|\omega A-(1-\omega)B| < \max\{|A|,|B|\}$. This inequality is applicable above because $\alpha_{j,k} \in (0,1)$ and $P\{(\hat{\theta}_j-\theta(P_j))/\tau_j = (\hat{\theta}_k-\theta(P_k))/\tau_k\} = 0$ for all $(j,k) \in S_{\text{all}}$. Therefore, the desired claim in (i) follows.

Part (ii) follows from analogous arguments to those in the proof of Theorem 3.3 and in Remark 3.6.

Consider part (iii). Notice that, for $p=2$, $q_{1-\alpha}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution. Suppose there is a $k \in J$ such that $\widetilde{C}_n(1-\alpha,k)$ lies entirely above $\widetilde{C}_n(1-\alpha,j)$, i.e.,

$$\hat{\theta}_k - \tau_k\widetilde{q}_{1-\alpha} > \hat{\theta}_j + \tau_j\widetilde{q}_{1-\alpha}$$

or, equivalently,

$$\hat{\theta}_j - \hat{\theta}_k + (\tau_j+\tau_k)\widetilde{q}_{1-\alpha} < 0 \ .$$

This implies that $C_n(1-\alpha,S_{\text{all}},(j,k))$ lies entirely below zero if $\tau_j+\tau_k > \sqrt{\tau_j^2+\tau_k^2}$ and $\widetilde{q}_{1-\alpha} > q_{1-\alpha}$. The former condition obviously holds because $\tau_j > 0$ for all $j \in J$. The latter follows because, since $\alpha \in (0,1)$,

$$\frac{1+\sqrt{1-\alpha}}{2} > 1-\frac{\alpha}{2} \ .$$

Therefore, we have shown that $|\widetilde{N}_j^-| \le |N_j^-|$ for all $j \in J$. Similarly, we can show that $|\widetilde{N}_j^+| \le |N_j^+|$ for all $j \in J$. For $|\widetilde{N}_j^-| < |N_j^-|$ to occur for some $j \in J$, there must exist a $k \in J$ such that

$$\hat{\theta}_j - \hat{\theta}_k + \sqrt{\tau_j^2+\tau_k^2}q_{1-\alpha} < 0 < \hat{\theta}_j - \hat{\theta}_k + (\tau_j+\tau_k)\widetilde{q}_{1-\alpha}$$

or, equivalently,

$$\hat{\theta}_k - \hat{\theta}_j \in \left(\sqrt{\tau_j^2+\tau_k^2}q_{1-\alpha}, (\tau_j+\tau_k)\widetilde{q}_{1-\alpha}\right) \ .$$

This event occurs with positive probability because the interval has positive length, as shown above, and the difference in the estimators is normally distributed. Similarly, we can show that $|\widetilde{N}_j^+| < |N_j^+|$ for some $j \in J$ occurs with positive probability, so the desired claim follows.

Finally, consider part (iv). Suppose there is a $k \in J$ such that $\widetilde{C}_n(1-\alpha,k)$ lies entirely above $\widetilde{C}_n(1-\alpha,j)$, i.e.,

$$\hat{\theta}_j - \hat{\theta}_k + \frac{2\sigma(P)}{\sqrt{n}}\widetilde{q}_{1-\alpha} < 0 \ .$$

Notice that $q_{1-\alpha}$ is the $(1-\alpha)$-quantile from the distribution of $\max_{(j,k)\in S_{\text{all}}} \frac{1}{\sqrt{2}}|Z_j-Z_k|$, where $Z_1,\ldots,Z_p$ are i.i.d. $N(0,1)$ random variables. This quantile satisfies $\sqrt{2}q_{1-\alpha} < 2\widetilde{q}_{1-\alpha}$ for all $\alpha \in (0,1)$ so that

$$\hat{\theta}_j - \hat{\theta}_k + \frac{\sqrt{2}\sigma(P)}{\sqrt{n}}q_{1-\alpha} < \hat{\theta}_j - \hat{\theta}_k + \frac{2\sigma(P)}{\sqrt{n}}\widetilde{q}_{1-\alpha} < 0 \ ,$$

which means that $C_n(1-\alpha,S_{\text{all}},(j,k))$ lies entirely below zero. Therefore, $|\widetilde{N}_j^-| \le |N_j^-|$ for all $j \in J$. Similarly, we can show that $|\widetilde{N}_j^+| \le |N_j^+|$ for all $j \in J$. The remainder of the proof is then similar to that of part (iii). ■

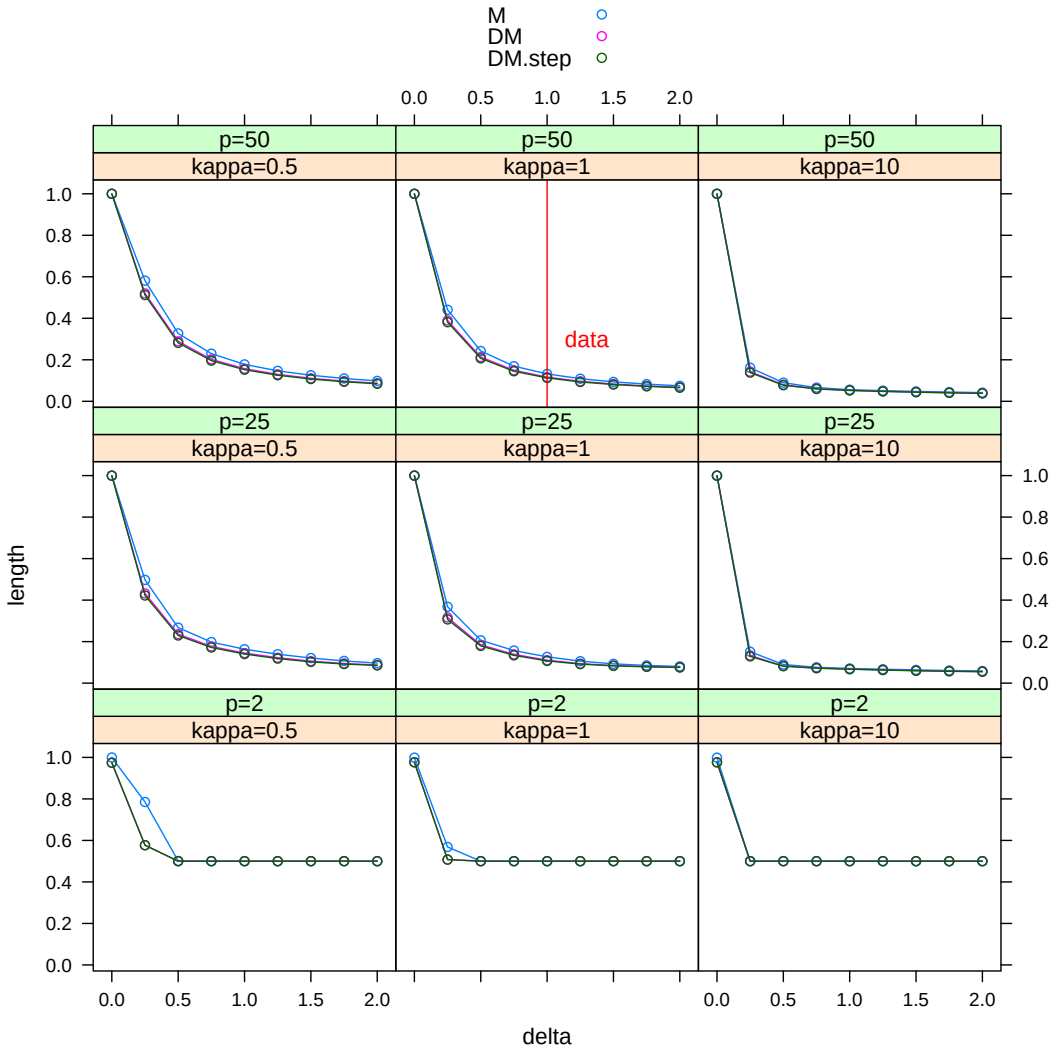

Figure 20: Joint confidence sets for the ranks of all CZs: relative length for the correlational design

Appendix F Additional Simulation Results

In this section, we present the simulation results for simultaneous inference on the ranks of all populations and on the τ-best populations with $\tau = 2$. We also provide additional simulations comparing the single-step and stepwise methods.

F.1 Joint Confidence Sets for the Ranks of All Populations

The simulation design is the same as described in Section 4. The methods analyzed are also the same as described in Section 4, except that "DM" is now based on (23) rather than (13) and "DM.step" is now based on (25) computed through Algorithm 3.2 rather than (19) computed through Algorithm 3.1.

Table 6 and Figures 20–21 are the counterparts of Table 2 and Figures 6–7. The plotted relative length of the simultaneous confidence sets is computed as the average length across CZs, averaged across Monte

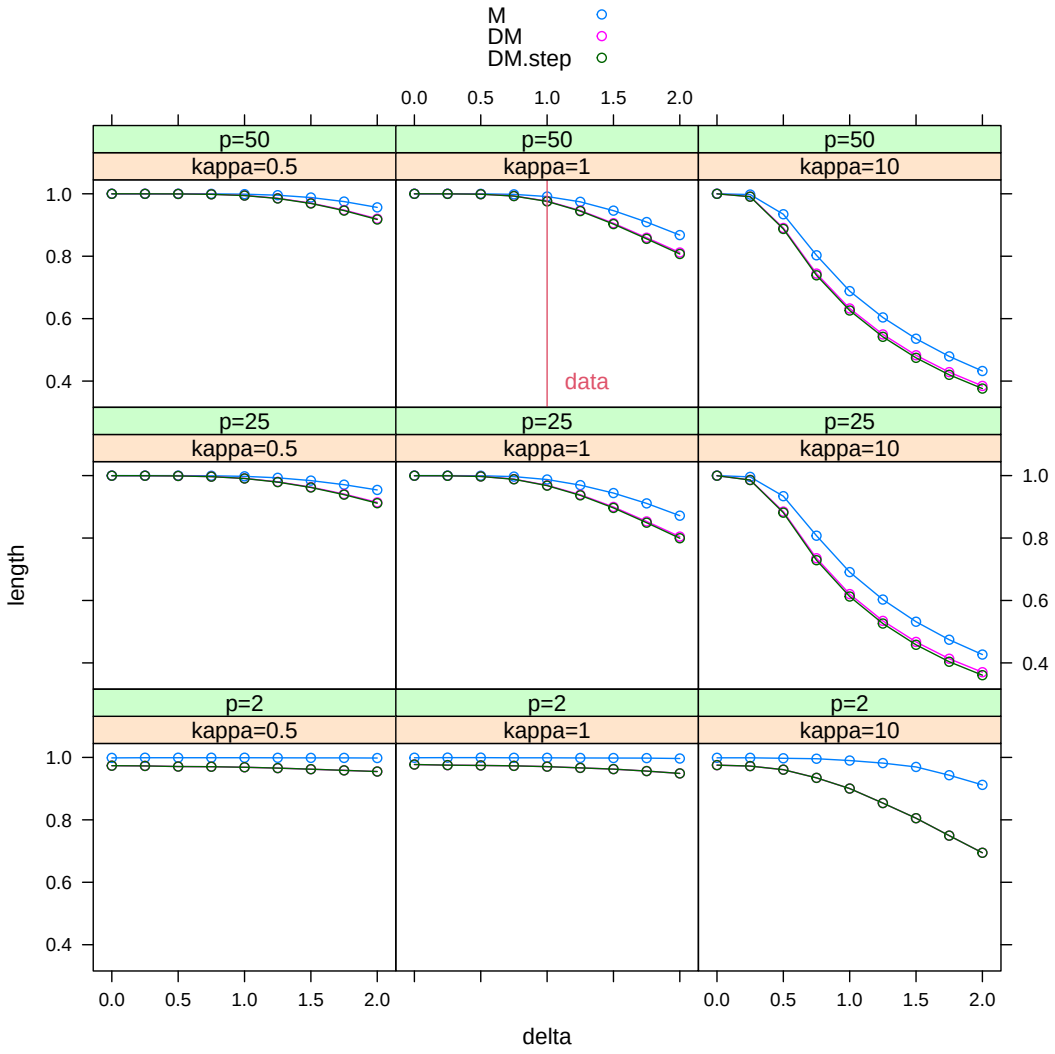

Figure 21: Joint confidence sets for the ranks of all CZs: relative length for the movers design

Carlo samples, and divided by p. The results are qualitatively the same as those for the rank of a single population. One additional insight is that the length of the simultaneous confidence sets for the ranks of all populations shrinks more slowly with the sample size and the degree of separation between the mobility measures, δ.

F.2 Confidence Sets for the τ-Best Populations

The simulation design is the same as described in Section 4 with one modification: instead of (31), θ_j is now defined so that the distance between the two top-ranked CZs and the remaining ones can be varied by the

parameter δ. Specifically,

$$\theta_j \equiv \begin{cases} \hat{\theta}_{t,j}, & j = 1,2 \\ \hat{\theta}_{t,3} + (1-\delta)\left(\hat{\theta}_{t,2} - \hat{\theta}_{t,3}\right), & j = 3 \\ \hat{\theta}_{t,3} + \delta \sum_{k=3}^{j-1}(\hat{\theta}_{t,k+1} - \hat{\theta}_{t,k}) + (1-\delta)\left(\hat{\theta}_{t,2} - \hat{\theta}_{t,3}\right), & j > 3 \end{cases}.$$

When $\delta = 1$, then as in the main text θ_j is equal to the mobility estimate from the data. When $\delta = 0$, then $\theta_2 = \ldots = \theta_p$, which means the two top-ranked CZs are not separated from the remaining CZs and the set of τ-best CZs contains all CZs. Values of δ between 0 and 1 correspond to cases in which the top-two ranked CZs are distinct from the remaining CZs, but not as well separated as in the data. Similarly, a value of δ larger than 1 corresponds to a case in which the separation is larger than in the data.

We compare the following methods:

"DM": the projection confidence set in (27), based on confidence sets for the differences in means as in (10), where $L^{-1}_{\text{upper,n}}(1-\alpha, S, \hat{P}_n)$ is the empirical $(1-\alpha)$-quantile of the 1,000 draws of $\max_{(j,k)\in S_{\text{all}}}(Z_k - Z_j)/\hat{\sigma}_{j,k}$.

"DM.step": the projection confidence set "DM" except that stepwise improvements as in Remark 3.9 are applied.

"T": the test inversion procedure from Section 3.4, using the test statistic in (28), where the critical value $\hat{c}_n(1-\alpha, I)$ is the maximum (over $K \in \mathcal{K}$) of the empirical $(1-\alpha)$-quantile of the 1,000 draws of $\max_{j\in J} \max_{k\in J\setminus K}\{Z_k - Z_j\}$.

"M": the projection confidence set in (27), based on symmetric confidence sets for the differences in means as in "M" above.

Table 7 and Figures 22–23 are the counterparts of Table 2 and Figures 6–7. First, Table 7 shows that all methods control the coverage frequency at the desired nominal level for small and large sample sizes, regardless of whether the mobility measures of the top-two CZs are well-separated from the remaining ones, and regardless of whether there are few or many populations to be ranked.

For both the correlational and the movers design, all methods' coverage frequencies are close to one for most parameter combinations except when $\delta = 0$. In the correlational design, coverage is actually close to the nominal level when $\delta = 0$ and the sample size is large. However, unlike for inference on ranks, coverage does not necessarily increase monotonically as the top-two become more separated (δ grows). Simulations further below show that the coverage frequency may be closest to the nominal level at a strictly positive value of δ.

