Measuring Distances in High Dimensional Spaces

Why Average Group Vector Comparisons Exhibit Bias, And What to Do About it

Breanna Green[*], William Hobbs[†], Sofia Avila[‡],
Pedro Rodriguez[§], Arthur Spirling[¶], Brandon M. Stewart[‖]

June 26, 2024

Abstract

Analysts often seek to compare representations in high-dimensional space, e.g. embedding vectors of the same word across groups. We show that the distance measures calculated in such cases can exhibit considerable statistical bias, that stems from uncertainty in the estimation of the elements of those vectors. This problem applies to Euclidean distance, cosine similarity, and other similar measures. After illustrating the severity of this problem for text-as-data applications, we provide and validate a bias correction for the squared Euclidean distance. This same correction also substantially reduces bias in ordinary Euclidean distance and cosine similarity estimates, but corrections for these measures are not quite unbiased and are (non-intuitively) bimodal when distances are close to zero. The estimators require obtaining the variance of the latent positions. We (will) implement the estimator in free software, and we offer recommendations for related work.

The response memo and supporting information are attached with this submission and available for review.

Number of words:

3166

[*]PhD candidate, Information Science, Cornell University, https://bregreen.github.io/, begreen@infosci.cornell.edu
[†]Assistant Professor, Department of Psychology and Department of Government, Cornell University, https://hobbs.human.cornell.edu, hobbs@cornell.edu. Corresponding author.
[‡]PhD student, Department of Sociology, Princeton University, sofiaavila@princeton.edu
[§]Visiting Scholar, Center for Data Science, New York University, United States; and International Faculty, Instituto de Estudios Superiores de Administración, Venezuela, (pedro.rodriguez@nyu.edu)
[¶]Professor, Department of Politics, Princeton University, https://arthurspirling.org/, arthur.spirling@princeton.edu
[‖]Associate Professor, Department of Sociology and Office of Population Research, Princeton University, brandonstewart.org, bms4@princeton.edu

1 Motivation

Social scientists routinely represent entities as vectors in high-dimensional spaces where the elements of those vectors have been estimated (e.g., Mozer et al., 2020; Nyarko and Sanga, 2022; Rossiter, 2022; van Loon et al., 2022; Rodriguez, Spirling and Stewart, 2023; Kraft and Klemmensen, 2023). For instance, they might represent documents in terms of the modeled topic proportions they contain, or Members of Congress in terms of their estimated positions in several dimensions of ideological space. From these representations, researchers draw conclusions about the (dis)similarity, between the documents or actors in question. They do this via measured *distances* between the vectors. This calculation is typically trivial: for example, it takes very little computational effort to compare two word embedding vectors in terms of their Euclidean distance from one another. But this simple "plugin" estimator can be misleading in practice. This is because the elements of the vector are *estimated with error*, yet this uncertainty is not properly incorporated into the distance calculation. The result is an upward bias in the measurement of that distance; and this bias is worse when the vectors are more poorly estimated. This problem has been observed in several fields, but the remedies are not well known or implemented (e.g., Weir, Wheatcroft and Price, 2012; Walther et al., 2016; Logan et al., 2018; Gentzkow, Shapiro and Taddy, 2019).[1] Thus our treatment below.

We explain the problem and show that statistical bias can be large and consequential, especially in comparisons where one group-wise distance has been estimated with greater uncertainty than another distance. This might be due to different sample sizes. Less intuitively, it might be due to sample sizes that are *imbalanced* across comparisons: e.g. a (very imbalanced) majority v minority group vector distance versus a (balanced) 50:50 group distance. This is in contrast with bias in (balanced) pairwise distances of non-averaged vectors (Mozer et al., 2020; Rossiter, 2022; Kraft and Klemmensen, 2023), where a) researchers *intend* for distances to capture levels of measurement error (and not a document's expected value) or b) such bias may have more limited effects on later inferences – since researchers *might* plausibly assume more or less equal measurement error across studied pairs.

We derive an estimator that does not suffer from this bias, and we show that it performs well in a variety of settings. Finally, we provide solutions to practical issues that arise in the embedding regression setting (e.g. statistical testing and inference) and incorporate the solution into the `conText` package in `R`.

[1] In practice, distances are typically not corrected for statistical bias (see, for example, uncorrected estimates in Rodriguez, Spirling and Stewart 2023). When bias is addressed, approaches include using cross-validation (Walther et al., 2016) (an approach not yet used in political science – that is functionally equivalent to our simple correction, but difficult to extend to complex designs) or using the mean of pairwise estimates and then comparing those means of a kind of permutation distribution (Kraft and Klemmensen, 2023) (without correcting bias). Our work is distinct from research that identifies and attempts to correct for term-frequency bias in word embedding association tests (e.g. van Loon et al., 2022; Kindel, 2023) (due to biases in position); however, the form of bias we correct here *might* also contribute to observed frequency-related biases.

2 Why Uncertainty in Position Leads to Bias

To fix ideas, suppose one had a relatively short vector—of length 2—representing a word embedding. For the word "immigration" (say) that embedding is *estimated* (e.g. because it is based on a sample of speech) to be $\widehat{v_D} = (-0.1, 1.2)$ for Democrats in Congress. The inferential task is to compare it with some other embedding vector which, for now, we will assume is *known* (not estimated) and is $v_R = (0, 1)$ (the noiseless, expected value of the "immigration" embedding for Republicans in Congress). The Euclidean distance between $\widehat{v_D}$ and v_R is 0.224. The cosine similarity between them is 0.997. We might ask: is the (true) Euclidean distance plausibly zero, and is the cosine similarity plausibly 1?

To see the problem, suppose our estimate of v_D is noisy. This might be because we do not have many Democrats in our sample, and thus there is more uncertainty over each (averaged) element in the $\widehat{v_D}$ vector. If we increase the noise in the estimated $\widehat{v_D}$ when there is *no* (true) difference between v_D and v_R, we always move it *further* from $(0, 1)$. But this bias also applies *in expectation* when there *is* some (true) difference between v_D and v_R. For instance suppose that across samples or noisy measurements, the values that we estimate for v_D (the unobserved sampling distribution of $\widehat{v_D}$) are sometimes greater than corresponding elements in the v_R vector and sometimes less. Then our element distances are nonetheless always positive; thus in expectation, the estimated distance is greater than the true distance. We illustrate this effect in SI Figure C.1, and also there expand our bias explanation.

To characterize the bias more fully and precisely, consider measuring the Euclidean distance between two (estimated) length-K vectors $\hat{\theta}$ and $\hat{\phi}$—of which our word embeddings vectors above were just specific examples. This is, by definition, the L_2 norm of the difference between those vectors, $\|\hat{\theta}-\hat{\phi}\|_2 = \sqrt{\sum_{k=1}^{K}(\hat{\theta}_k - \hat{\phi}_k)^2}$. Now, for presentation reasons (though, as noted below, this will also be our preferred norm), suppose we square that norm. That is, we are working with $\|\hat{\theta} - \hat{\phi}\|_2^2 = \sum_{k=1}^{K}(\hat{\theta}_k - \hat{\phi}_k)^2$. Taking expectations on both sides we have

$$E\left[\|\hat{\theta} - \hat{\phi}\|_2^2\right] = E\left[\sum_{k=1}^{K}(\hat{\theta}_k - \hat{\phi}_k)^2\right] \tag{1}$$

$$= \sum_{k=1}^{K} E[(\hat{\theta}_k - \hat{\phi}_k)]^2 + V[\hat{\theta}_k - \hat{\phi}_k] \tag{2}$$

$$= \|\theta - \phi\|_2^2 + \underbrace{\sum_{k=1}^{K} V[\hat{\theta}_k - \hat{\phi}_k]}_{\text{Bias}} \tag{3}$$

where line 2 follows because $E[X^2] = E[X]^2 + V[X]$ for a random variable X. Importantly, variance here is the variance of the (unobserved) distribution of the estimator (i.e., the squared standard error). The point is that the bias (for the squared norm) is the (sum of) the variances of the differences between the vectors' elements. And those variances result from uncertainty in estimation of ϕ and θ. Only if the elements of those vectors are estimated without error is there no bias. To be clear, there is no obligation to use the *squared* Euclidean norm—one can use the unsquared version (as in Rodriguez, Spirling and Stewart 2023, where this version is used but uncorrected), cosine similarity, or some other metric. But some version of the bias will remain—and the form of the bias is not straightforward to write down or fully correct (see SI Sections C.2 and C.2.1).

Linking this back to our initial motivating example, we need to subtract the variances (i.e., the squared standard errors) of $\widehat{v_{D1}}$–v_{R1} and $\widehat{v_{D2}}$–v_{R2} from 0.224^2 $((\widehat{v_{D1}}\text{–}v_{R1})^2 + (\widehat{v_{D2}}\text{–}v_{R2})^2)$ for an unbiased estimate of the squared Euclidean distance between v_D and v_R.

3 Why This Matters, Even in Large Samples

Inspecting Equation (3), when could we realistically expect the *absolute* bias to be small or zero? It is when we have a very large amount of data such that our estimates of the (elements of the) vectors are close to their true population values. And, for descriptive *relative* comparisons such as "the difference between Democrat women and men on this issue is larger than for Republican women and men" we must describe the distances correctly if measurement or sampling error may be unequal across these comparisons. But to clarify, the issue is *not* that small samples make claims about whether one vector is statistically significantly different to another harder to assess; they do, but that is a separate matter. The issue is that the claimed (point) difference between the vectors is reported as being larger than it really is irrespective of hypothesis test concerns.

Of course, it is hard to know in advance whether one has "enough" data or not. However, the problem will remain *in absolute terms* when the dimension of the vector is high relative to the uncertainty. And, below, we show that it can remain *in relative terms* when two compared group-wise differences are relatively imbalanced.

To illustrate absolute versus relative bias, we use a 10% sample of the Twitter voter panel described in Hughes et al. (2021) and compare random groups (where the true difference must be zero by construction) of varying relative sizes and for the same overall sample size —for example, in more practical terms, a

50-50 party affiliation comparison based on 1000 total observations versus a 90-10 majority-minority racial group comparison based on 1000 total observations. Specifically, we compare these groups by estimating embeddings of the word `children` (derived by the methods in Rodriguez, Spirling and Stewart, 2023) for them, restricting our analysis to a single tweet each from a random sample of users and only tweets containing a single use of the word `children`—we address more complex designs later in this paper. We then calculate the squared Euclidean norm of the difference between (the average) embedding vector of `children` for the 50-50 comparison (e.g., gender or U.S. political party) versus 90-10 (e.g., majority-minority race or religion). In this, increasing group imbalance leads to increased estimate uncertainty.

Figure 1 shows the squared Euclidean norm (with jackknifed confidence interval) for both the balanced and imbalanced random groups. While the size of the estimated difference—note the y-axis scale is much smaller as we move across the page—is decreasing as the sample size grows (from 200 to 400 and upwards to 10000 instantiations of the term), the difference between the balanced and unbalanced case remains essentially the same in relative terms. This issue matters because the absolute size of the norm often has no clear interpretation and scholars use the relative scale of the differences to benchmark their results. Simply put, for practical reasons of interpretation, large sample sizes will not fix this issue.

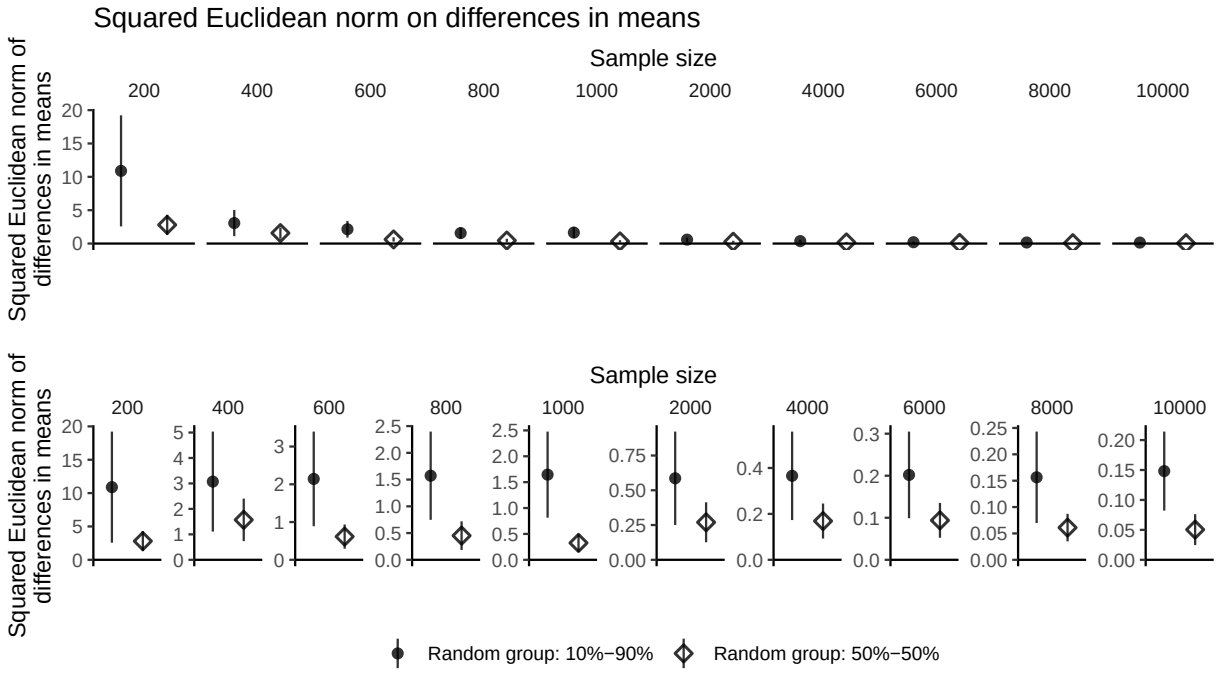

Figure 1: Smaller sample sizes and larger group imbalance both lead to increased estimate uncertainty, and so artificially inflate distance estimates. This can exaggerate majority-minority group differences relative to equally sized group differences.

4 A Correction Based on Variance Estimation

Below, we derive a correction for both simple (a two-group comparison) and complex designs (e.g., non-independent observations and controls). Our applied setting is embedding regression though it will work in other settings too.

Consider the generic case where the model on our latent vectors is parameterized by a vector β. In the context of two groups' embeddings this would just be the difference between their average vectors ($\beta = \theta - \phi$) but we can think of it as a more general regression parameter. If our summary of this vector is the squared Euclidean norm, this would lead to the estimator,

$$\widehat{\|\beta\|_2^2} = \sum_{k=1}^{K} (\hat{\beta_k}^2 - \hat{V}[\hat{\beta}_k]) \tag{4}$$

which is unbiased given an unbiased estimator of the variance $\hat{V}[\cdot]$ (and an unbiased estimator for β).[2]

For the (non-squared) Euclidean norm, because the corrected value can be negative, we can take the square root of the absolute value $\widehat{\|\beta\|_2^2}$ and then apply the sign of the estimate. Let that sign of $\widehat{\|\beta\|_2^2}$ —i.e. literally whether the quantity $\sum_{k=1}^{K} \hat{\beta_k}^2 - \sum_{k=1}^{K} \hat{V}[\hat{\beta}_k]$ is positive or negative—be denoted as sgn. Then we have the estimator, $\widehat{\|\beta\|} = \text{sgn}\sqrt{\text{abs}\left(\sum_{k=1}^{K} \hat{\beta_k}^2 - \sum_{k=1}^{K} \hat{V}[\hat{\beta}_k]\right)}$. However, this estimator is no longer unbiased, as we will show. And, unlike the squared version, it is strongly bimodal. We illustrate this bimodality and potential interpretation problems in SI Figures C.3 and C.4. Nonetheless, it does provide an estimate closer to a null of no difference, and with bias far smaller than for the uncorrected norm. Further, in this form, the quantity can be used to (mostly) correct bias that arises in cosine similarity measures (which itself arises due to bias in the Euclidean distance denominator of those measures; see SI Section C.7). While we think it is reasonable for some researchers to prefer the ordinary Euclidean distance (and cosine similarity), authors who use these corrected distance measures should be careful to fully explain to readers their bimodal distributions and correspondingly skewed confidence intervals around 0.

These estimators rely on a way of estimating the variance of the estimator for β. For the simplest case, for comparing embeddings of dimensions K, one can run K separate (linear) regressions, each with n observations corresponding to the number of instantiations of the term in question. One then has immediate access to the relevant $\hat{\beta}$s and the standard errors (and thus variances) of the same. We illustrate this debiasing using `R`'s `lm` function (`R`'s main linear regression function) in SI section A, where the de-biasing

[2]This can result in a negative estimate if the value of the variance is large enough. That case should be substantively interpreted as implying "no difference" between the vectors.

step is simply `estimate`^{2} `- std.error`^{2}.