Second, the coverage frequency of "M" is approximately equal to one in all scenarios whereas our methods "T", "DM" and "DM.step" tend to have coverage frequency closer to the nominal level at least when $\delta = 0$. In consequence, our methods, in particular the stepwise method "DM.step", tend to lead to confidence sets for the 2-best populations that are not larger than those of "M" and smaller in some scenarios. While the differences are small in the correlational design, the methods "DM" and "DM.step" may lead to significantly shorter confidence sets in the movers design. The direct "T" method dominates "M" in most scenarios, but may also perform worse than "M" in terms of size of the confidence set when the sample size is small relative

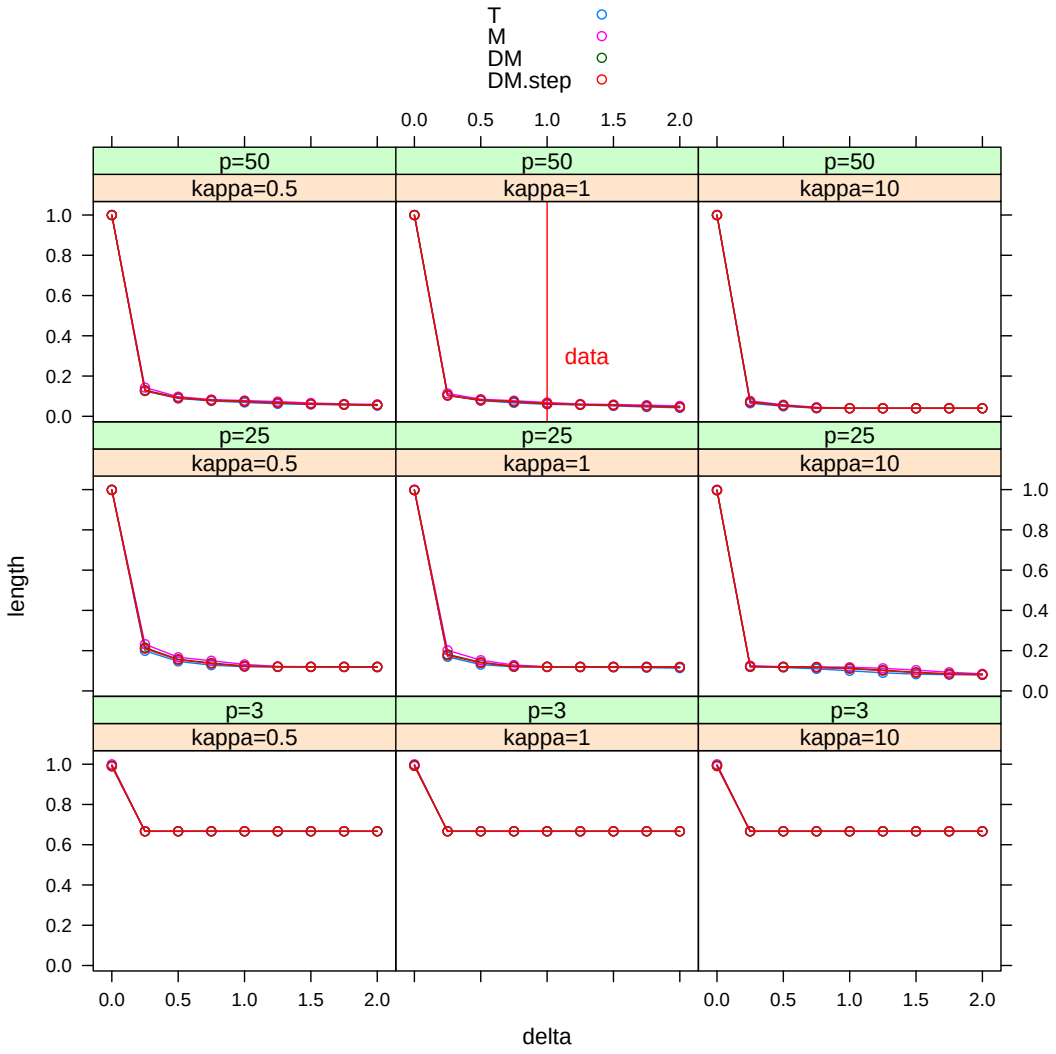

Figure 22: Confidence sets for the τ-best CZs: relative length for the correlational design

to the number of CZs (see the top row of Figure 23). In an alternative simulation design further below, we show that the differences between our proposals "DM"/"DM.step" and "M" can be much more substantial. In addition, "T" may significantly outperform all other methods.

Third, as the two top-ranked CZs are more separated from the remaining ones (δ increases), the size of the confidence sets decreases. Comparing the bottom rows of Figures 22 and 23, we see that, in the correlational design, the top-two CZs can be distinguished from the third-ranked CZ even for small values of δ whereas, in the movers design, all three CZs are in the confidence set for the 2-best regardless of the value of δ.

Fourth, comparing the left and the right columns in Figures 22 and 23 shows that an increase in sample size leads to smaller confidence sets, as expected.

To highlight some differences in performance across the four methods that could not be seen in the simulation design calibrated to the data, we now present additional simulation results from a different

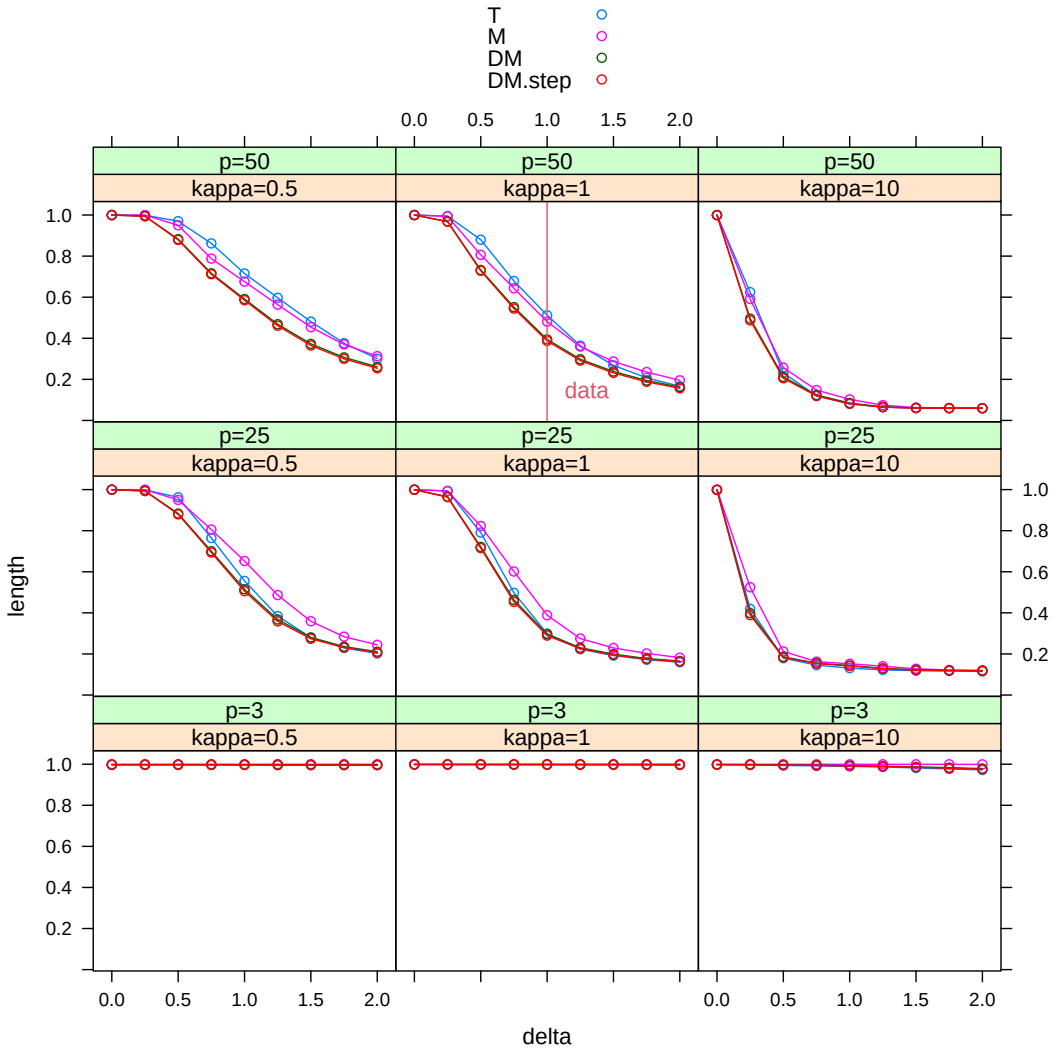

Figure 23: Confidence sets for the τ-best CZs: relative length for the movers design

design. For each population $j = 1, \ldots, p$, we generate an i.i.d. sample $X_{j,1}, \ldots, X_{j,n}$ from $N(\theta_j, 1)$ so that all samples across populations are mutually independent. The parameter $\theta \equiv (\theta_1, \ldots, \theta_p)'$ is defined as follows:

$$\theta_1 = \delta_1$$
$$\theta_2 = \theta_3 = \frac{\delta_1}{\delta_2}$$
$$\theta_4 = \ldots = \theta_p = 0$$

In all simulation scenarios, the first three populations, $j = 1, 2, 3$, possess ranks less than or equal to 2 and are thus elements of the set of 2-best populations. All other elements of θ are equal to zero. Suppose $\delta_2 = 1$. Then, the magnitude of δ_1 determines how well the top three populations are separated from the remaining ones. A value of $\delta_2 \geq 1$ allows us to consider data-generating processes in which the populations at rank equal to two (i.e. θ_2, θ_3) are separated from that of rank equal to one (i.e. θ_1). When $\delta_1 = 0$,

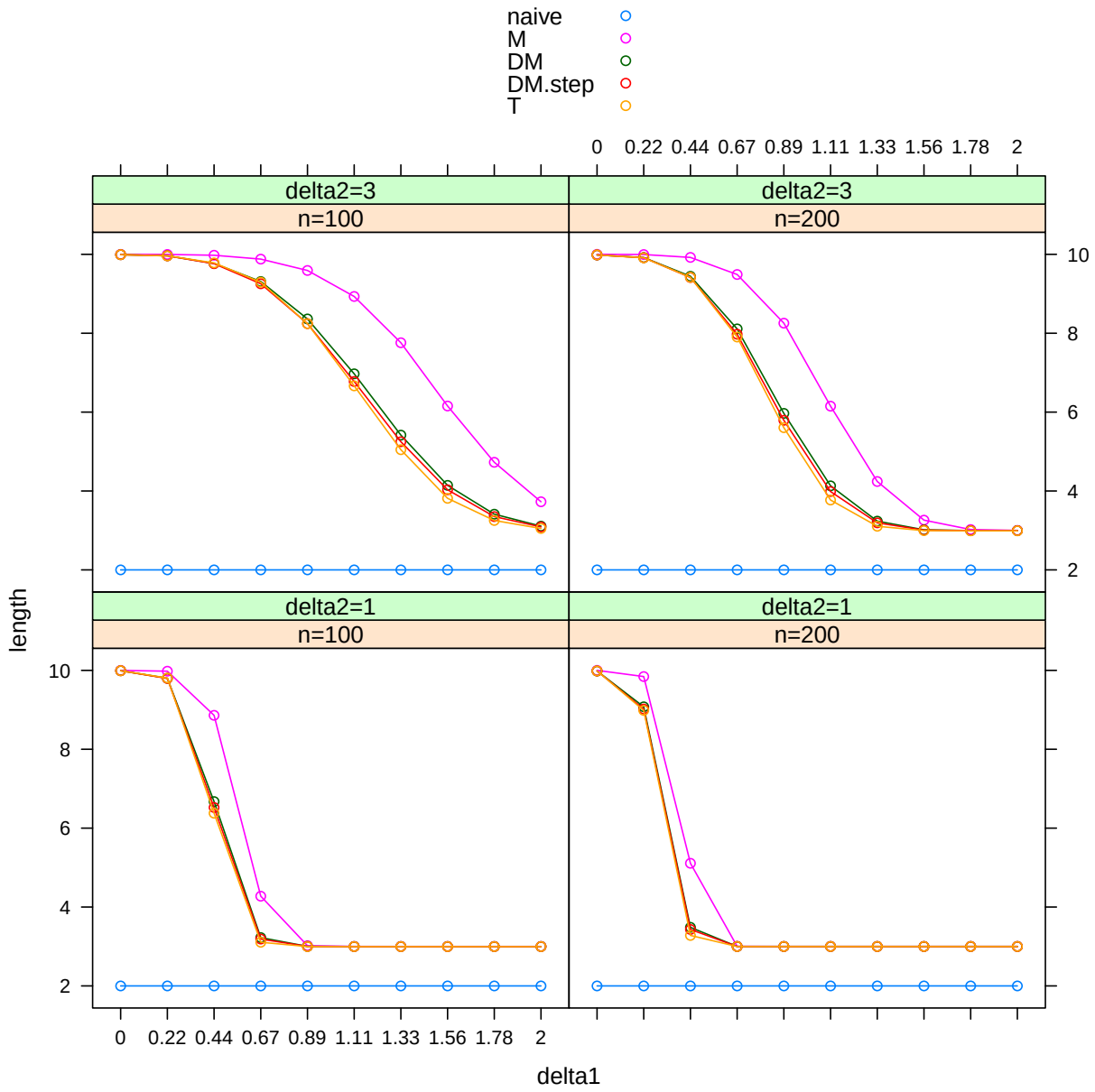

Figure 24: Confidence sets for the τ-best: length of the confidence sets for $p = 10$.

then all populations are tied with all elements of θ equal to 0. All simulations are based on $1,000$ Monte Carlo samples and nominal coverage of 95%. The methods we compare are the same as described above, but in addition we also consider the "naive" procedure that simply picks the two populations with the largest estimated means.

Table 8 shows the coverage frequencies, where coverage is computed as in the other simulations, i.e. as in Remark 3.6. Figure 24 shows the length (not the relative length as in the tables above) of the confidence sets, averaged over the Monte Carlo samples, for $p = 10$. Several aspects of the simulation results are similar to those above, so we focus the discussion on differences. First, the "naive" method does not control the coverage frequency at the desired nominal level. It never covers the set of τ-best populations because,

by definition it only selects two populations even though, in all scenarios, there are at least three τ-best populations.

Second, the coverage frequency of "M" is approximately equal to one in all scenarios whereas our methods "T", "DM" and "DM.step" tend to have coverage frequency closer to the nominal level. In consequence, our methods tend to lead to confidence sets for the 2-best populations that are not larger than those of "M" and substantially smaller in most scenarios. For instance, the top row of graphs in Figure 24 shows that the confidence sets of "M" may be up to about 50% larger than those of our proposed methods. The method "T" generally produces even smaller confidence sets than the projection methods "DM" and "DM.step", but the gains are modest. Similarly, the stepwise improvements ("DM.step") help shorten the confidence sets relative to "DM", but the gains are modest. Overall, the three methods "DM", "DM.step", and "T" perform similarly well.

Third, as in the case of inference on a single rank and on all ranks, comparing the top and the bottom rows of Figure 24 shows that as the first three populations' means become better separated from the others (δ_2 decreases or δ_1 increases). The lengths of the confidence sets shrink because it is easier to recover the populations with rank less than or equal to two (exactly those first three populations).