To demonstrate the efficacy of the correction, using variances estimated from the K linear regressions method above, we simulate differences in means by sampling vectors from a multivariate normal and comparing the corrected and uncorrected norms for different group imbalances (50%-50% or 10%-90%). In this, we sample $k = 50$ 'embedding' dimensions, for varying sample sizes. Across these 50 dimensions, the locations of the two groups are offset by a value of $\pm c$, where c is a small or large number, on half (25) of the dimensions. The variance of each dimension is selected by a random draw from a non-central (specifically $\lambda = 1$) χ^2 distribution with one degree of freedom, meaning that variance is not equal across dimensions. Figure 2 displays the uncorrected plugin estimator for the squared Euclidean norm. It shows that whatever the group imbalance, the corrected estimator is unbiased: on average, it recovers the true distance.

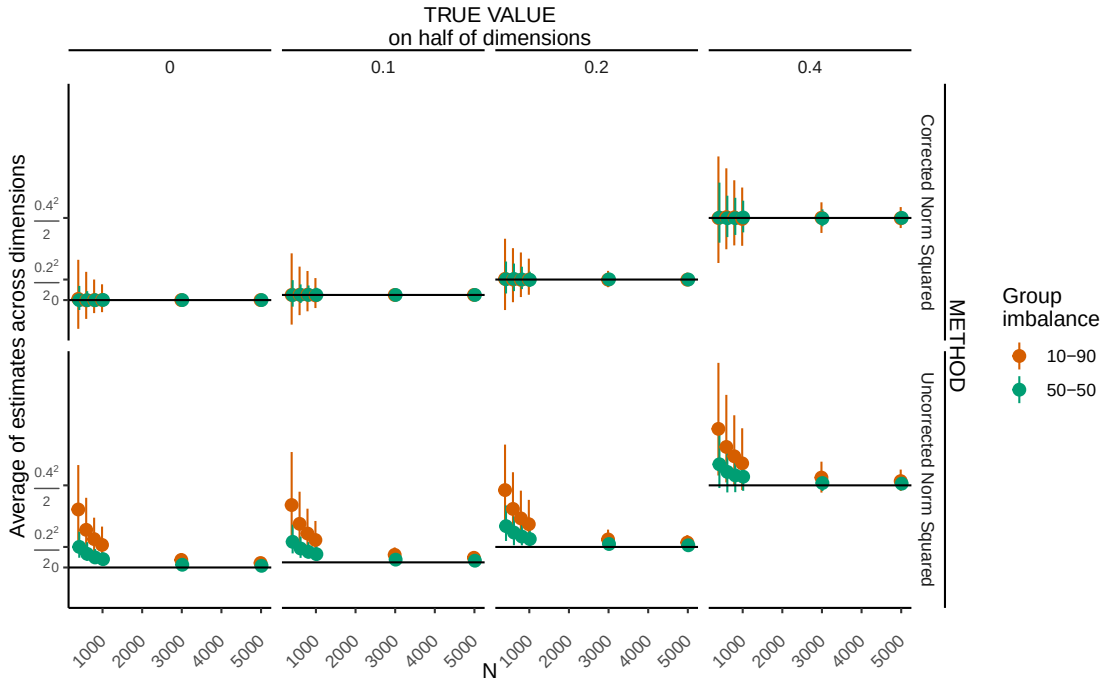

Figure 2: This figure shows simulation results for the squared Euclidean norm divided by the number of dimensions (i.e., the square of the true β's). The horizontal black lines represent the *true* Euclidean norm squared, divided by the number of dimensions (50). Points represent the averages of the simulations and intervals are the 2.5% to 97.5% quantiles of the sampling distributions. Small sample size and greater group imbalance increase estimation uncertainty (i.e., the standard error/variance of $\hat{\beta}$). The effect of greater k can be determined by multiplying the y axis scale by k.

Figure 2 in the SI is the same analysis for the non-squared (i.e. the usual) Euclidean norm. This corrected estimator has a small negative bias, but it still substantially outperforms the plug-in estimator.

Clustering and complex sampling designs

In some cases, such as embedding regression, observations are not independent. This can cause a naive estimator to underestimate the variance of the difference *and* permutation tests to be inaccurate.[3] In the SI, we provide and validate through simulation straightforward solutions to these problems, demonstrating that a) we can use sandwich-style standard errors to estimate the variance under clustering and other complex designs when computing our debiased estimator, b) clustered permutation appropriately controls Type I error under clustering (see SI Section C.6, and C.6.2), and c) residual permutation controls Type I error with control variables (see SI Sections C.6 and C.6.2).

We illustrate performance under a realistic setting with clustering and covariates in Figure 3. For this, we use a Twitter data set linked to voter records and voter demographics, described in Hughes et al. (2021). In a 10% sample (approximately 150,000 users out of 1.5 million), we analyzed tweets between January 2019 and February 2023 that contained the word "people" (specifically 72,389 users, who posted 5.5 million tweets containing 'people'). To simulate sampling distributions with this 'population', we then created sub-samples of users for varying sample sizes (n), taking 500 samples for each n. For these sub-samples (and on the full sample/'population'), we ran embedding regression with covariates—party (Republican or Democrat), age group, gender, and race—and with each user weighted equally in the models (rather than by tweet frequency). In the figure, the points indicate the mean of the squared Euclidean norm for party, with intervals for the 2.5% to 97.5% quantiles of the sampling distribution.

Here, in real data and with many repeat documents for authors, the correction accurately measures distance for small sample sizes. However, we also see a wide sampling distribution, suggesting that relatively large sample sizes and/or large effect sizes are likely to be needed to reliably measure differences between groups.

Constructing confidence intervals is difficult

Despite the feasibility of the debiased estimator and the accuracy of related permutation tests (i.e., accurate calculations of the distribution of a null), the construction of confidence intervals with proper coverage at all values of the latent distance—and for complex designs in particular—is challenging. While the natural impulse is to use resampling methods with the debiased estimator, this is unfortunately an understood failure case for the bootstrap (Dodd and Korn, 2007), which we illustrate in SI Section C.4. Bootstrapped confidence intervals tend to contain more than nominal coverage, meaning that a 95% CI will have greater

[3] Permutation tests are unaffected by estimate corrections in simple designs.

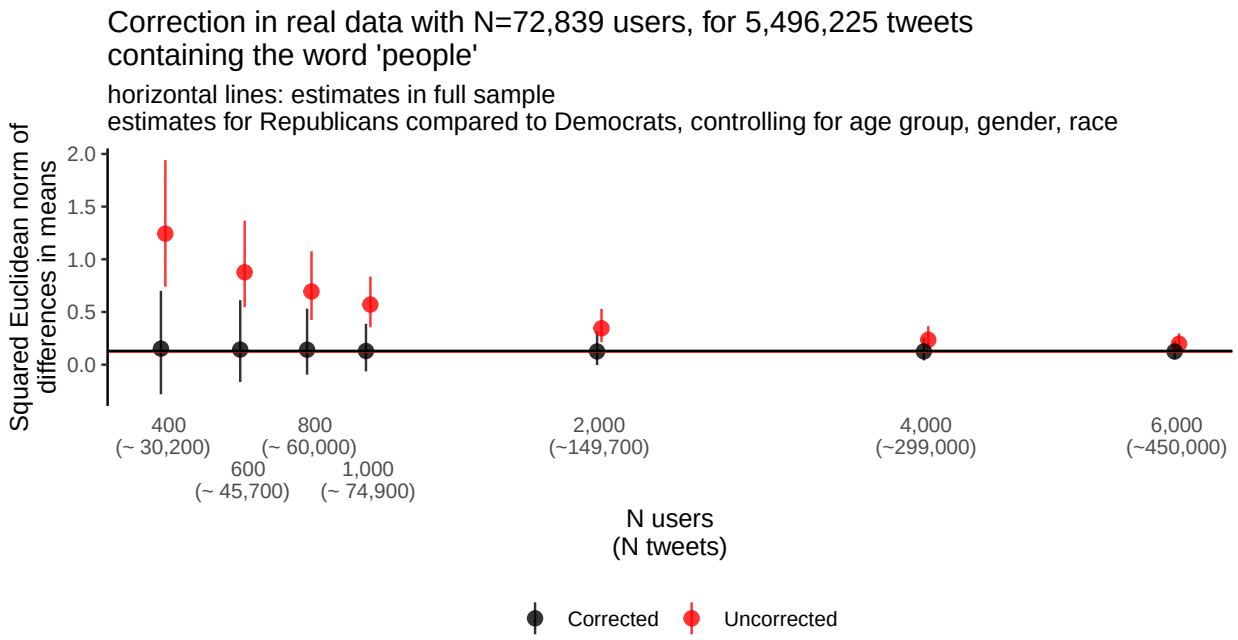

Figure 3: Estimator performance on sub-samples of Twitter data set.

than 95% coverage, especially for small effect sizes. We instead recommend the jackknife which outperforms the bootstrap (see, again, SI Section C.4), but does still over-cover when the true distance is small.