Finally, Table 8 shows that the coverage frequency is not necessarily closest to the nominal level when all populations are tied ($\delta_1 = 0$). For instance, when $\delta_2 = 3$ and $p = 3$, then the coverage frequency of our procedures is close to one when all populations are tied ($\delta_1 = 0$), but takes values near 0.98 when δ_1 is large.

			correlational design						movers design					
κ	p	method	$\delta = 0$	$\delta = 0.25$	$\delta = 0.5$	$\delta = 0.75$	$\delta = 1$	$\delta = 2$	$\delta = 0$	$\delta = 0.25$	$\delta = 0.5$	$\delta = 0.75$	$\delta = 1$	$\delta = 2$
0.5	2	M	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.999	0.999	1.000	1.000	1.000
		DM	0.949	1.000	1.000	1.000	1.000	1.000	0.947	0.979	0.980	0.984	0.986	0.991
		DM.step	0.949	1.000	1.000	1.000	1.000	1.000	0.947	0.979	0.980	0.984	0.986	0.991
	25	M	0.997	1.000	1.000	1.000	1.000	1.000	0.995	1.000	1.000	1.000	1.000	1.000
		DM	0.950	1.000	1.000	1.000	0.999	1.000	0.947	0.997	0.999	0.999	0.999	0.999
		DM.step	0.950	1.000	1.000	0.999	0.999	0.999	0.947	0.997	0.999	0.999	0.999	0.999
	50	M	0.996	1.000	1.000	1.000	1.000	1.000	0.995	1.000	1.000	1.000	1.000	1.000
		DM	0.948	1.000	1.000	1.000	1.000	1.000	0.956	1.000	1.000	1.000	1.000	1.000
		DM.step	0.948	1.000	1.000	1.000	1.000	1.000	0.956	1.000	1.000	1.000	1.000	1.000
1	2	M	0.998	1.000	1.000	1.000	1.000	1.000	0.998	1.000	1.000	1.000	1.000	1.000
		DM	0.953	1.000	1.000	1.000	1.000	1.000	0.955	0.981	0.984	0.986	0.988	0.995
		DM.step	0.953	1.000	1.000	1.000	1.000	1.000	0.955	0.981	0.984	0.986	0.988	0.995
	25	M	0.997	1.000	1.000	1.000	1.000	1.000	0.993	1.000	1.000	1.000	1.000	1.000
		DM	0.951	1.000	1.000	1.000	1.000	1.000	0.944	0.998	0.999	1.000	1.000	1.000
		DM.step	0.951	0.999	0.999	0.999	0.999	1.000	0.944	0.998	0.999	1.000	1.000	1.000
	50	M	0.995	1.000	1.000	1.000	1.000	1.000	0.995	1.000	1.000	1.000	1.000	1.000
		DM	0.954	1.000	1.000	1.000	1.000	1.000	0.958	1.000	1.000	1.000	1.000	1.000
		DM.step	0.954	1.000	1.000	1.000	1.000	1.000	0.958	1.000	1.000	1.000	1.000	1.000
10	2	M	0.998	1.000	1.000	1.000	1.000	1.000	0.998	1.000	1.000	1.000	1.000	1.000
		DM	0.953	1.000	1.000	1.000	1.000	1.000	0.951	0.987	0.994	0.998	0.999	1.000
		DM.step	0.953	1.000	1.000	1.000	1.000	1.000	0.951	0.987	0.994	0.998	0.999	1.000
	25	M	0.997	1.000	1.000	1.000	1.000	1.000	0.996	1.000	1.000	1.000	1.000	1.000
		DM	0.949	1.000	1.000	1.000	1.000	1.000	0.947	1.000	1.000	1.000	1.000	1.000
		DM.step	0.949	1.000	1.000	1.000	1.000	1.000	0.947	1.000	1.000	1.000	1.000	1.000
	50	M	0.993	1.000	1.000	1.000	1.000	1.000	0.997	1.000	1.000	1.000	1.000	1.000
		DM	0.952	1.000	0.999	0.999	1.000	1.000	0.953	1.000	1.000	1.000	1.000	1.000
		DM.step	0.952	0.999	0.999	0.999	0.999	1.000	0.953	1.000	1.000	1.000	1.000	1.000

Table 6: Joint confidence sets for the ranks of all CZs: coverage

			correlational design						movers design					
κ	p	method	$\delta = 0$	$\delta = 0.25$	$\delta = 0.5$	$\delta = 0.75$	$\delta = 1$	$\delta = 2$	$\delta = 0$	$\delta = 0.25$	$\delta = 0.5$	$\delta = 0.75$	$\delta = 1$	$\delta = 2$
0.5	3	T	0.971	1.000	1.000	1.000	1.000	1.000	0.990	0.999	0.999	0.999	0.999	0.999
		M	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		DM	0.979	1.000	1.000	1.000	1.000	1.000	0.994	0.999	1.000	1.000	1.000	1.000
		DM.step	0.968	1.000	1.000	1.000	1.000	1.000	0.993	0.999	0.999	0.999	0.999	1.000
	25	T	0.969	1.000	1.000	1.000	1.000	1.000	0.995	1.000	1.000	1.000	1.000	1.000
		M	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		DM	0.963	1.000	1.000	1.000	1.000	1.000	0.992	1.000	1.000	1.000	1.000	1.000
		DM.step	0.958	1.000	1.000	1.000	1.000	1.000	0.993	1.000	1.000	1.000	1.000	1.000
	50	T	0.986	1.000	1.000	1.000	1.000	1.000	0.988	1.000	1.000	1.000	1.000	1.000
		M	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		DM	0.988	1.000	1.000	1.000	1.000	1.000	0.989	1.000	1.000	1.000	1.000	1.000
		DM.step	0.988	1.000	1.000	1.000	1.000	1.000	0.990	1.000	1.000	1.000	1.000	1.000
1	3	T	0.978	1.000	1.000	1.000	1.000	1.000	0.994	0.999	0.999	0.999	0.999	0.999
		M	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		DM	0.986	1.000	1.000	1.000	1.000	1.000	0.996	0.998	0.998	0.998	0.999	1.000
		DM.step	0.975	1.000	1.000	1.000	1.000	1.000	0.996	0.998	0.998	0.998	0.999	0.999
	25	T	0.957	1.000	1.000	1.000	1.000	1.000	0.992	1.000	1.000	1.000	1.000	1.000
		M	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		DM	0.954	1.000	1.000	1.000	1.000	1.000	0.985	1.000	1.000	1.000	1.000	1.000
		DM.step	0.951	1.000	1.000	1.000	1.000	1.000	0.983	1.000	1.000	1.000	1.000	1.000
	50	T	0.992	1.000	1.000	1.000	1.000	1.000	0.991	1.000	1.000	1.000	1.000	1.000
		M	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000
		DM	0.990	1.000	1.000	1.000	1.000	1.000	0.991	1.000	1.000	1.000	1.000	1.000
		DM.step	0.991	1.000	1.000	1.000	1.000	1.000	0.992	1.000	1.000	1.000	1.000	1.000
10	3	T	0.976	1.000	1.000	1.000	1.000	1.000	0.993	0.999	1.000	1.000	1.000	1.000
		M	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		DM	0.982	1.000	1.000	1.000	1.000	1.000	0.994	1.000	1.000	1.000	1.000	1.000
		DM.step	0.971	1.000	1.000	1.000	1.000	1.000	0.994	0.999	1.000	1.000	1.000	1.000
	25	T	0.948	1.000	1.000	1.000	1.000	1.000	0.973	1.000	1.000	1.000	1.000	1.000
		M	0.996	1.000	1.000	1.000	1.000	1.000	0.998	1.000	1.000	1.000	1.000	1.000
		DM	0.948	1.000	1.000	1.000	1.000	1.000	0.972	1.000	1.000	1.000	1.000	1.000
		DM.step	0.945	1.000	1.000	1.000	1.000	1.000	0.971	1.000	1.000	1.000	1.000	1.000
	50	T	0.991	1.000	1.000	1.000	1.000	1.000	0.984	1.000	1.000	1.000	1.000	1.000
		M	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000
		DM	0.989	1.000	1.000	1.000	1.000	1.000	0.976	1.000	1.000	1.000	1.000	1.000
		DM.step	0.987	1.000	1.000	1.000	1.000	1.000	0.971	1.000	1.000	1.000	1.000	1.000

Table 7: Confidence set for the τ-best CZs: coverage

				δ_1									
n	δ_2	p	test	0	0.22	0.44	0.67	0.89	1.11	1.33	1.56	1.78	2
100	1	3	naive	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			M	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			DM	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996
			DM.step	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995
			T	0.993	0.993	0.993	0.993	0.993	0.993	0.993	0.993	0.993	0.993
		10	naive	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			M	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			DM	0.992	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
			DM.step	0.991	0.999	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998
			T	0.992	0.998	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997
		50	naive	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			M	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			DM	0.985	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			DM.step	0.984	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			T	0.990	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	3	3	naive	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			M	1.000	0.999	0.999	0.998	0.998	0.998	0.998	0.998	0.998	0.998
			DM	0.996	0.992	0.982	0.979	0.979	0.979	0.979	0.979	0.979	0.979
			DM.step	0.995	0.989	0.978	0.974	0.973	0.968	0.968	0.968	0.968	0.968
			T	0.993	0.990	0.981	0.979	0.979	0.979	0.979	0.979	0.979	0.979
		10	naive	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			M	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			DM	0.992	0.998	0.998	0.997	0.997	0.997	0.997	0.997	0.997	0.997
			DM.step	0.991	0.998	0.998	0.997	0.997	0.996	0.996	0.996	0.996	0.996
			T	0.992	0.997	0.997	0.997	0.995	0.995	0.995	0.995	0.995	0.995
		50	naive	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			M	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			DM	0.985	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			DM.step	0.984	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			T	0.990	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.998	0.998
200	1	3	naive	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			M	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			DM	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995
			DM.step	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995
			T	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995
		10	naive	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			M	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			DM	0.985	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			DM.step	0.985	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			T	0.988	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
		50	naive	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			M	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			DM	0.986	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			DM.step	0.985	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			T	0.991	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	3	3	naive	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			M	1.000	1.000	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.999
			DM	0.995	0.988	0.985	0.984	0.984	0.984	0.984	0.984	0.984	0.984
			DM.step	0.995	0.987	0.983	0.980	0.980	0.980	0.980	0.980	0.980	0.980
			T	0.995	0.986	0.982	0.981	0.981	0.981	0.981	0.981	0.981	0.981
		10	naive	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			M	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			DM	0.985	0.996	0.995	0.996	0.996	0.996	0.996	0.996	0.996	0.996
			DM.step	0.985	0.996	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995
			T	0.988	0.996	0.995	0.993	0.992	0.990	0.991	0.990	0.990	0.990
		50	naive	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			M	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			DM	0.986	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			DM.step	0.985	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			T	0.991	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 8: Confidence sets for the τ-best populations: coverage

F.3 Single-Step Versus Stepwise Methods

In this section, we show through simulations that the stepwise procedure "DM.step" may provide significant improvements in length of the confidence set for the rank of a population relative to the single-step method "DM".

The simulation design is different from the one in the main text. For each population $j = 1, \ldots, p$, we generate an i.i.d. sample $X_{j,1}, \ldots, X_{j,n}$ from $N(\theta_j, 1)$ so that all samples across populations are mutually independent. The parameter $\theta \equiv (\theta_1, \ldots, \theta_p)'$ is defined as follows:

$$\theta_1 = \frac{1}{2} + \delta$$

$$\theta_2, \ldots, \theta_p \text{ lie on an equally-spaced grid from 0 to } \frac{1}{2}$$

The magnitude of δ determines how well the first population, on whose rank we perform inference, is separated from the remaining ones. All simulations are based on $1,000$ Monte Carlo samples, $n = 100$, $p = 100$, and nominal coverage of 95%. We compare the two methods "DM" and "DM.step" as described in Section 4.

Tables 9 and 10 show the coverage frequency and the length (not relative length as in some previous tables) of the marginal confidence set for the first population. In these simulations, most populations are well-separated (even when δ is small) and thus coverage is close to one. The length of the confidence sets may differ significantly between the two methods. For instance, for $\delta = 0.5$ the stepwise procedure leads to a confidence set that is $1 - 7.91/9.84 \approx 20\%$ shorter than that of the single-step method.

				δ								
n	δ_2	p	test	0	0.125	0.25	0.375	0.5	0.625	0.75	0.875	1
100	1	100	DM	1.000	1.000	1.000	1.000	0.998	0.998	0.998	0.998	0.998
			DM.step	1.000	1.000	1.000	0.999	0.998	0.998	0.998	0.998	0.998

Table 9: Marginal confidence set for the rank of the most populous CZ: coverage

				δ								
n	δ_2	p	test	0	0.125	0.25	0.375	0.5	0.625	0.75	0.875	1
100	1	100	DM	84.10	65.62	43.53	23.28	9.84	3.83	2.27	2.01	2.00
			DM.step	83.28	63.91	40.92	20.40	7.91	3.18	2.15	2.00	2.00

Table 10: Marginal confidence set for the rank of the most populous CZ: length

Appendix G Supporting Results for the Empirical Applications

G.1 PISA Student Tests in OECD Countries: Math and Science Proficiency

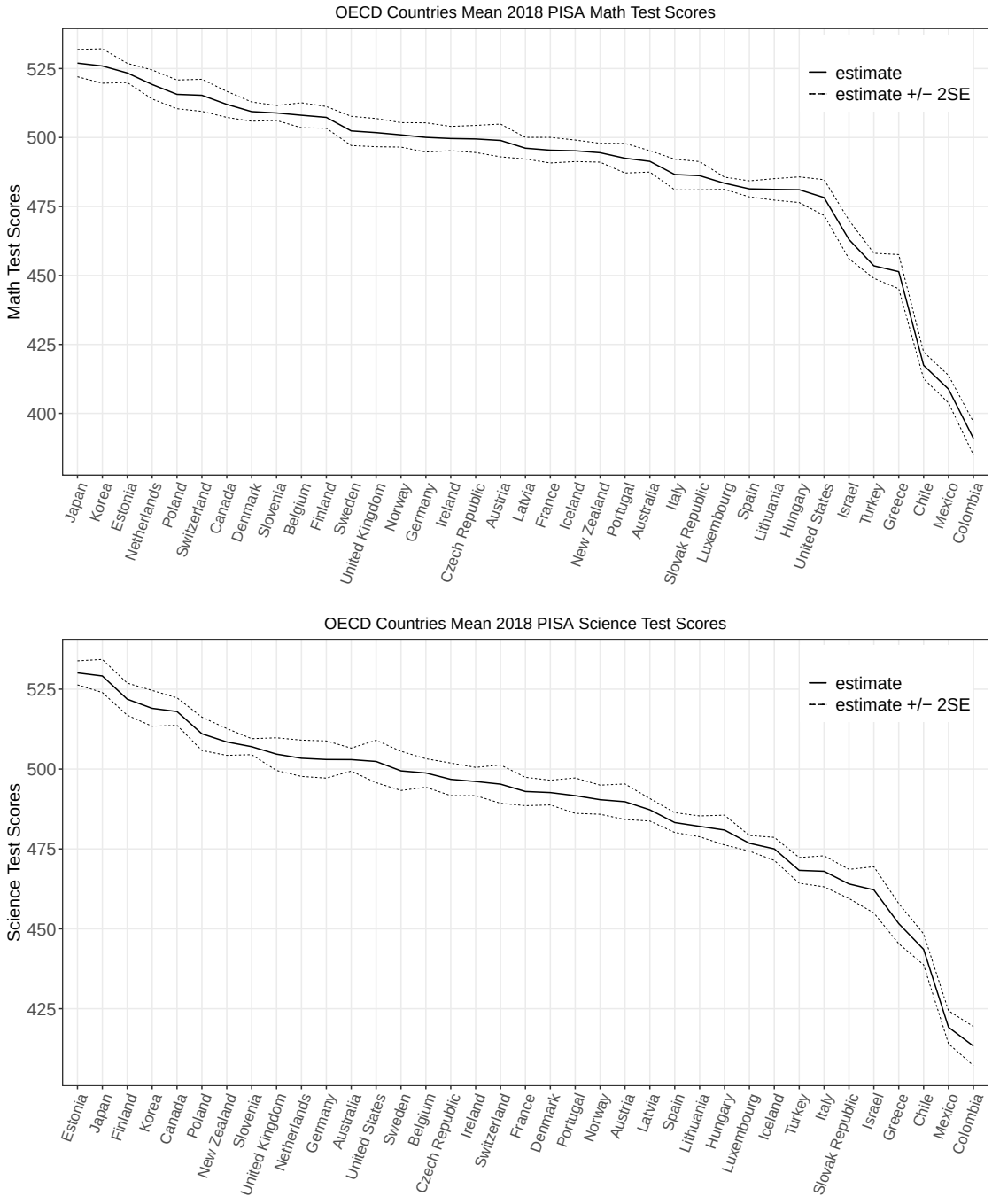

Figure 25: Mean PISA Test Scores with marginal confidence intervals (estimates plus or minus twice the standard errors) for the sample of OECD countries. The PISA scale is normalized to approximately fit a normal distribution with the mean of 500 and standard deviation of 100. **Top:** Math Test Score; **Bottom:** Science Proficiency Test Score.

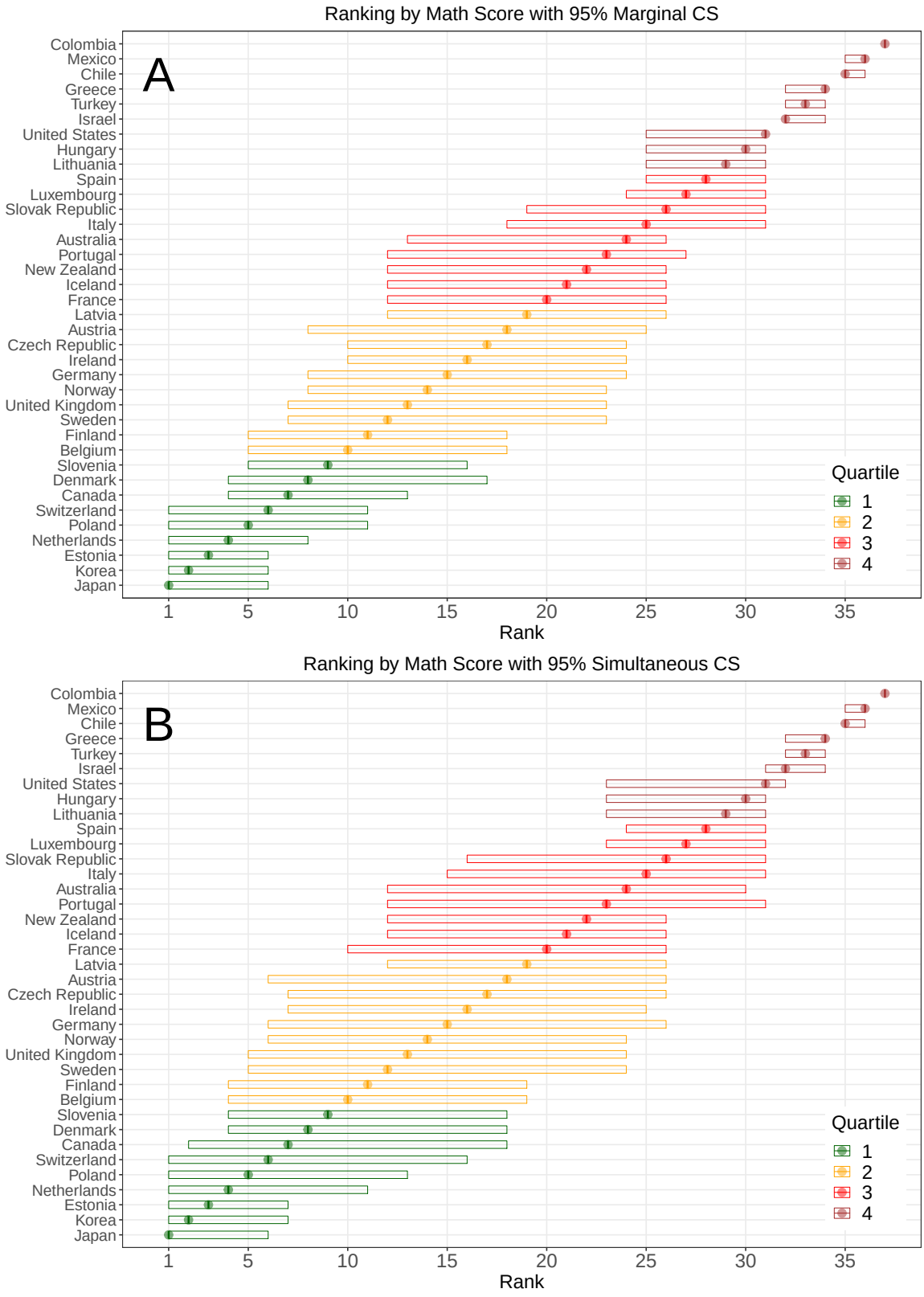

Figure 26: **Panel A:** for each OECD country, we plot its rank by math score and the 95% marginal confidence set ("CS"). **Panel B:** for each OECD country, we plot its rank by math score and the 95% simultaneous confidence set ("CS"). Different quartiles of the rankings are indicated with different colors.

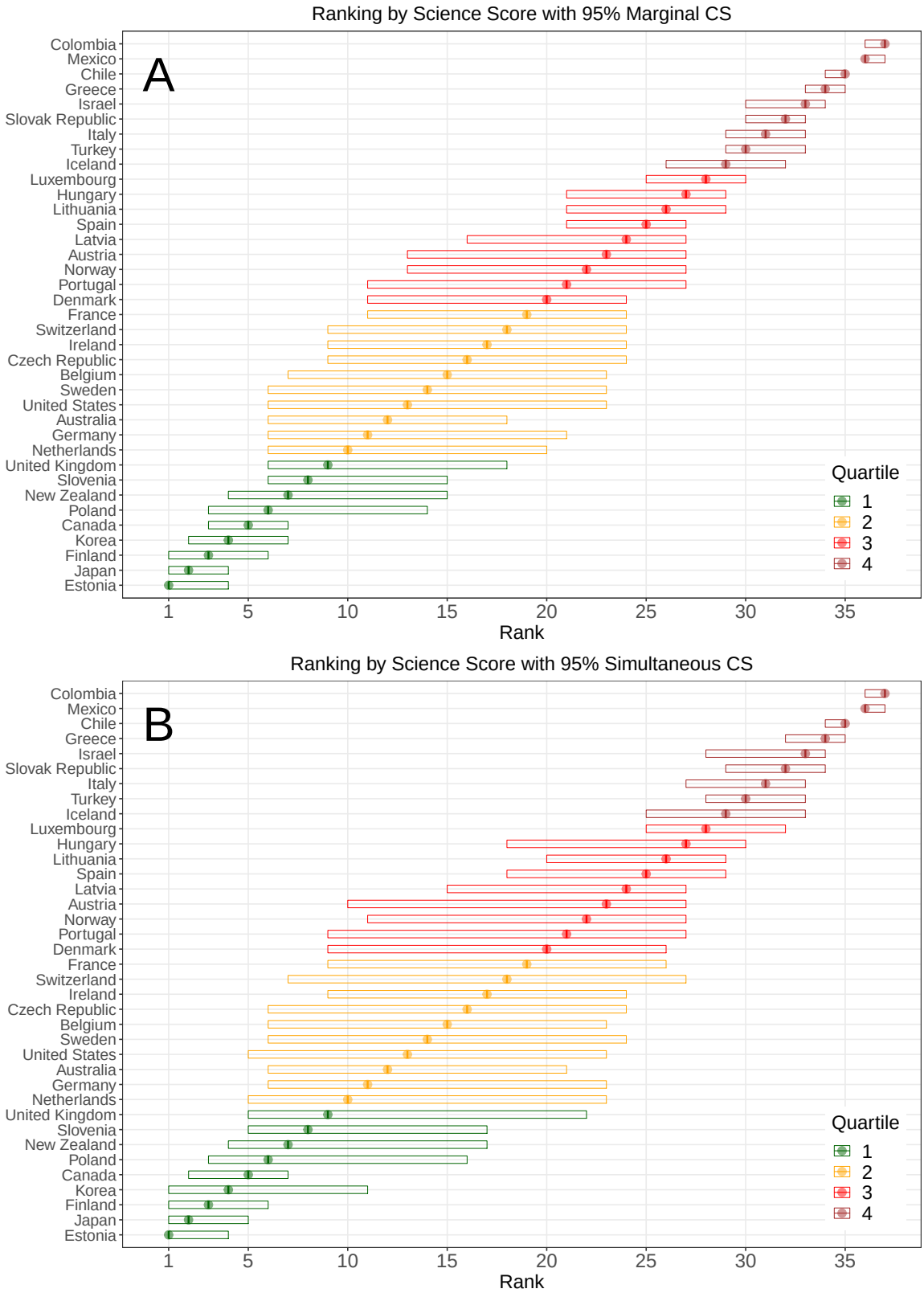

Figure 27: **Panel A:** for each OECD country, we plot its rank by science proficiency score and the 95% marginal confidence set ("CS"). **Panel B:** for each OECD country, we plot its rank by science proficiency score and the 95% simultaneous confidence set ("CS"). Different quartiles of the rankings are indicated with different colors.

G.2 Applying Andrews et al. (2018) to the PISA Data

This subsection provides more details and the results for the application of Andrews et al. (2018) to the PISA data as discussed in Remark 5.1.

The results are reported in Table 11. Following Andrews et al. (2018), we consider two types of confidence sets. While we refer to their paper for details, it is useful to observe how these confidence sets differ. The first type of confidence set is valid conditional on the target selected, in this case on the identity of the country with the highest point estimate, Estonia (the "winner" in the sample). Conditional validity is more demanding but may be desirable in the PISA setting, for example if one wants to ensure validity conditional on the recommendation made to a policy maker in Estonia. The second type of confidence set is unconditional and requires validity on average over the countries that statistically could have had the highest point estimate. For unconditional validity, Andrews et al. (2018) develop two procedures for constructing confidence sets, a projection and a hybrid procedure, the latter typically delivering shorter confidence sets than the former while satisfying the same unconditional coverage guarantee. Confidence sets for the sample "losers" are construct in an analogous fashion.

We find that both the conditional and the unconditional hybrid confidence sets are relatively narrow while the unconditional projection confidence set is wide. For example, the 95% conditional confidence set for the expected reading score of the sample "winner", Estonia, is (517.9, 526.6). Only 4 out of 35 other countries have higher point estimates than the lower endpoint of this confidence set. In conclusion, one can be confident that the sample "winner" truly has a high reading score. However, just like Estonia's marginal confidence interval displayed in Figure 8, Andrews et al. (2018)'s confidence set does not allow us to draw any conclusions about what is the true rank of Estonia nor which country has true rank one. On the other hand, our marginal confidence sets for the rank of Estonia tell us that (with 95% probability) its true rank lies between 1 and 5. In addition, our τ-best confidence set for $\tau = 1$ shows that (with 95% probability) there are 6 countries in total that could be the best.

Type of CS	Reading			
	CS on PISA score for sample "winner"	Number of countries with sample score above CS lower bound	CS on PISA score for sample "loser"	Number of countries with sample score below CS upper bound
Conditional:	(517.9, 526.6)	4	(406.0, 419.3)	1
Unconditional:				
Projection	(458.6, 587.4)	31	(298.6, 526.1)	31
Hybrid	(517.8 526.7)	4	(405.8, 419.5)	1

Table 11: Andrews et al. (2018) equal-tailed 95% confidence sets for the expected score of the countries with the highest (sample "winner") and the lowest (sample "loser") point estimates on PISA Reading test among the OECD countries.
"Number of countries with sample score above CS lower bound" shows the number of point estimates above the lower bound of the corresponding 95% confidence set for the sample "winner". "Number of countries with sample score below CS upper bound" shows the number of point estimates below the upper bound of the corresponding 95% confidence set for the sample "loser".
Conditional 95% CS provides coverage conditional on a particular country with the highest (lowest) point estimate. Projection 95% CS provides coverage on average over the countries that could statistically get the highest or the lowest point estimates. Hybrid 95% CS combines conditional and unconditional approaches to provide unconditional coverage with improved length of the CS.

G.3 50 Most Populous Commuting Zones and Counties

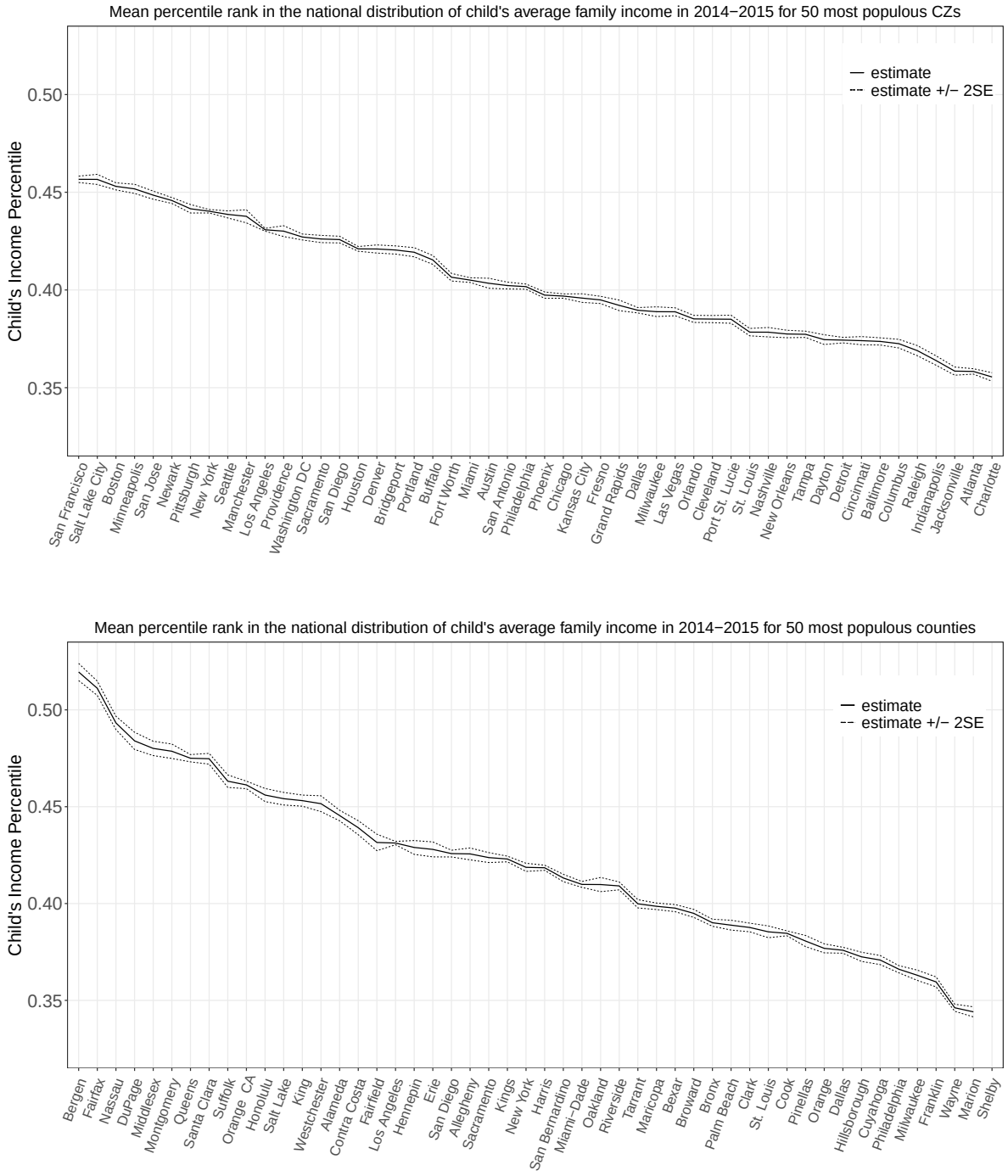

Figure 28: Estimates of the mean percentile rank of child's average household income for 2014-2015 in the national distribution of her cohort ($\bar{y}_{c25}$) with marginal confidence intervals (estimates plus or minus twice the standard errors) from Chetty et al. (2018) for the 50 most populous Commuting Zones (**Top Panel**) and the 50 most populous counties (**Bottom Panel**).