5 Discussion

When social scientists compute distance between vectors the risk of bias is considerable. We studied this bias and suggested ways to mitigate it. These methods work in real and simulated settings, and will be implemented in free statistical software.

6 Data availability statement

All code to reproduce the simulations and data analyses in this paper will be made publicly available in our replication materials and posted to an online repository (e.g., `osf.io`). The social media data is publicly viewable but, due to new API access restrictions, the text of shareable tweet ID's can no longer be downloaded in bulk through Twitter's academic API. We will be unable to share raw text data to reproduce analyses, except as aggregated model output.

7 Competing Interests

The authors have no competing interests to report.

8 Research with Human Subjects

Analysis of Twitter data linked to voter records was approved by the Cornell University Institutional Review Board (IRB #143475, exempt).

References

Dodd, Lori E and Edward L Korn. 2007. "The Bootstrap Variance of the Square of a Sample Mean." *The American Statistician* 61(2):127–131.

Gentzkow, Matthew, Jesse M. Shapiro and Matt Taddy. 2019. "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech." *Econometrica* 87(4):1307–1340.

Hughes, Adam G, Stefan D McCabe, William R Hobbs, Emma Remy, Sono Shah and David M J Lazer. 2021. "Using Administrative Records and Survey Data to Construct Samples of Tweeters and Tweets." *Public Opinion Quarterly* 85(S1):323–346.

Kindel, Alexander T. 2023. "Geometrically consistent estimation of multidimensional word associations in text corpora.".

Kraft, Patrick W. and Robert Klemmensen. 2023. "Lexical Ambiguity in Political Rhetoric: Why Morality Doesn't Fit in a Bag of Words." *British Journal of Political Science* pp. 1–19.

Logan, John R, Andrew Foster, Jun Ke and Fan Li. 2018. "The uptick in income segregation: Real trend or random sampling variation?" *American Journal of Sociology* 124(1):185–222.

Mozer, Reagan, Luke Miratrix, Aaron Russell Kaufman and L. Jason Anastasopoulos. 2020. "Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality." *Political Analysis* 28(4):445–468.

Nyarko, Julian and Sarath Sanga. 2022. "A Statistical Test for Legal Interpretation: Theory and Applications." *The Journal of Law, Economics, and Organization* 38(2):539–569.

Rodriguez, Pedro L., Arthur Spirling and Brandon M. Stewart. 2023. "Embedding Regression: Models for Context-Specific Description and Inference." *American Political Science Review* pp. 1–20.

Rossiter, Erin L. 2022. "Measuring Agenda Setting in Interactive Political Communication." *American Journal of Political Science* 66(2):337–351.

van Loon, Austin, Salvatore Giorgi, Robb Willer and Johannes Eichstaedt. 2022. Negative Associations in Word Embeddings Predict Anti-black Bias across Regions–but Only via Name Frequency. In *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 16 pp. 1419–1424.

Walther, Alexander, Hamed Nili, Naveed Ejaz, Arjen Alink, Nikolaus Kriegeskorte and Jörn Diedrichsen. 2016. "Reliability of Dissimilarity Measures for Multi-Voxel Pattern Analysis." *NeuroImage* 137:188–200.

Weir, Jason T., David J. Wheatcroft and Trevor D. Price. 2012. "The Role of Ecological Constraint in Driving the Evolution of Avian Song Frequency Across a Latitudinal Gradient: Evolution of Birdsong." *Evolution* 66(9):2773–2783.

Supporting information for:

"Measuring Distances in High Dimensional Spaces

Why Average Group Vector Comparisons Exhibit Bias, And What to Do About it"

Table of Contents

A Illustration of correction with R code

```
library(MASS)
library(tidyverse)
library(broom)
```

Example debiasing function

This is a simplified illustration of our debiasing method. The `conText` package will be implemented more efficiently and robustly. **The debiasing step in the function below is highlighted with ####.**

```
debiased_estimates <- function(
    mod # e.g., model output from lm() -- R's main linear regression function
    # for independent observations
    # or the estimatr package's lm_robust() to calculate clustered standard errors
    # for non-independent observations
) {
  mod_df <- tidy(mod) # convert model summary to data frame
  #
  # (unbiased) beta hats to (biased) squared beta hats
  mod_df$biased_sqrd_beta <- mod_df$estimate^2
  #
  mod_df$beta_variance <- mod_df$std.error^2 # beta standard error to beta variance
  #
  ##### This is debiasing step: ####
  # subtract estimated beta variance from squared beta hats
  mod_df$debiased_sqrd_beta <- mod_df$biased_sqrd_beta - mod_df$beta_variance
  #
  return(mod_df)
}
```

Simulate data

In this simulation, the true value (i.e., the expected value) of the squared Euclidean norm on the difference between embeddings vectors is 0 – because the groups have been randomly assigned. We demonstrate that the debiased estimator returns 0 in the section "Average of 10,000 estimates" below.

```
simulate_data_k2 <- function(n = 500, group_imbalance = c(0.9, 0.1)) {
  list(
    # a (random) dummy variable for group membership
    random_groups = sample(c(0, 1), size = n, replace = TRUE, prob = group_imbalance),
    # these k=2 "embeddings" are just the example in the help file of mvrnorm()
    embeddings =  mvrnorm(n = n, mu = rep(0, 2), matrix(c(10,3,3,2),2,2))
  )
}
```

Single dimension illustration

```
set.seed(987654321)

simulated_data <- simulate_data_k2()
```

```r
embeddings <- simulated_data[["embeddings"]]
random_groups <- simulated_data[["random_groups"]]

mod_d1 <- lm(
  # run a regression with the group indicator as x and the first embedding dimension as y
  embeddings[,1] ~ random_groups
)
```

```r
debiased_estimates(mod_d1) |>
  select(term, biased_sqrd_beta, beta_variance, debiased_sqrd_beta) |>
  filter(term != "(Intercept)")
# we remove the intercept estimate for this illustration but it can be used in
# intercept only models, e.g., y ~ 1 and one of these intercept regressions for each
# compared group, to correct the denominator of a cosine similarity calculation
```

term	biased_sqrd_beta	beta_variance	debiased_sqrd_beta
random_groups	0.24	0.21	0.03

Multiple dimension illustration

```r
mod_d2 <- lm( # run a separate regression with the second embedding dimension as y
  embeddings[,2] ~ random_groups
)

all_debiased_sqrd_betas <- bind_rows(
  # stack estimates from models 1 and 2 for the squared Euclidean norm below
  debiased_estimates(mod_d1),
  debiased_estimates(mod_d2)
) |>
  filter(term != "(Intercept)")
```

```r
all_debiased_sqrd_betas |>
  # calculate the squared Euclidean norm for each x variable (here, only 1 of them)
  group_by(term) |>
  summarize(
    biased_sqrd_euclidean_norm = sum(biased_sqrd_beta),
    debiased_sqrd_euclidean_norm = sum(debiased_sqrd_beta)
  )
```

term	biased_sqrd_euclidean_norm	debiased_sqrd_euclidean_norm
random_groups	0.28	0.02

Average of 10,000 estimates

Repeating the above code for 10,000 simulated samples, mean estimates are:

term	biased_sqrd_euclidean_norm.mean	debiased_sqrd_euclidean_norm.mean
random_groups	0.27	0

B Twitter data tests

For the Twitter data tests, we use data from a panel of Twitter users, described in (Hughes et al., 2021). Users in this panel were linked to voter records, which included basic demographic information and vote histories. We down-sample the large panel to only users whose user IDs ended with eight, and analyzed tweets between January 2019 and February 2023 that contained the words "children" (illustration of bias in Figure 1), "people" (large sample illustration of bias and correction in Figure 3), or "racism" (Tables C.8 and C.9; Figures C.12 and C.13 – assessing correction for plausibly larger main effects and covariate effects).

Context-dependent word embeddings are drawn from the a la carte embedding approach described in (Rodriguez, Spirling and Stewart, 2023). This approach assigns context-dependent word embeddings (the 200d Twitter embeddings from GloVe (Pennington, Socher and Manning, 2014): https://nlp.stanford.edu/projects/glove/) for the word 'people' based on the words that appear near the word people in each tweet. Our analyses study the squared Euclidean norm of distances across groups for these embeddings.

C Supplementary figures and tables

C.1 Illustration of bias from folding

If, across samples or noisy measurements, the values that we estimate for β (the unobserved sampling distribution of $\hat{\beta}$) are sometimes greater than their true values and sometimes less than, our distances are nonetheless always positive – and so, in expectation, greater than the true value of β.