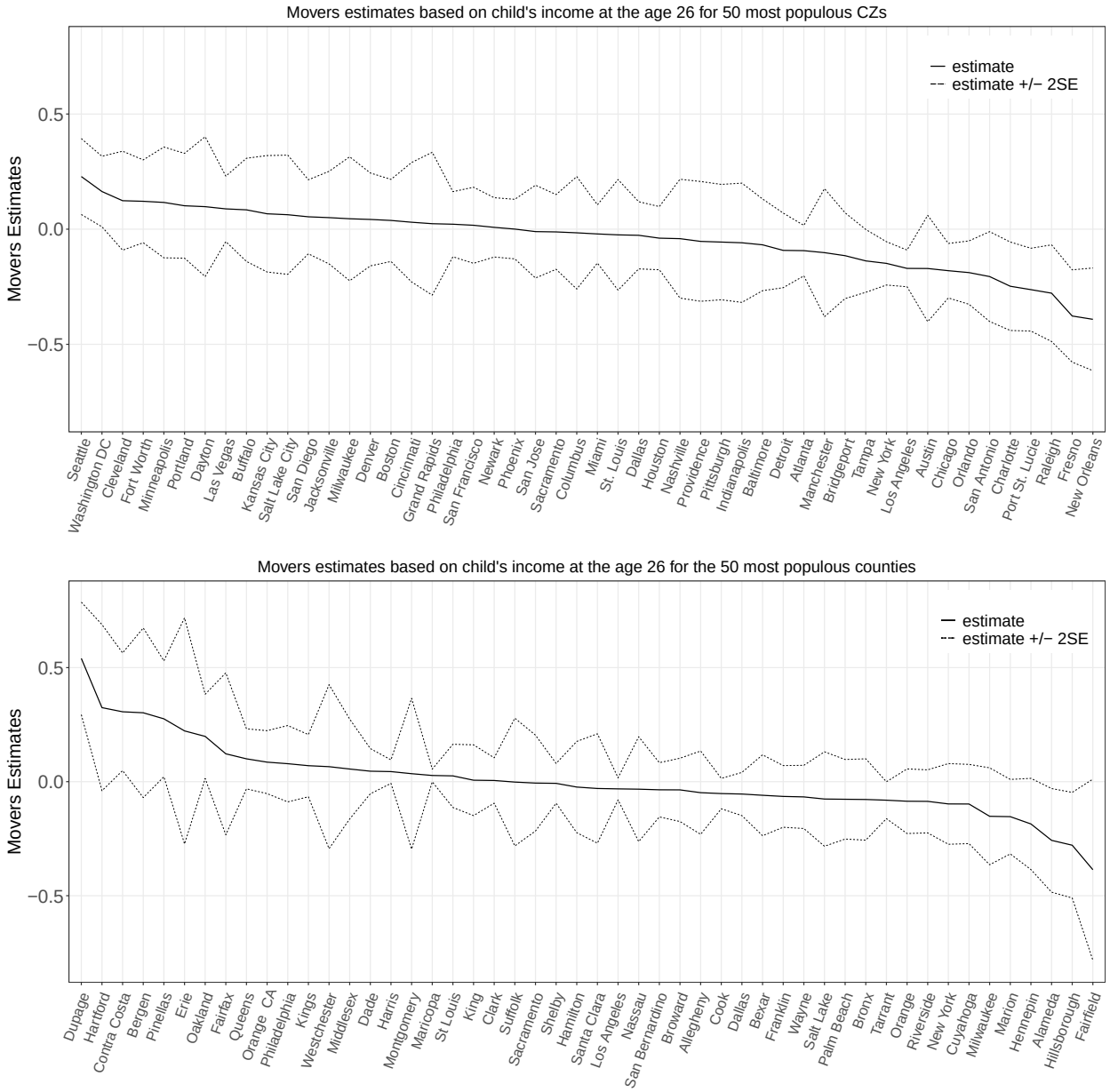

Figure 29: Movers estimates of the exposure effects (μ_{c25}) with marginal confidence intervals (estimates plus or minus twice the standard errors) from Chetty and Hendren (2018) for the 50 most populous CZs (**Top Panel**) and for the 50 most populous counties (**Bottom Panel**).

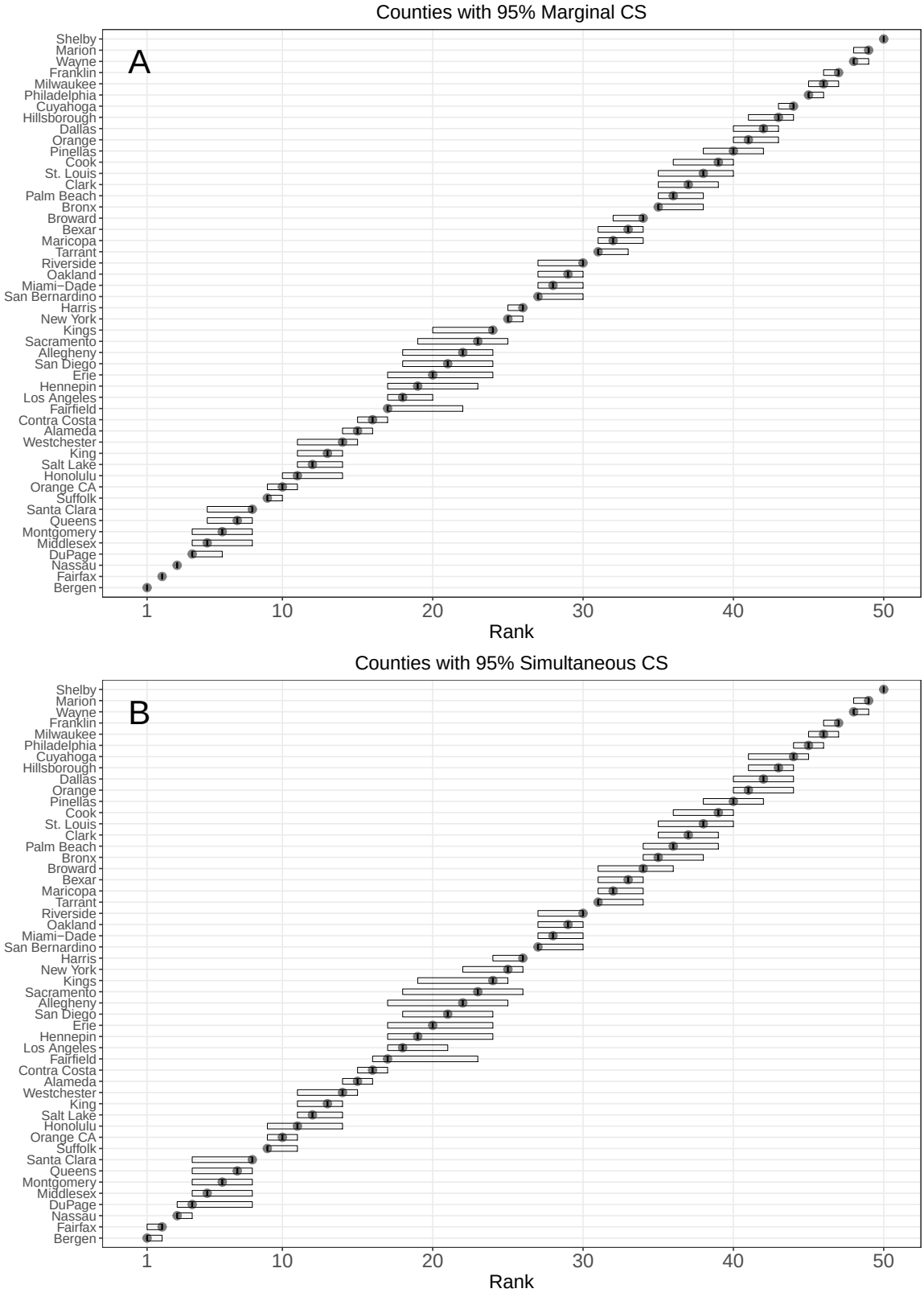

Figure 30: **Panel A:** point estimates and the 95% marginal confidence sets ("CS") for the ranking of the 50 most populous counties by $\bar{y}_{c25}$. **Panel B:** point estimates and the 95% simultaneous confidence sets ("CS") for the ranking of the 50 most populous counties by $\bar{y}_{c25}$.

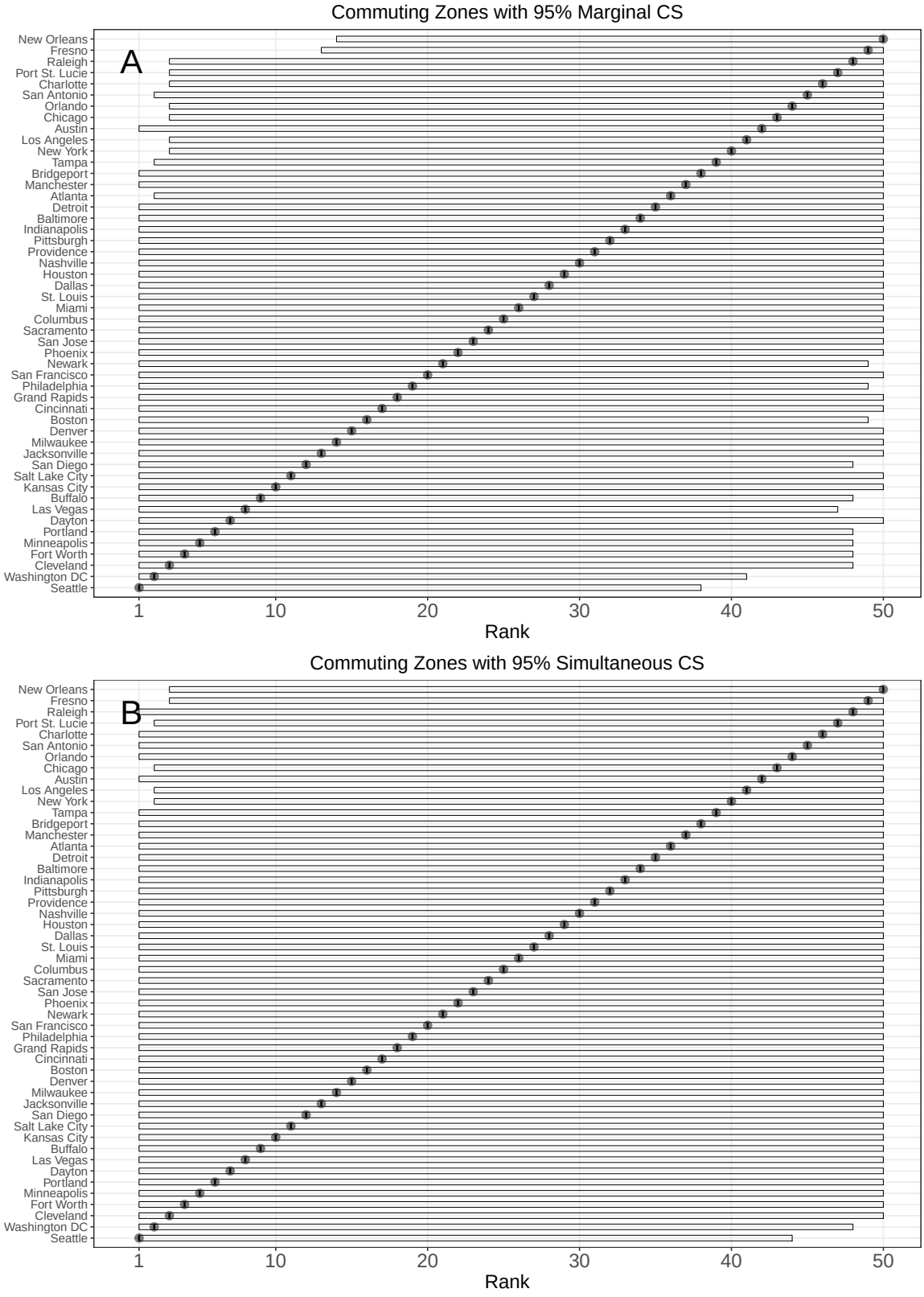

Figure 31: **Panel A:** point estimates and the 95% marginal confidence sets ("CS") for the ranking of the 50 most populous CZs by μ_{c25}. **Panel B:** point estimates and the 95% simultaneous confidence sets ("CS") for the ranking of the 50 most populous CZs by μ_{c25}.

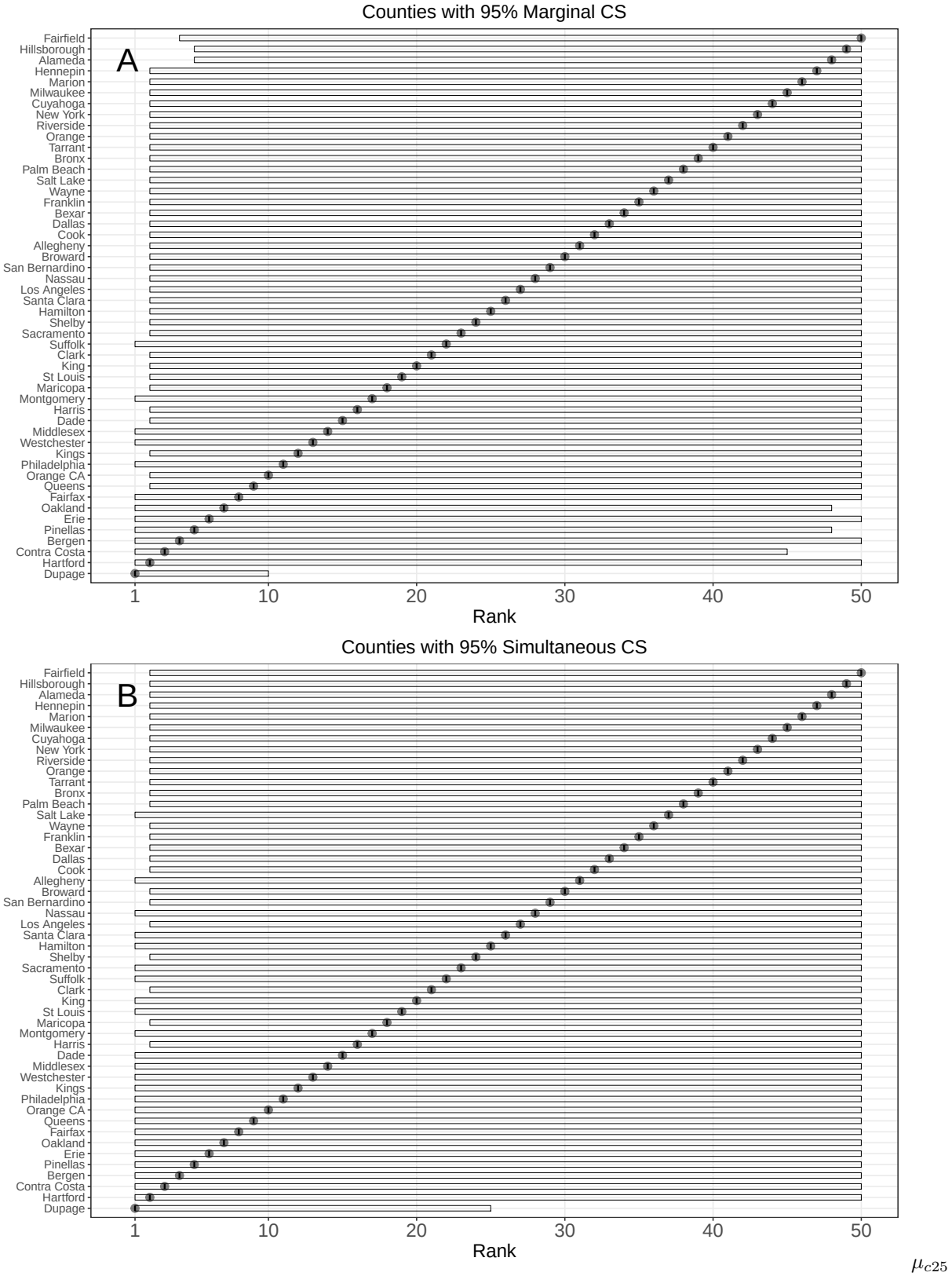

Figure 32: **Panel A:** point estimates and the 95% marginal confidence sets ("CS") for the ranking of the 50 most populous counties by μ_{c25}. **Panel B:** point estimates and the 95% simultaneous confidence sets ("CS") for the ranking of the 50 most populous counties by μ_{c25}.

Panel A: Top 5

		Correlational					Movers				
Rank	τ	County	$\hat{y}_{c25}$	SE	95% CS	τ-best	County	$\hat{\mu}_{c25}$	SE	95% CS	τ-best
1	1	Bergen	0.520	0.002	[1, 1]	2	DuPage	0.540	0.123	[1, 10]	25
2	2	Fairfax	0.511	0.002	[2, 2]	2	Hartford	0.325	0.182	[1, 50]	50
3	3	Nassau	0.493	0.002	[3, 3]	4	Contra Costa	0.306	0.129	[1, 45]	50
4	4	DuPage	0.484	0.002	[4, 6]	8	Bergen	0.302	0.186	[1, 50]	50
5	5	Middlesex	0.480	0.002	[4, 8]	8	Pinellas	0.276	0.127	[1, 48]	50

Panel B: Bottom 5

		Correlational					Movers				
Rank	τ	County	$\hat{y}_{c25}$	SE	95% CS	τ-worst	County	$\hat{\mu}_{c25}$	SE	95% CS	τ-worst
46	5	Milwaukee	0.363	0.001	[45, 47]	6	Marion	-0.153	0.082	[2, 50]	49
47	4	Franklin	0.360	0.001	[46, 47]	5	Hennepin	-0.185	0.100	[2, 50]	49
48	3	Wayne	0.346	0.001	[48, 49]	3	Alameda	-0.257	0.114	[5, 50]	49
49	2	Marion	0.344	0.001	[48, 49]	3	Hillsborough	-0.279	0.116	[5, 50]	49
50	1	Shelby	0.318	0.001	[50, 50]	1	Fairfield	-0.386	0.199	[4, 50]	49

Table 12: **Panel A:** Top 5 among the 50 most populous counties ranked by the correlational estimates on the left and by the movers estimates on the right. **Panel B:** Bottom 5 among the 50 most populous counties ranked by the correlational estimates on the left and by the movers estimates on the right. "95% CS" refers to the 95% marginal confidence set for the rank, and "τ-best" and "τ-worst" refer to the size of the 95% confidence sets for the "τ-best" and "τ-worst" counties.

Confidence level	CZ		County	
	Correlational	Movers	Correlational	Movers
0.95	1	0	4	0
0.90	1	0	7	0
0.75	1	0	7	0
0.50	1	0	7	0
0.25	3	0	7	0
0.10	4	0	8	0
0.05	4	0	8	0

Table 13: The number of places with the same rank at the upper and the lower endpoints of the 95% confidence sets for a given confidence level. The analyses are done for both the correlational and the mover estimates, using the 50 most populous CZs and the 50 most populous counties.

G.4 Applying Andrews et al. (2018) to the Intergenerational Mobility Data

This subsection provides more details and the results for the application of Andrews et al. (2018) to the intergenerational mobility data as discussed in Remark 5.3.

We focus on correlational estimates for CZs, as these are the most precise estimates and, thus, offer the best opportunity to reach informative conclusions. Table 14 reports the results. Following Andrews et al. (2018), we consider the 95% conditional and unconditional (hybrid and projection) confidence sets for a neighborhood's true mobility. As discussed in Section G.2, the former is valid conditional on the identity of the neighborhood with the highest point estimate, while the latter requires validity on average over all the neighborhoods that statistically could have had the highest point estimate. Conditional validity is more demanding but may be desirable if the goal is to help low-income families move to the neighborhood with highest estimated mobility.

We first consider inference on the CZ with the highest point estimate in the national ranking.[21] The 95% conditional confidence set for true mobility of the "winning" CZ is (-2.70, 0.66). Thus, we may expect, with 95% confidence, an income rank between 0 and 66 among children who grew up in the CZ with the highest point estimate of mobility with parents' income at the 25th percentile.[22] Since the confidence set includes zero (the smallest possible value of the mobility measure), one cannot be confident that the sample "winner" truly has high mobility. This conclusion holds both if we consider the conditional confidence set and unconditional (hybrid or projection) confidence sets.

Next, we restrict the inference problem to the 50 largest CZs by population size. The 95% conditional confidence set on upward mobility of the CZ with the highest point estimate, San Fransisco, is (0.389, 0.457). Thus, we may expect, with 95% confidence, an income rank between 39 and 46 of the children who grew up in San Francisco with parents at the 25th percentile. The lower endpoint of this confidence set is lower than the point estimates of 31 of the 49 other CZs. By comparison, the 95% unconditional hybrid confidence set is narrower than the conditional ones, but still contains the point estimates of 19 of the 49 other CZs.

Taken together, the results from Andrews et al. (2018) suggest there is considerable statistical uncertainty about the true value of upward mobility at the top of the estimated ranking of CZs, even if one restricts the study to the 50 largest CZs by population size.

However, just like the marginal confidence interval for the "winner" displayed in Figure 10, Andrews et al. (2018)'s confidence sets do not allow us to draw any conclusions about what is the true rank of the sample "winner" nor which CZ has true rank one. In contrast, our confidence set for the rank of San Francisco among the 50 most populous CZs tells us that (with 95% probability) its true rank lies between 1 and 2. In addition, our τ-best confidence set for $\tau = 1$ shows that (with 95% probability) there are only 4 CZs that could be the best. It is interesting to note that our confidence sets for the ranks are very narrow even though Andrews et al. (2018)'s confidence sets for true mobility of the "winner" are fairly wide. This finding illustrates that it can be possible to achieve a statistically informative ranking even if one cannot draw firm conclusions about the true value of the sample "winner".

[21] We have also performed inference on the neighborhoods with the highest point estimate within each state. In many cases, the lower endpoint of the 95% confidence sets include the point estimates of a majority of the other neighborhoods within the state.

[22] Although percentiles can only take values between 0 and 1, the estimated confidence sets of Andrews et al. (2018) may take negative value. Following their suggestion, we therefore trim the confidence sets when interpreting the results by substituting negative lower endpoints with zero.

Type of CS	All 741 CZs		50 most populous CZs	
	CS on correlational mobility for sample "winner"	Number of CZs with correlational estimates above CS lower bound	CS on correlational mobility for sample "winner"	Number of CZs with correlational estimates above CS lower bound
Conditional:	(-2.70, 0.66)	741	(0.389, 0.457)	31
Trimmed	(0, 0.66)	741		
Unconditional:				
Projection	(-14.95, 16.28)	741	(0.417, 0.496)	19
Trimmed	(0, 16.28)	741		
Hybrid	(-2.79, 0.66)	741	(0.416, 0.457)	19
Trimmed	(0, 0.66)	741		

Table 14: Andrews et al. (2018) equal-tailed 95% confidence sets for true upward mobility of the CZ with the highest correlational point estimate (sample "winner") among all 741 CZs and among the 50 most populous CZs.
"Number of CZs with correlational estimates above CS lower bound" refers to the number of point estimates above the lower bound of the corresponding confidence set for the sample "winner".
Conditional 95% CS provides coverage conditional on the identity of the CZ with the highest mobility point estimate. Projection 95% CS provides coverage on average over the CZs that could statistically have had the highest point estimate. Hybrid 95% CS combines conditional and unconditional approaches to provide unconditional coverage with improved length of the CS.

G.5 Heat Maps of $\hat{\bar{y}}_{c25}$ for Counties and Structure of the Rankings

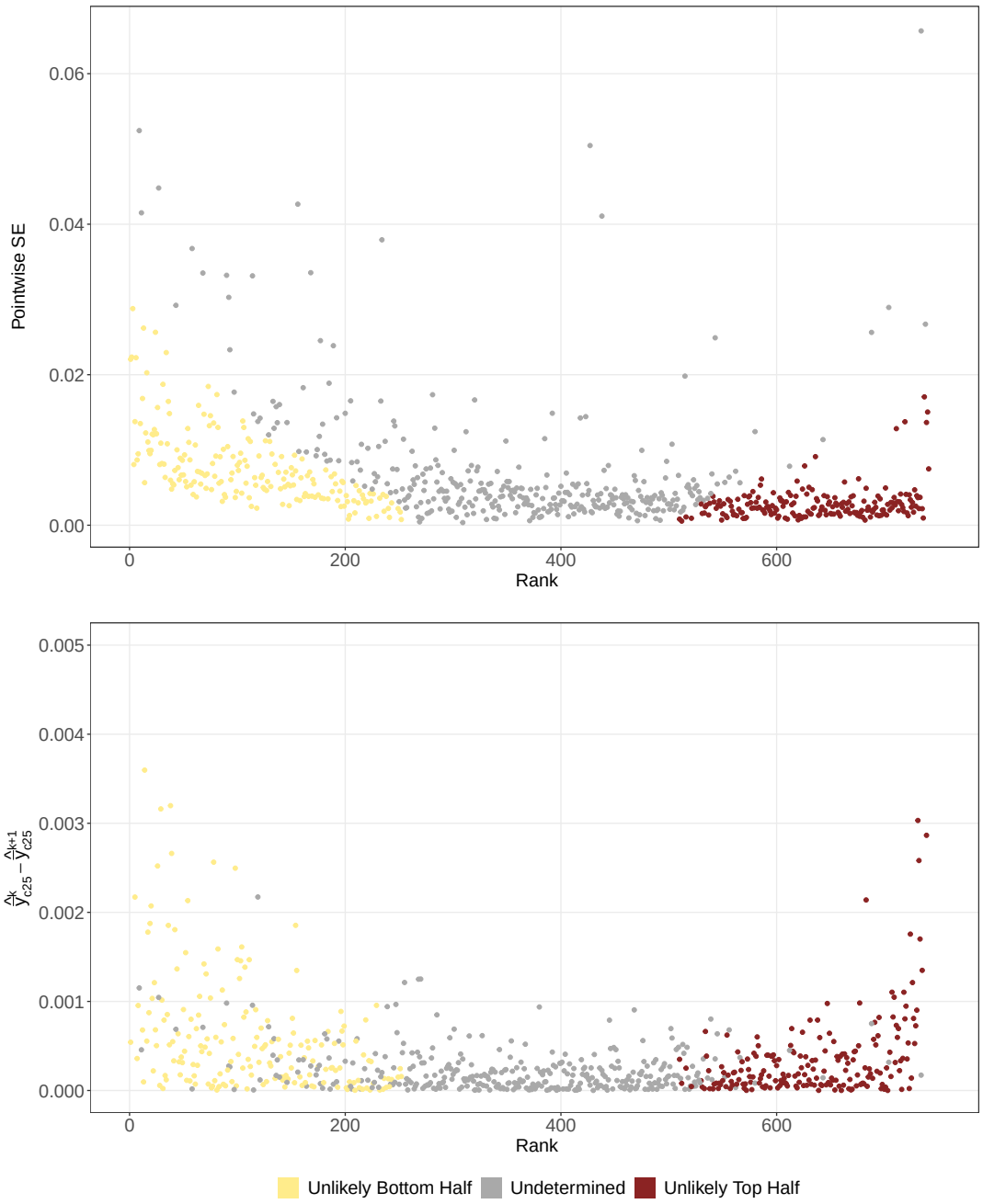

Figure 33: **Top Panel** for each CZ we plot the standard error ("SE") against the rank of the CZ. **Bottom Panel** for each CZ we compute the difference in estimated mobility ($\hat{\bar{y}}_{c25}$) between the CZ ($\hat{\bar{y}}^k_{c25}$) and the next CZ ($\hat{\bar{y}}^{k+1}_{c25}$) in the estimated ranking. Next, we plot these differences against the estimated ranks of CZs. Each dot on both panels represents a CZ. The CZ is assigned to a high mobility group (light color) if the upper endpoint of its simultaneous confidence set is in the top half of the national ranking. The CZ is assigned to a low mobility group (red color) if the lower endpoint of its simultaneous confidence set is in the bottom half of the national ranking. Grey colors represent the CZs with simultaneous confidence sets such that the places cannot be assigned to either the high or the low mobility group. We emphasize two points: 1) the middle of the ranking does not have particularly large SEs; 2) in the middle of the ranking estimates of mobility are more similar.

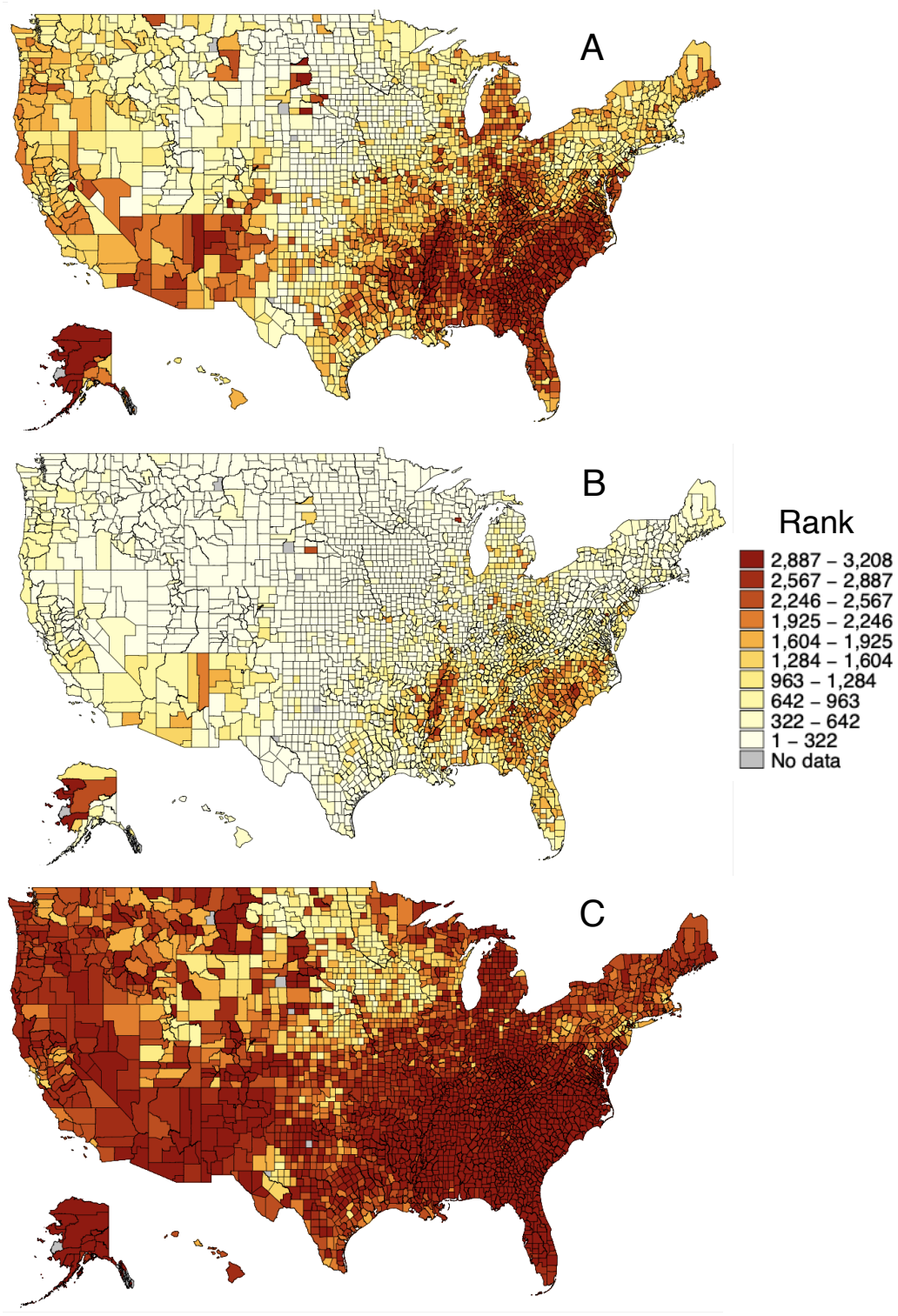

Figure 34: Ranking of counties by point estimates and lower and upper endpoints of simultaneous confidence sets. The heat maps are based on estimates of $\bar{y}_{c25}$, the mean percentile rank of child's average household income for 2014-2015, for the full set of counties. **Panel A:** the map is constructed by dividing the counties into deciles based on the estimated values of $\bar{y}_{c25}$, and shading the areas so that lighter colors correspond to higher absolute mobility or, equivalently, lower ("better") rank. **Panel B** (**Panel C**) shows the lower (upper) endpoint of the 95% simultaneous confidence sets for the ranks of all counties, using the same color coding as for the estimated ranks in Panel A.