We illustrate this folding effect in Figure C.1.

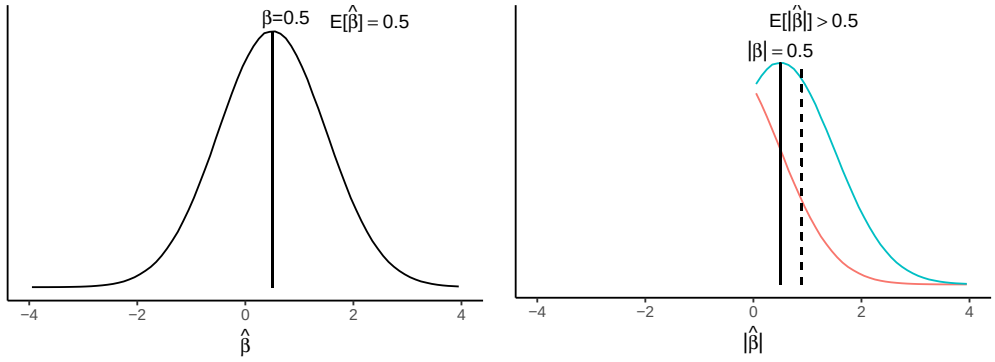

Figure C.1: Illustration of bias from folding. In the right panel, negative values (in red) for $\hat{\beta}$ become positive after squaring and then taking the square root (equivalent to $|\hat{\beta}|$ in this uni-dimensional illustration).

Less intuitively, *squared* Euclidean distance estimates are biased even when the sampling (or measurement error) distribution (unrealistically) never spans 0. If we split a positively or negatively bounded (and non-constant) distribution in exactly half at its expected value, the expected value of the half further from 0 will increase *more* (or, if originally less than 1, decrease less) after squaring than the expected value of the half closer to it will.

C.2 (Unsquared) corrected Euclidean distance

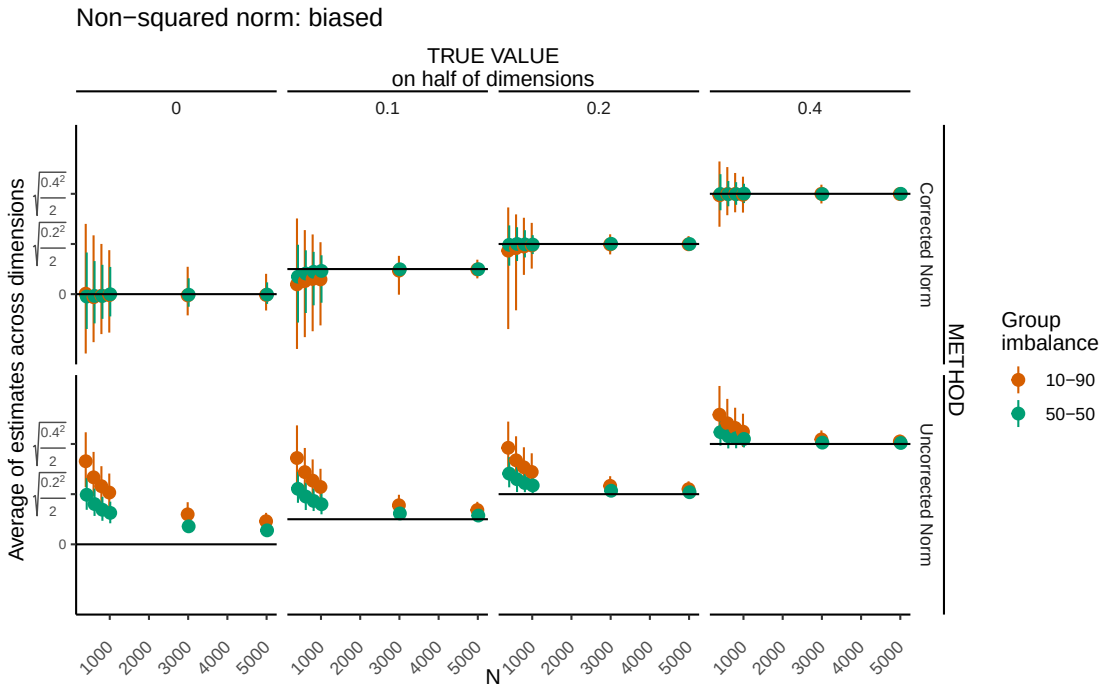

Figure C.2: This figure shows simulation results for the ordinary (unsquared) Euclidean norm. The horizontal black lines represent the true Euclidean norm, divided by the number of dimensions (50). Points represent average of the simulations and intervals are the 2.5% to 97.5% quantiles of the sampling distribution.

C.2.1 (Unsquared) Euclidean distance bias

An expression of the bias for the Euclidean norm must, to our knowledge, be distribution dependent. For example, for the case of $k = 1$, we can use the properties of the half normal distribution (for a $\hat{\beta}$ that is normally distributed for large N, by the central limit theorem) to get an expression for the expected value of the absolute value of $\hat{\beta}$: $E[|\hat{\beta}|] = \sigma\sqrt{\frac{2}{\pi}}e^{\frac{-\beta^2}{2\sigma^2}} + \beta\text{erf}\left(\frac{\beta}{\sqrt{2\sigma^2}}\right)$, where erf indicates the error function and σ the standard deviation of $\hat{\beta}$ (i.e., the standard error). $E[\hat{\beta}] = \beta$. This reduces to $E[|\hat{\beta}|] = \sigma\sqrt{\frac{2}{\pi}}$ when $\beta = 0$.

C.3 Bimodality of corrected Euclidean distance

Unlike the squared Euclidean distance estimator, the ordinary Euclidean distance estimator is strongly bimodal. We suspect the bimodality in particular may make this estimator somewhat difficult for many readers to interpret. We illustrate this bimodality and potential interpretation problem in Figures C.3 and C.4, where we show the *same* estimates with and without squaring. In the squared version, we think that the distribution resembles what an average reader would expect to see for estimates of no difference. In the unsquared version, some readers may interpret estimates further from 0 as being more distinct from 0 than they really are – they are far from 0 only because the distribution of this estimator has low density close to zero.

Given this, and while we think it is reasonable to prefer the ordinary Euclidean distance, authors who use this corrected distance measure may need to be careful to fully explain its bimodal distribution – and the reason for heavily skewed confidence intervals.

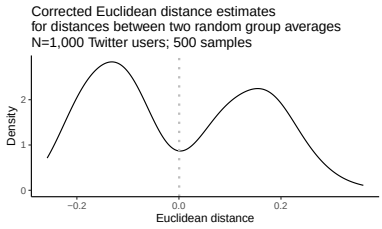

Figure C.3: Distribution of corrected Euclidean distance estimates for N=1,000 across 500 samples from Twitter data for term 'people'.

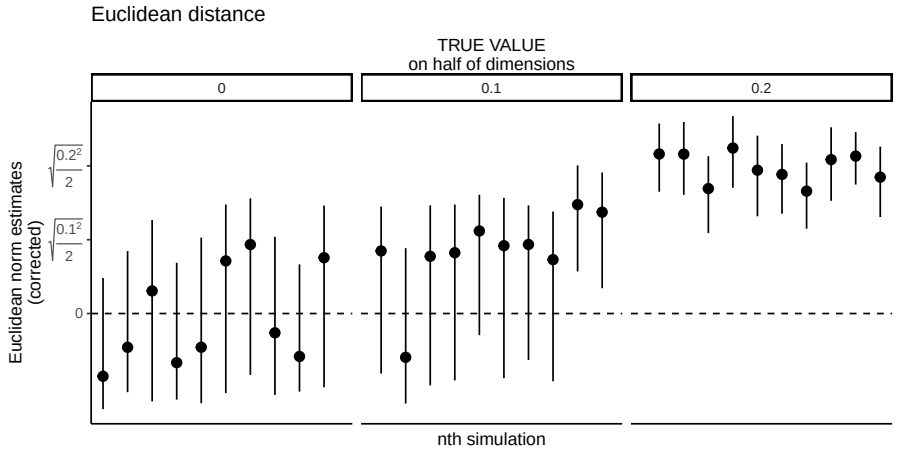

Figure C.4: 10 corrected Euclidean distance estimates for N=1,000, equal group comparisons, and different effect sizes – from simulations shown in Figure C.2.

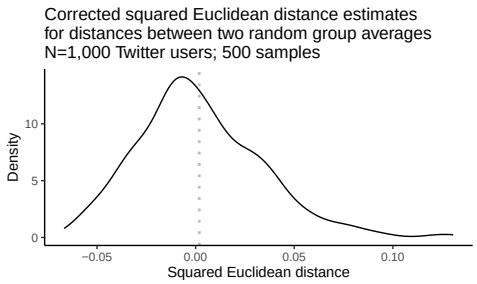

Figure C.5: Distribution of corrected *squared* Euclidean distance estimates for N=1,000 across 500 samples from Twitter data for term 'people'.

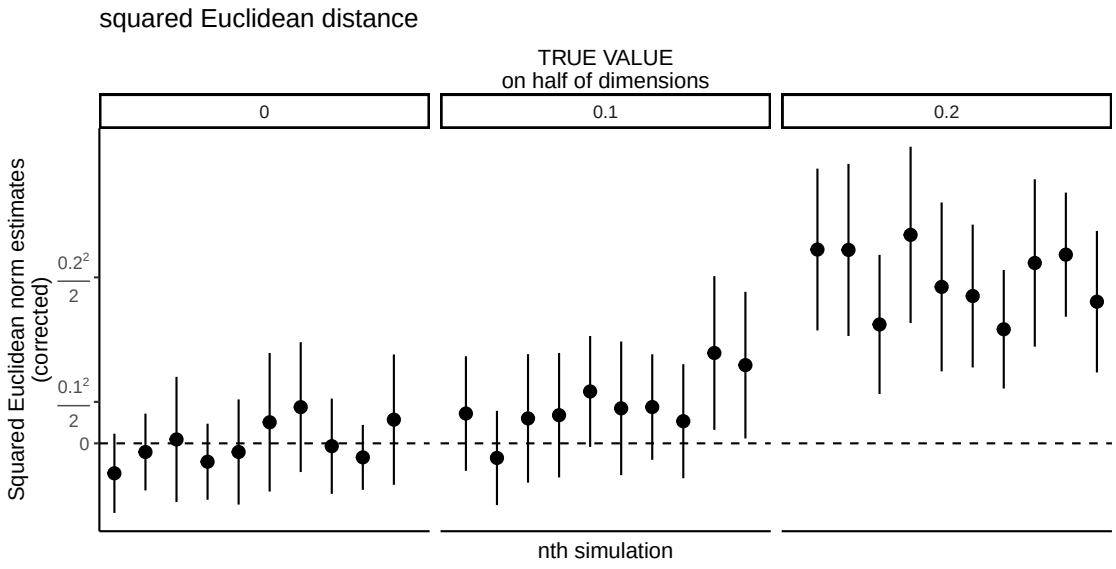

Figure C.6: 10 corrected squared Euclidean distance estimates for N=1,000, equal group comparisons, and different effect sizes – from simulations shown in Figure 2.

C.4 Bootstrapping: challenges, coverage of confidence intervals

We assess whether bootstrapping and/or the jackknife can be used to construct confidence intervals for the squared Euclidean norm.

For the bootstrap, we calculate the coverage of a bootstrapped confidence interval with 500 replicates for our main simulations described in the main text (see Figure 2) for the case of N=1,000. Meaning, we calculate the fraction of (true) squared Euclidean norms that fall within the range of 2.5% to 97.5% quantiles of the bootstrap distribution (after subtracting double the calculated variance of each estimate – since, as we show in Figure C.7, the mean of the bootstrap distribution is biased by double the variance).

For the jackknife, we use the leave-one-out method to construct standard errors and confidence intervals.

These results are shown in Figures C.8 and C.9. For effect size values less than around 0.5, "95%" confidence intervals contain more than 95% of the true/assigned effect size. The jackknife appears to have closer to nominal coverage because for effect sizes less than 0.1 it has coverage of around 98% for a "95%" confidence interval while the bootstrap is around 100%.

In Figure C.10, we show similar coverage for the method in Hyodo, Watanabe and Seo (2018).

We also test the jackknife using Congressional Record data from Sessions 111-114 (Gentzkow, Shapiro and Taddy, 2018). To do this, we select target words with varying degrees of gender and partisan differences and obtain locally trained embeddings with context window size six. We fit an embedding regression with party or gender as a covariate and define the (non-deflated) squared Euclidean norm of the coefficients as the true parameter. We simulate sampling distributions from this 'population' of embeddings by taking sub-samples of varying sizes ($n = 100, n = 500, n = 1000$) and estimate the same regression, using the jackknife to calculate confidence intervals. For each target word and sub-sample size, we replicate the simulation process 1000 times and calculate the jackknife coverage as described above. Coverage results for each embedding regression specification are shown in Table C.1. Similar to the coverage we obtain using simulated data, the jackknife has a coverage of around 98% for a "95%" confidence interval for effect sizes close to 0, but has closer to nominal coverage for larger effect sizes.

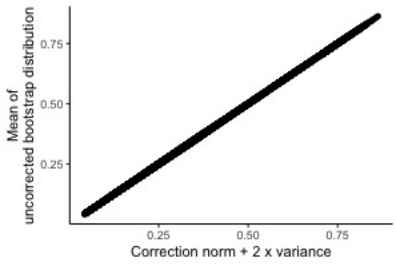

Figure C.7: Bootstrapping doubles the variance bias (from simulations in main text Figure 2).

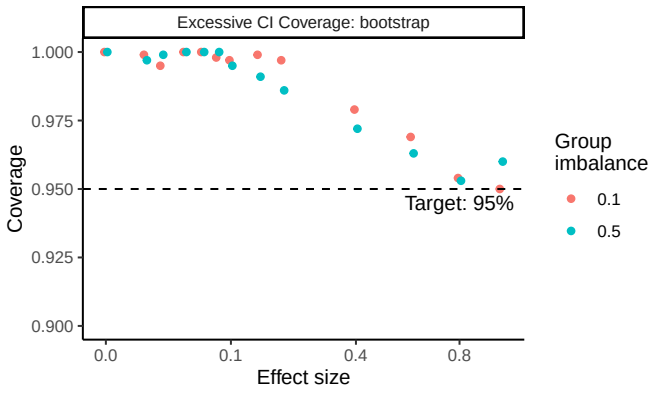

Figure C.8: Coverage of bootstrapped and doubly corrected norm for N=1,000.

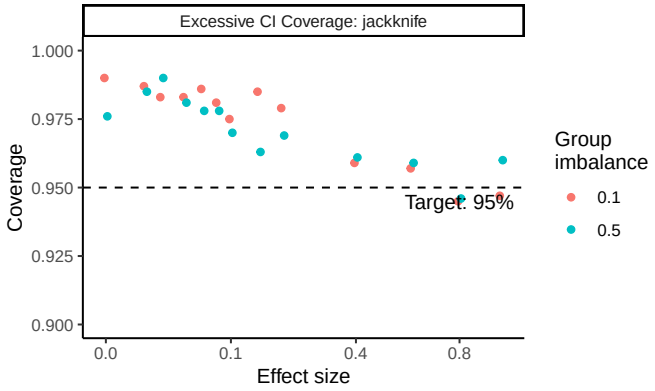

Figure C.9: Coverage of jackknife for N=1,000.

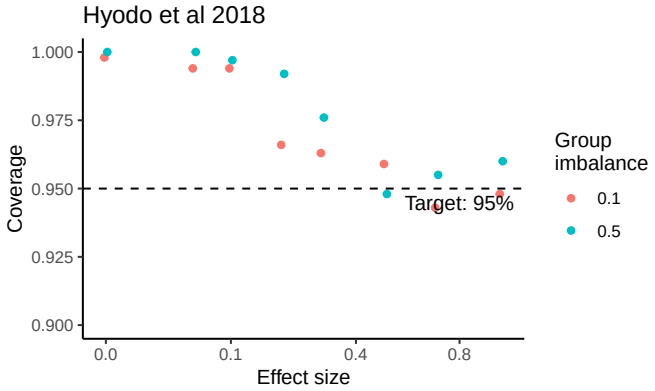

Figure C.10: Coverage of Hyodo, Watanabe and Seo (2018) confidence intervals for N=1,000.

Embedding Regression	Squared Norm Full Sample Estimate	Total observations	Coverage by Sub-sample Size		
			n = 100	n = 500	n = 1000
children ∼ gender	0.62	50,191	0.994	0.984	0.980
nation ∼ party	0.82	49,777	0.989	0.976	0.975
president ∼ party	3.08	220,944	0.965	0.941	0.935
health ∼ party	3.16	133,797	0.978	0.952	0.941
women ∼ gender	5.01	46,802	0.952	0.946	0.958
abortion ∼ party	6.57	6,670	0.982	0.96	0.956
climate ∼ party	6.98	12,641	0.936	0.915	0.943
hispanic ∼ party	7.45	1,565	0.983	0.956	0.955
black ∼ party	11.77	6,945	0.975	0.960	0.957
unemployment ∼ party	12.01	21,398	0.945	0.951	0.944
wage ∼ party	21.75	6,471	0.910	0.942	0.926
gun ∼ party	22.64	10,446	0.956	0.958	0.943
immigrants ∼ party	24.99	4,677	0.928	0.946	0.952

Table C.1: Coverage of jackknife on full Congressional Record data for N=100, N=500, and N=1000.