Figure 35: For each county, we compute the difference between the upper and the lower endpoint of the 95% simultaneous confidence set. Next, we plot these differences against the estimated ranks of the counties. To ease interpretation, we normalize the differences by the number of counties. Thus, a difference of 1 means one cannot tell whether a county has the highest or the lowest income mobility in the United States. By comparison, a difference of 0 means we can be confident in the exact rank of the county. Each dot in the graph represents a county. The county is assigned to a high mobility group (light color) if the upper endpoint of its simultaneous confidence set is in the top half of the national ranking. The county is assigned to a low mobility group (red color) if the lower endpoint of its simultaneous confidence set is in the bottom half of the national ranking. Grey colors represent the counties with simultaneous confidence sets such that the places cannot be assigned to either the high or the low mobility group.

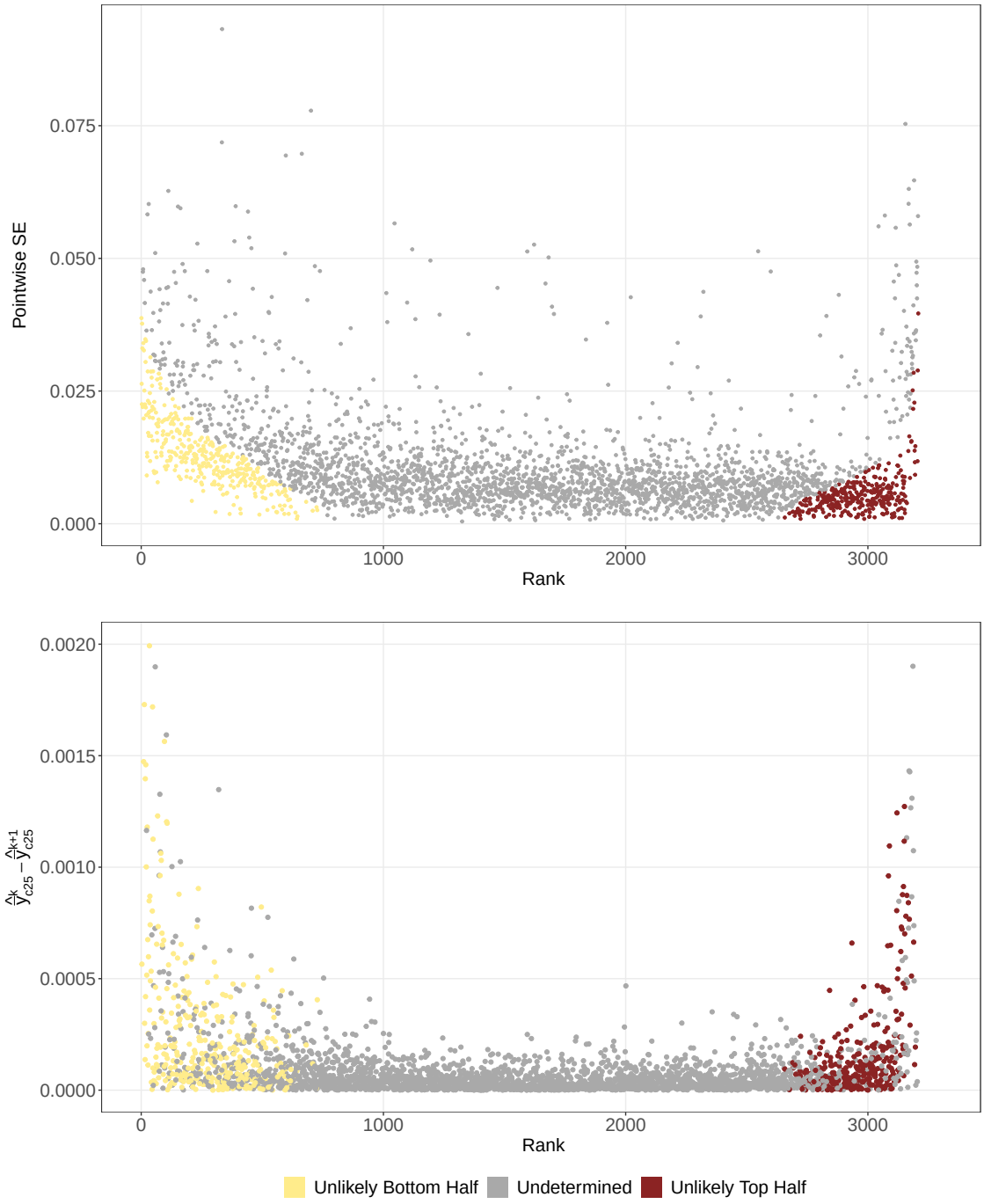

Figure 36: **Top Panel** for each county we plot the standard error ("SE") against the rank of the county. **Bottom Panel** for each county we compute the difference in estimated mobility ($\hat{\bar{y}}_{c25}$) between the county ($\hat{\bar{y}}^k_{c25}$) and the next county ($\hat{\bar{y}}^{k+1}_{c25}$) in the estimated ranking. Next, we plot these differences against the estimated ranks of the counties. Each dot on both panels represents a county. The county is assigned to a high mobility group (light color) if the upper endpoint of its simultaneous confidence set is in the top half of the national ranking. The county is assigned to a low mobility group (red color) if the lower endpoint of its simultaneous confidence set is in the bottom half of the national ranking. Grey colors represent the counties with simultaneous confidence sets such that the places cannot be assigned to either the high or the low mobility group. We emphasize two points: 1) the middle of the ranking does not have particularly large SEs; 2) in the middle of the ranking estimates of mobility are more similar.

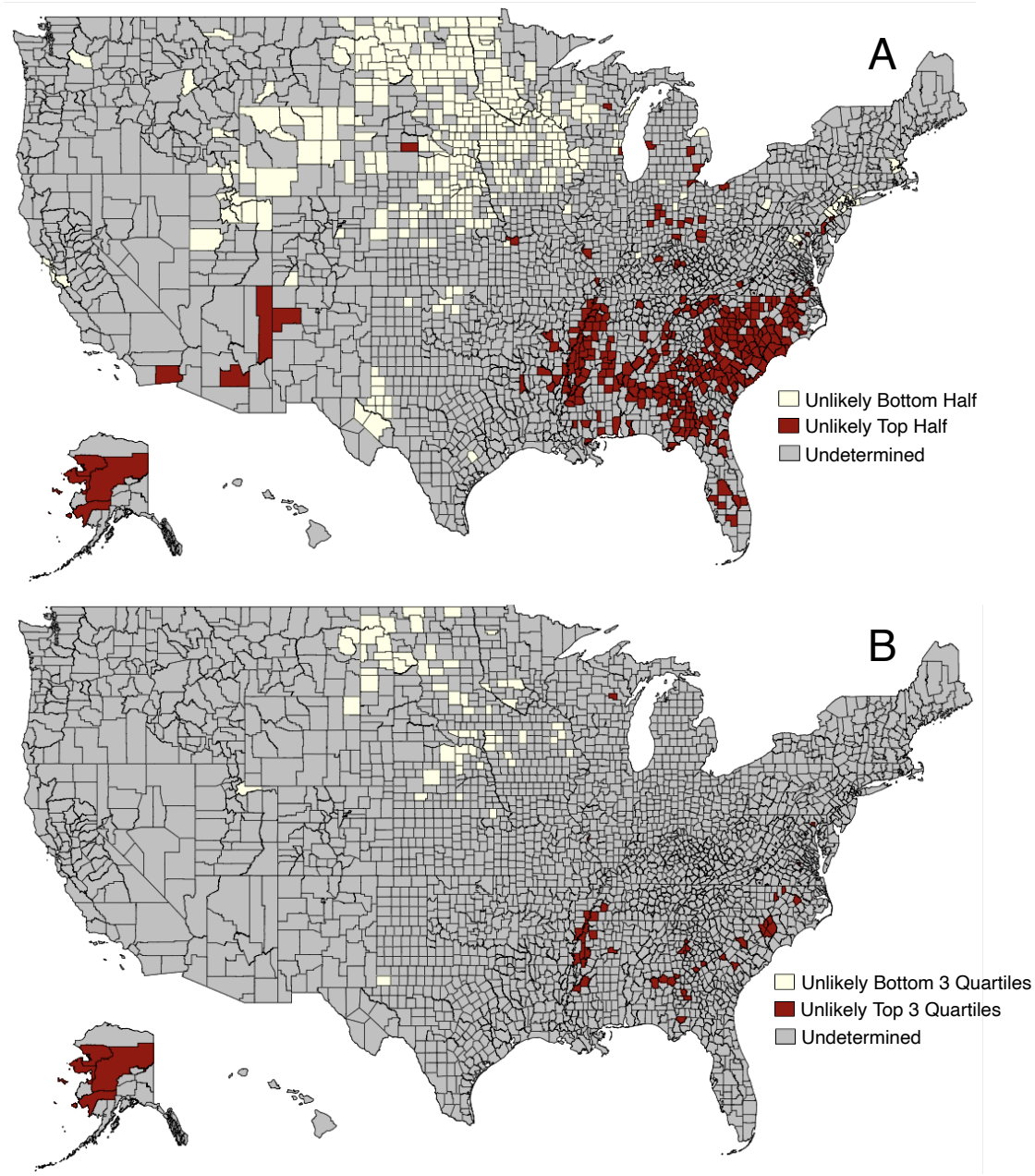

Figure 37: The heat map in **Panel A** is constructed by assigning the counties to groups depending on the lower and upper endpoints of the simultaneous confidence sets. A county is assigned to a high mobility group, **Unlikely Bottom Half**, if the upper endpoint of its simultaneous confidence set is in the top half of the national ranking of counties, i.e. when the confidence set lies entirely in the top half of the ranking, indicating high mobility. A county is assigned to a low mobility group, **Unlikely Top Half**, if the lower endpoint of its simultaneous confidence set is in the bottom half of the national ranking of counties, i.e. when the confidence set lies entirely in the bottom half of the ranking, indicating low mobility. Grey colors represent the counties with simultaneous confidence sets such that the places cannot be assigned to either the high or the low mobility group, i.e. the **Undetermined** counties. The heat map in **Panel B** is constructed in the same way, except the high and low mobility groups are now defined in terms of top and bottom quartiles in the national ranking of the counties. Thus, we refer to these groups as **Unlikely Bottom 3 Quartiles** and **Unlikely Top 3 Quartiles**.

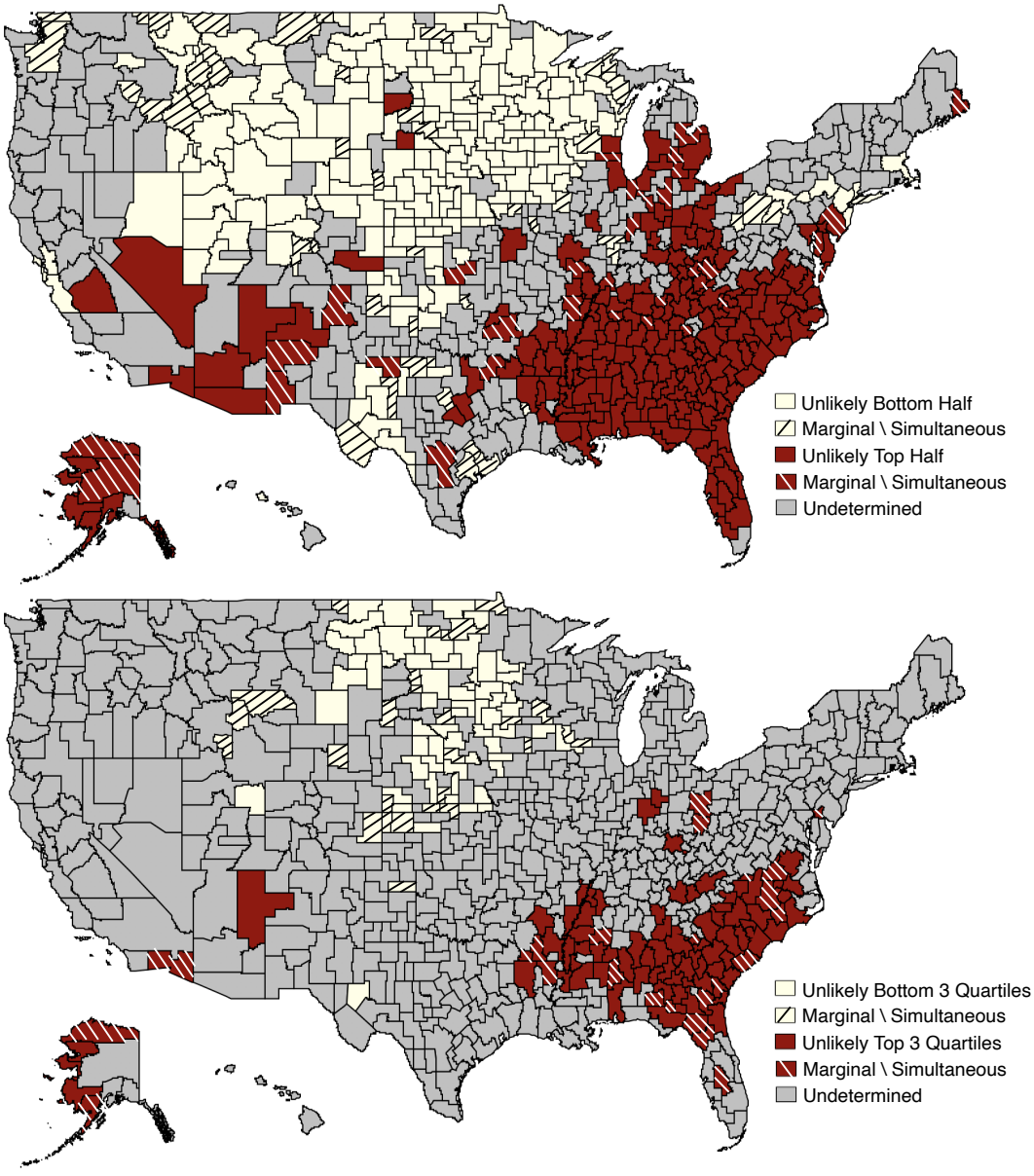

Figure 38: The heat map in **Panel A** is constructed by assigning the CZs to groups depending on the lower and upper endpoints of the simultaneous and marginal confidence sets. A CZ is assigned to a high mobility group, **Unlikely Bottom Half**, if the upper endpoint of its simultaneous confidence set is in the top half of the national ranking of CZs, i.e., when the confidence set lies entirely in the top half of the ranking, indicating high mobility. A CZ is assigned to a high mobility group, **Marginal \ Simultaneous**, if the upper endpoint of its marginal confidence set is in the top half of the national ranking of CZs but the upper endpoint of its simultaneous confidence set is in the bottom half of the national ranking of CZs. A CZ is assigned to a low mobility group, **Unlikely Top Half**, if the lower endpoint of its simultaneous confidence set is in the bottom half of the national ranking of CZs, i.e., when the confidence set lies entirely in the bottom half of the ranking, indicating low mobility. A CZ is assigned to a low mobility group, **Marginal \ Simultaneous**, if the lower endpoint of its marginal confidence set is in the bottom half of the national ranking of CZs but the lower endpoint of its simultaneous confidence set is in the top half of the national ranking of CZs. Grey colors represent the CZs with marginal and simultaneous confidence sets such that the places cannot be assigned to either the high or the low mobility group, i.e., the **Undetermined** CZs. The heat map in **Panel B** is constructed in the same way, except the high and low mobility groups are now defined in terms of top and bottom quartiles in the national ranking of the CZs. Thus, we refer to these groups as **Unlikely Bottom 3 Quartiles** and **Unlikely Top 3 Quartiles**.