C.5 Effects of whitening embeddings

Our method corrects bias related to the variance of an estimated $\hat{\beta}$ rather than variance in the data itself. If we equalize variance in the data, like by whitening a matrix and *then* calculating Euclidean distance (i.e., Mahalanobis distance), this re-introduces bias. Intuitively, this introduces bias because (large) differences between groups increase the variance of the data without altering the variance of an estimator. Further, whitening the embeddings of groups separately prior to comparing them would place them into different and incomparable embedding spaces.

In Figure C.11, we re-run our main simulation shown in Figure 2 but whiten the matrix prior to calculating distances. This whitening step equalizes the variance of every embedding dimension and removes covariance across embedding dimensions.

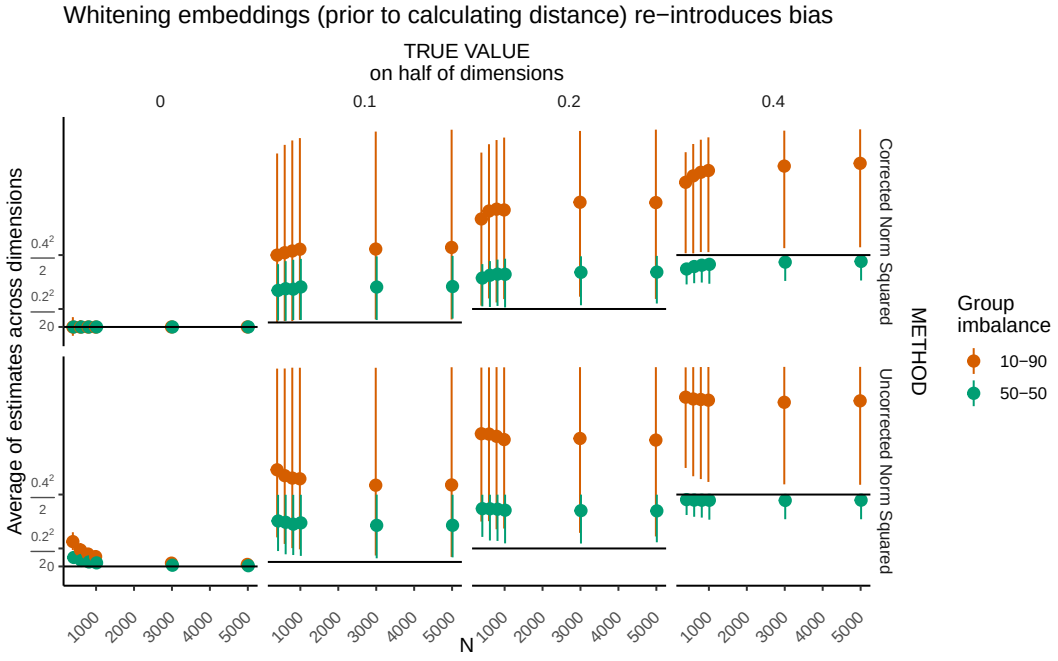

Figure C.11: Estimated squared Euclidean distance on a whitened embedding matrix. In this analysis, the simulated embeddings are whitened prior to calculating squared Euclidean distance. Whitening the embeddings of groups separately prior to comparing them would place them into different and incomparable embedding spaces.

C.6 Covariates and clustering: corrections and simulations

Clustering

If responses are not independent, then we can under-estimate the variance of our $\hat{\beta}$, just as in ordinary linear regression. The solution for this is straightforward – we can cluster our standard errors using standard practices. We demonstrate in Table C.2 that a) not accounting for clustering biases estimates and b) we can fix that bias through the approaches just described. For estimating clustered standard errors, we use the 'estimatr' R package (Blair et al., 2024) and "stata" (CR1) type cluster-robust standard errors.

Further, we must also permute our outcomes at the cluster level to return valid p-values. Without accounting for clustering, we will tend to over-reject the null due to a permutation distribution that is too narrow and that also has a downward bias. These problems and their fix in simulations is are shown in Tables C.4 and C.7.

Expanding our main text derivation to

$$E\left[\|\hat{\theta}-\hat{\phi}\|_2^2\right] = \|\theta-\phi\|_2^2 + \sum_{k=1}^{K} V[\hat{\theta}_k - \hat{\phi}_k] \tag{5}$$

$$= \|\theta-\phi\|_2^2 + \sum_{k=1}^{K} \left(V[\hat{\theta}_k] + V[\hat{\phi}_k] - 2Cov[\hat{\theta}_k, \hat{\phi}_k]\right) \tag{6}$$

our clustered standard error approach on the *difference* corrects for both inaccurately estimated variances *and* the covariance term. This covariance term can be non-zero when, for example, the same author's text embeddings are included in the averages of both compared vectors. We demonstrate the efficacy of this covariance correction in the "non-independent contrast" results in SI Table C.4, where the same errors are included in both compared vectors in a cross-over design.

Multiple regression

If we use a regression approach, like (Rodriguez, Spirling and Stewart, 2023), then we need to account for the possibility that highly predictive variables will reduce the variability of other estimates. In permutation tests that permute our outcomes, we remove the effect of that increased precision (setting all associations to 0 on average) and, if we do have predictive variables, then over-estimate the variance of our $\hat{\beta}$'s.

The primary solution to this issue is a) to use standard errors from the regression rather than using permutation to estimate variance and b) permuting the residuals from our regression rather than the

outcome. We demonstrate that full model residual permutation produces accurate estimates and valid p-values in Tables C.3 and C.6.

Below, we conduct simulations that are the same as those in the main paper, but we restrict our sample size to 1,000, use 1,000 replicates (rather than 500), and also for:

- (maximum) clustering: duplicate each observation (each observation appears twice)
- (strong) covariate: assign a covariate with $c = 10$ (a very large effect size)
- non-independent contrasts (a crossover design): duplicate each observation – but with the duplicate observation in the opposite group as the original

In reporting estimates, we calculate the average estimates using the squared norm before taking the pseudo square root. Meaning, the estimates below are for the unbiased correction – we have only applied a pseudo square root so that we can still see bias (and lack of bias after correction) in the uncorrected squared norm for effect sizes equal to 0.

C.6.1 Simulation estimates

True value	Uncorrected estimate	Subtract regression variances	**Subtract clustered regression variances**
0.00^2	0.09^2	0.06^2	0.01^2
0.71^2	0.71^2	0.71^2	0.71^2

Table C.2: Normed estimates: (maximum) clustering only

True value	Uncorrected estimate	Subtract regression variances	**Subtract clustered regression variances**
0.00^2	0.09^2	0.06^2	0.01^2
0.71^2	0.71^2	0.71^2	0.71^2

Table C.3: Normed estimates: (maximum) clustering and (strong) covariate

	True value	Uncorrected estimate	Subtract regression variances	**Subtract clustered regression variances**
no covariate	0.00^2	0.04^2	-(0.06^2)	-(0.00^2)
strong covariate	0.00^2	0.04^2	-(0.06^2)	0.00^2
no covariate	0.71^2	0.71^2	0.70^2	0.71^2
strong covariate	0.71^2	0.71^2	0.70^2	0.71^2

Table C.4: Normed estimates: (maximum) clustering and non-independent contrasts (i.e., a crossover design)

C.6.2 Simulation p-values

True fraction < 0.05	Permutation test	**Clustered permutation test**	**Clustered residuals permutation test**
0.05	0.73	0.04	0.04

Table C.5: P-values: (maximum) clustering only

True fraction < 0.05	Permutation test	Clustered permutation test	**Clustered residuals permutation test**
0.05	0.00	0.00	0.05

Table C.6: P-values: (maximum) clustering only and (strong) covariate

	True fraction < 0.05	Permutation test	**Clustered permutation test**	**Clustered residuals permutation test**
no covariate	0.05	0.00	0.05	0.05
strong covariate	0.05	0.00	0.06	0.05

Table C.7: P-values: (maximum) clustering only and non-independent contrasts (i.e., a crossover design)

C.6.3 Twitter p-values

We further assessed the performance of the clustered permutations on the Twitter data using the same sampling procedure as used in Figure 3 for the term 'racism'. 'Racism' is less common than the terms 'people' and 'children' and it is also more likely to be strongly associated with covariates (which can affect the performance of permutation tests). In these tests, each tweet is weighted inversely proportional to the number of tweets that a user posted in the sample overall (e.g., each observation of a user who posted twice will receive a weight of $\frac{1}{2}$). Embeddings are permuted at the user level, whether or not a user has posted the same number of tweets, and each tweet is then re-weighted using a user's new number of tweets after permutation.