G.6 Heat Maps for the Movers Estimates of the Exposure Effects

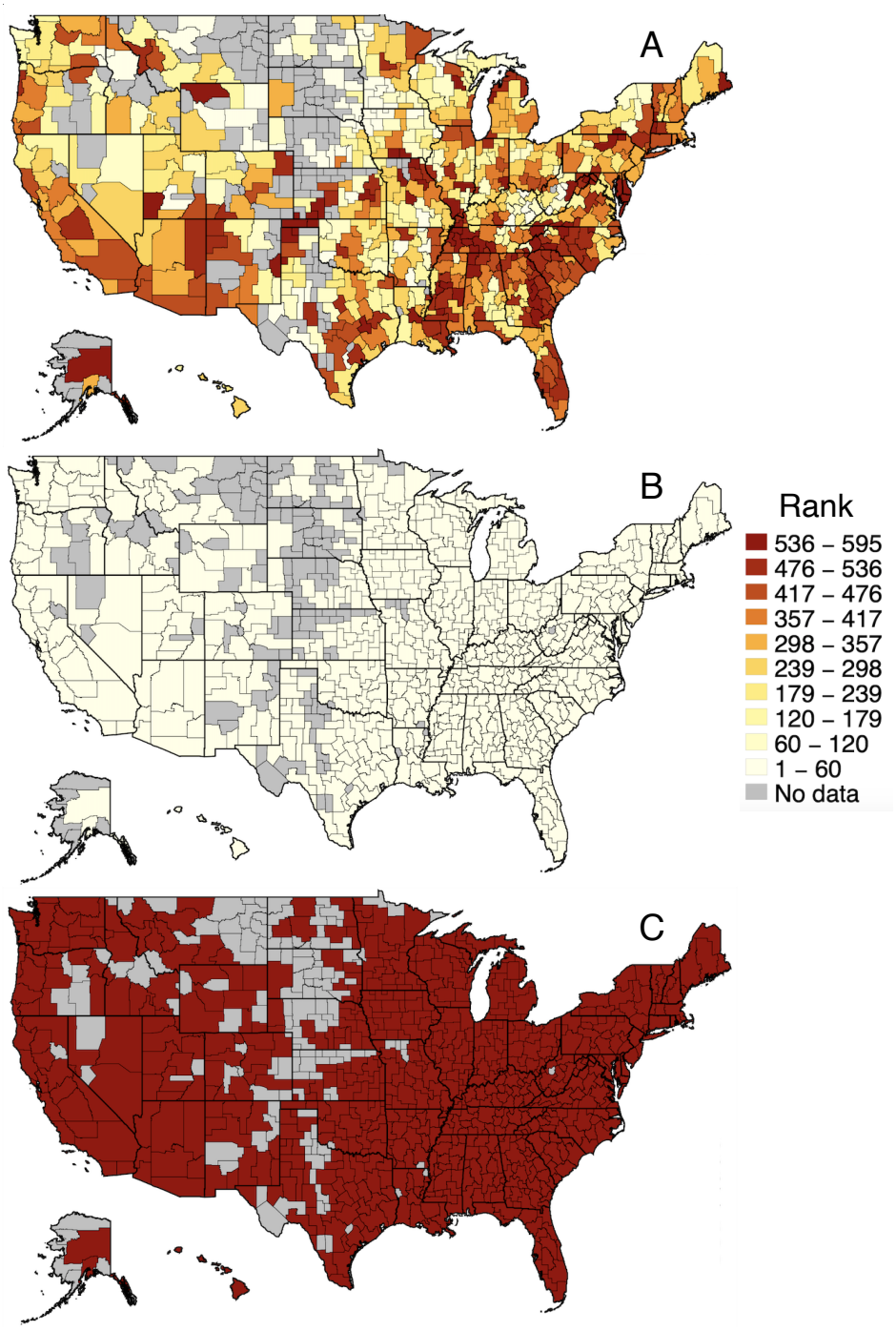

Figure 39: Ranking of Commuting Zones by point estimates and lower and upper endpoints of simultaneous confidence sets. The heat maps are based on movers estimates of the exposure effects μ_{c25}. **Panel A:** the map is constructed by dividing the CZs into deciles based on the estimated values of μ_{c25}, and shading the areas so that lighter colors correspond to higher values of exposure effects or, equivalently, lower ("better") rank. **Panel B** (**Panel C**) shows the lower (upper) endpoint of the 95% simultaneous confidence sets for the ranks of all CZs, using the same color coding as for the estimated ranks in Panel A.

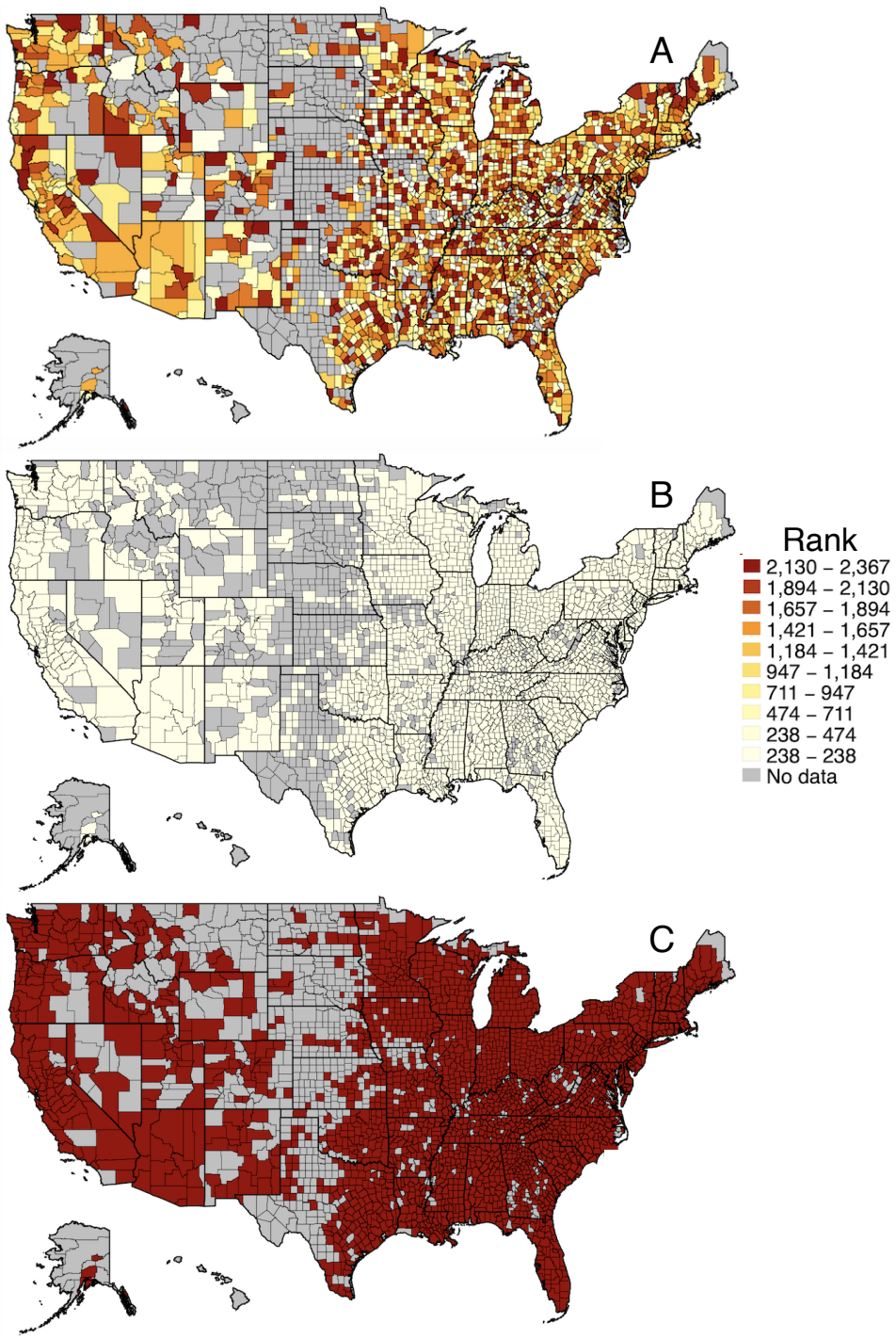

Figure 40: Ranking of counties by point estimates and lower and upper endpoints of simultaneous confidence sets. The heat maps are based on movers estimates of the exposure effects μ_{c25}. **Panel A:** the map is constructed by dividing the counties into deciles based on the estimated values of μ_{c25}, and shading the areas so that lighter colors correspond to higher values of exposure effects or, equivalently, lower ("better") rank. **Panel B** (**Panel C**) shows the lower (upper) endpoint of the 95% simultaneous confidence sets for the ranks of all counties, using the same color coding as for the estimated ranks in Panel A.

References

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

Andrews, I., Kitagawa, T. and McCloskey, A. (2018). Inference on winners. Working Paper CWP 31/18, CeMMAP.

Bai, Y., Santos, A. and Shaikh, A. (2019). A practical method for testing many moment inequalities. *University of Chicago, Becker Friedman Institute for Economics Working Paper*.

Bauer, P., Hackl, P., Hommel, G. and Sonnemann, E. (1986). Multiple testing of pairs of one-sided hypotheses. *Metrika*, **33** 121–127.

Bergman, P., Chetty, R., DeLuca, S., Hendren, N., Katz, L. F. and Palmer, C. (2019). Creating moves to opportunity: Experimental evidence on barriers to neighborhood choice. Working Paper 26164, NBER.

Breakspear, S. (2012). The policy impact of pisa: An exploration of the normative effects of international benchmarking in school system performance. *OECD Education Working Papers*.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, **41** 2786–2819.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2017). Central limit theorems and bootstrap in high dimensions. *Annals of Probability*, **45** 2309–2352.

Chernozhukov, V., Chetverikov, D., Kato, K. and Koike, Y. (2019). Improved central limit theorem and bootstrap approximations in high-dimensions. Tech. rep.

Chetty, R. (April 1, 2014). Improving opportunities for economic mobility in the united states. *Budget Committee United States Senate*.

Chetty, R., Friedman, J. N., Hendren, N., Jones, M. R. and Porter, S. R. (2018). The opportunity atlas: Mapping the childhood roots of social mobility. Working Paper 25147, NBER.

Chetty, R., Friedman, J. N., Hendren, N., Jones, M. R. and Porter, S. R. (2020). The opportunity atlas: Mapping the childhood roots of social mobility. Tech. rep.

Chetty, R. and Hendren, N. (2018). The impacts of neighborhoods on intergenerational mobility ii: County-level estimates. *The Quarterly Journal of Economics*, **133** 1163–1228.

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** 1553–1624.

Connor, D. S. and Storper, M. (2020). The changing geography of social mobility in the united states. *Proceedings of the National Academy of Sciences*, **117** 30309–30317.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, **50** 1096–1121.

Goldstein, H. and Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **159** 385–443.

Grandhi, A., Guo, W. and Romano, J. P. (2019). Control of directional errors in fixed sequence multiple testing. *Statistica Sinica*, **29** 1047–1064.

Gu, J. and Koenker, R. (2020). Invidious comparisons: Ranking and selection as compound decisions. Tech. rep.

Guo, W. and Romano, J. P. (2015). On stepwise control of directional errors under independence and some dependence. *Journal of Statistical Planning and Inference*, **163** 21 – 33.

Gupta, S. S. (1956). *On a decision rule for a problem in ranking means*. Ph.D. thesis, Institute of Statistics, University of North Carolina, Chape Hill.

Gupta, S. S. and Panchapakesan, S. (1979). *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations*. John Wiley & Sons, New York.

Hall, P. and Miller, H. (2009). Using the bootstrap to quantify the authority of an empirical ranking. *The Annals of Statistics*, **37** 3929–3959.

Hubert, E. (2006). Educational standards and the changing discourse on education: the reception and consequences of the pisa study in germany. *Oxford Review of Education*, **32** 619–634.

Klein, M., Wright, T. and Wieczorek, J. (2020). A joint confidence region for an overall ranking of populations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **69** 589–606.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, New York.

Lehmann, E. L., Romano, J. P. and Shaffer, J. P. (2005). On optimality of stepdown and stepup multiple test procedures. *The Annals of Statistics*, **33** 1084–1108.

OECD (2017). Pisa 2015 technical report. Tech. rep., OECD.

OECD (2019). *PISA 2018 Results (Volume I): What Students Know and Can Do*. OECD Publishing, Paris.

Romano, J. P. and Shaikh, A. M. (2012). On the uniform asymptotic validity of subsampling and the bootstrap. *Annals of Statistics*, **40** 2798–2822.

Romano, J. P., Shaikh, A. M. and Wolf, M. (2014). A practical twostep method for testing moment inequalities. *Econometrica*, **82** 1979–2002.

Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, **73** 1237–1282.

Shaffer, J. P. (1980). Control of directional errors with stagewise multiple test procedures. *The Annals of Statistics*, **8** 1342–1347.

Spjøtvoll, E. (1972). On the optimality of some multiple comparison procedures. *Annals of Mathematical Statistics*, **43** 398–411.

Tukey, J. (1953). The problem of multiple comparisons. Mimeographed notes, Princeton University.

Westfall, P., Tobias, R., Rom, D., Wolfinger, R. and Hochberg, Y. (1999). *Multiple Comparisons and Multiple Tests*. SAS Institute, Cary, NC.

Xie, M., Singh, K. and Zhang, C.-H. (2009). Confidence intervals for population ranks in the presence of ties and near ties. *Journal of the American Statistical Association*, **104** 775–788.