In that data, clustered permutation appropriately controls type I error and ordinary permutation slightly over-rejects the null. Clustered residual permutation slightly over-rejects, though it over-rejects less than non-clustered residual permutation. Based on this, we suspect that the ordinary permutation test may perform relatively well on most data sets – except for cases where there is substantial duplication in the embedding observations (e.g., many observations drawn from one very short document), which would more closely resemble the extreme correlation across observations considered in the simulations in Section C.6.2.

A Hotelling T^2 test as well as an estimator (Chen and Qin, 2010) for settings where the number of embedding dimensions exceeds the number of observations can also be used for *simple design* significance tests (Chen and Qin, 2010; Hyodo, Watanabe and Seo, 2018), though with potentially restrictive assumptions. We are unaware of any such estimator for complex designs, however, and we see below that it performs poorly with non-independent observations.

	True fraction < 0.05	clustered permutation	non-clustered permutation	Hotelling T^2 test	Hotelling T^2 test (on single tweet per author)
Random group: 1%-99%	0.05	0.05	0.07	0.59	0.17
Random group: 10%-90%	0.05	0.05	0.08	0.94	0.05
Random group: 50%-50%	0.05	0.05	0.09	1.00	0.04

Table C.8: P-values: Twitter sampling (500 samples of 1,000 users) and distances calculated between random groups.

	True fraction < 0.05	clustered residual permutation	non-clustered residual permutation
Random group: 10%-90%	0.05	0.07	0.10
Random group: 50%-50%	0.05	0.06	0.09

Table C.9: P-values: Twitter sampling (500 samples of 1,000 users) and distances calculated between random groups. Controls: age group, race, gender, party.

C.7 Cosine similarity correction

We use the same down-sampling procedure as in Section C.6.3 (1,000 users' uses of the word 'racism' on Twitter) to analyze the performance of a corrected cosine similarity estimator, but without covariates. In this, we corrected the Euclidean distance in the denominator of the cosine similarity calculations (the Euclidean norm of each group's average embedding vector), and left the numerator untouched (assuming independence across the compared groups). Figure C.12 and C.13 display these results. Although there is a small upward bias in the corrected estimator, that bias is far smaller than the uncorrected estimator bias. Note, too, that this figure illustrates that group differences in embeddings surrounding the same terms may tend to be relatively small.

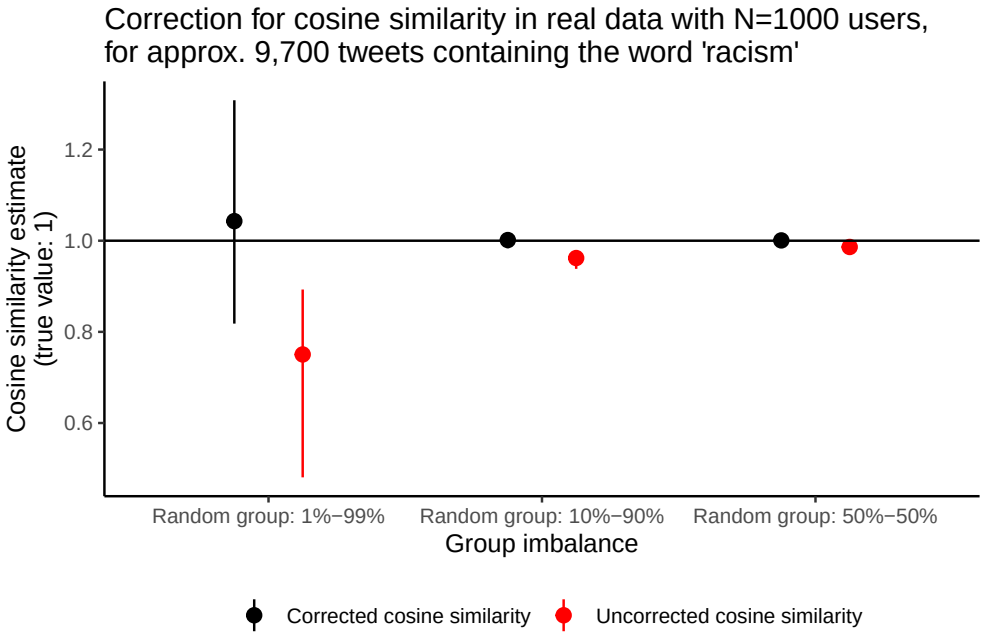

Figure C.12: Cosine similarity estimator performance on sub-samples of Twitter data set: random groups. For this data, we use sandwich-style standard errors to estimate the variance under clustering, to account for clustering at the user level given multiple tweets from the same users.

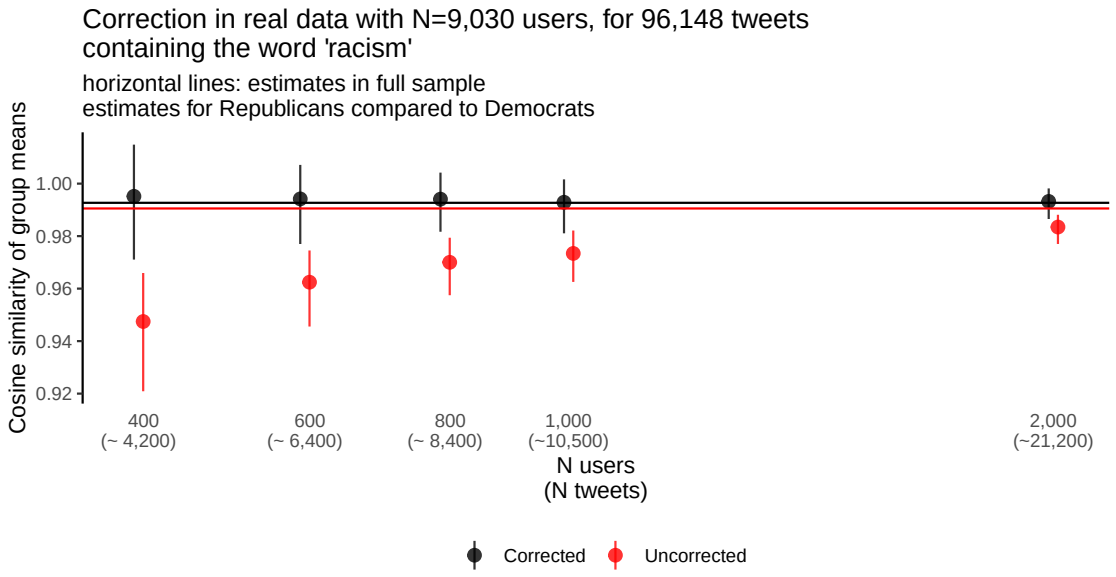

Figure C.13: Cosine similarity estimator performance on sub-samples of Twitter data set: Republican average versus Democrat average. For this data, we use sandwich-style standard errors to estimate the variance under clustering, to account for clustering at the user level given multiple tweets from the same users.

References

Blair, Graeme, Jasper Cooper, Alexander Coppock, Macartan Humphreys and Luke Sonnet. 2024. *Estimatr: Fast Estimators for Design-Based Inference.*

Chen, Song Xi and Ying-Li Qin. 2010. "A Two-Sample Test for High-Dimensional Data with Applications to Gene-Set Testing." *The Annals of Statistics* 38(2).

Gentzkow, Mathew, Jesse Shapiro and Matt Taddy. 2018. "Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts." *Stanford Libraries* .

Hughes, Adam G, Stefan D McCabe, William R Hobbs, Emma Remy, Sono Shah and David M J Lazer. 2021. "Using Administrative Records and Survey Data to Construct Samples of Tweeters and Tweets." *Public Opinion Quarterly* 85(S1):323–346.

Hyodo, Masashi, Hiroki Watanabe and Takashi Seo. 2018. "On Simultaneous Confidence Interval Estimation for the Difference of Paired Mean Vectors in High-Dimensional Settings." *Journal of Multivariate Analysis* 168:160–173.

Pennington, Jeffrey, Richard Socher and Christopher D Manning. 2014. "Glove: Global Vectors for Word Representation." *EMNLP* 14:1532–1543.

Rodriguez, Pedro L., Arthur Spirling and Brandon M. Stewart. 2023. "Embedding Regression: Models for Context-Specific Description and Inference." *American Political Science Review* pp. 1–20.