Geometrically consistent estimation of multidimensional word associations in text corpora

Alexander T. Kindel*

18 September 2023

Abstract

The Word Embedding Association Test (WEAT) is a popular model for measuring word associations in text corpora (e.g., biases, stereotypes, schemas). WEAT-like measurement models aim to estimate the difference in association between two concepts indexed by keyword lists over a set of word embeddings. I show that they do not consistently estimate this quantity. The underlying metric, mean cosine similarity, cannot discern what all of the keywords have in common: in keyword lists with at least four words, the metric does not guarantee that every sub-list containing at least three words is closest in association to itself. For this to be true, the Euclidean distance between at least one of the word pairs would have to be negative. The metric is geometrically inconsistent in the sense that this is impossible. The degree of inconsistency is partially predictable from the conditioning of the cosine similarity matrix. The inconsistency of mean cosine similarity is explained in comparison to a multidimensionally consistent similarity metric.

*Assistant Professor of Sociology, médialab, Sciences Po, Paris, France. Contact: alexander.kindel@sciencespo.fr. I am grateful to B. Stewart for providing extensive guidance over the lifespan of this project. I also wish to thank A. Berg, D. Choi, T. Hansen, J. Lockhart, N. Torres-Echevarry, B. Rohr, and F. Wherry for helpful conversations regarding prior drafts of this paper; and N. Zhou and N. West for introducing me to an aleatoric representation of multidimensional word association problems.

The Word Embedding Association Test (WEAT) is a popular method for measuring word associations in large text corpora using word embeddings (*1*).[1] WEAT is modeled on the Implicit Association Test (IAT) in social psychology (*2*); like the IAT, it is meant to estimate differences in association between concepts encoded in keyword lists. However, WEAT-like measurement models do not consistently estimate this quantity. Given two keyword lists of cardinality k that are subsets of a word vector space ($A, B \subset W$), the similarity metric underlying WEAT—mean cosine similarity (MCS)—is best understood as a certain estimator of the expected *unidimensional* word association: on average, how much does any pair of individual words $\{A_i, B_i\}$ have in common? This is a mathematically and conceptually distinct quantity from *multidimensional* word association: how much do all of the words in A and all of the words in B have in common? This paper explains why these two seemingly similar word association estimands must not be confused.

I demonstrate that the MCS metric is inadmissible as an estimator of multidimensional word association. Specifically, MCS is not geometrically consistent: given any list with more than three words, the metric is not guaranteed to determine that all of the sub-lists are closest in association to themselves, resulting in a geometrically self-contradictory measure. I contrast MCS with a multidimensional similarity metric for word associations with close links to canonical correlation analysis (*3*): the sum of the squared cosines of the principal angles between the subspaces spanned by the two word lists. I refer to this quantity as the canonical subspace metric.

Table 1 compares the performance of the two estimators on the ten WEAT measures of semantic bias presented in the original paper (see Materials and Methods). MCS-based WEAT measures tend to be inflated related to the canonical subspace metric. The MCS estimate for WEAT 8 is closest in magnitude but has the opposite sign. The median error in magnitude is approximately 4.2 times the corresponding canonical subspace metric, but magnitude errors in the other direction are possible. Additionally, the nested difference in means design of the test masks variation in the list-pairwise similarity metrics. MCS correctly estimates the order statistics for the four pairwise comparisons only for one test (WEAT 10).

To explain why MCS is an inconsistent estimator for multidimensional word association problems, this paper compares the two similarity estimators in greater mathematical and conceptual detail. I define what it means for a multidimensional similarity metric to be geometrically inconsistent; explain how canonical correlation analysis yields a consistent metric; and show that mean cosine similarity does not. I illustrate the difference with respect to a comparison between the "male" and "female" words in WEAT 7 and the

[1]Word embeddings are low-rank approximations to bivariate word association measures in text corpora. Embedding algorithms typically factor sparse $n \times n$ matrices of word association statistics into two dense rank p vector spaces W_R, W_C corresponding to the row and column spaces of the underlying measure. In practice, researchers use $W = W_R$ as the word vector space, or sometimes $W = W_R + W_C$ when the underlying measure is assumed to be symmetric. The results in this paper are agnostic to the choice of embedding algorithm.

"pleasant" and "unpleasant" words in WEAT 5 (see Table S1) using publicly available GloVe embeddings of English Wikipedia (*4*–*6*). In the supplemental materials, I provide results for other commonly used pre-trained word embeddings and discuss a few WEAT-like measurement models that are widely used in applied settings.

Minimality and geometric consistency in multidimensional similarity estimation. Many research designs in applied statistical text analysis employ similarity metrics as estimators of word association. A well-known requirement for geometric similarity measurement in this data analysis regime is the *minimality axiom*, which states that any object x is most similar to itself (*7*):

$$\forall x, y \in W : d(x, y) \geq d(x, x) = 0$$

This is really two statements: $d(x, y) \geq d(x, x)$ and $d(x, x) = 0$. The inequality can be generalized to multidimensional comparisons between subsets of W, but the zero-equivalence statement applies only to points. If there were not a non-trivial similarity structure internal to subsets of W, we would not be willing to accept the unidimensional metric in the first place. So, a multidimensional similarity metric should allow the self-similarity to equal any scalar as long as this is the greatest similarity over all of the possible comparisons.

The minimality inequality generalizes to many dimensions in the sense that it applies recursively to every $[1, k-1]$-dimensional comparison contained in any k-dimensional comparison. Define a multidimensional similarity metric as *geometrically consistent* if every q-subset of any k-subset of the rows of a real vector space W is more similar to itself than any other q-subset of the same k-subset. Formally, where $[W]^{kCq}$ denotes the set of q-subsets of some k-subset of W $(k > q)$, then the following statement must be true:

$$\forall i, \forall q < k : \arg\sup_{j} \text{Sim}([W]_i^{kCq}, [W]_j^{kCq}) = i$$

In other words, if I choose any set of keywords, then every subset I can make by removing one or more words from the set must be more similar to itself with respect to my chosen similarity metric than any of the other subsets with the same number of words.

A proposed metric without this property is inadmissible because it is self-contradicting. Consider the k-simplex with edges corresponding to the pairwise Euclidean differences between a set of vectors indexed by any keyword list X. This k-simplex is bounded by a lattice of q-simplical subspaces formed by subsets of

X ($q \in [1, k-1], q \in \mathbb{Z}$). An admissible similarity metric must satisfy the minimality inequality with respect to all of the q-simplices of equal dimensionality. As a concrete example, consider the tetrahedron formed by $X = \{\text{he, him, his, himself}\}$. X has triangular faces $X_A = \{\text{he, him, himself}\}$ and $X_B = \{\text{he, him, his}\}$. X_A and X_B coincide at the edge between {he, him}, so the dihedral angle between A and B is proportional to the pointwise distance between {his} and {himself}. Now consider a proposed metric $\text{Sim}^*_W(\cdot, \cdot)$ that yields the result $\text{Sim}^*_W(X_A, X_A) < \text{Sim}^*_W(X_A, X_B)$. This implies that the distance between {himself} and {his} must be less than zero. $\text{Sim}^*_W(\cdot, \cdot)$ is inadmissible because this is not possible.

Define the index of inconsistency $\mathcal{I}_{q,k}(\text{Sim}_W(\cdot, \cdot))$ for a similarity metric over W as the proportion of q-subsets of a keyword list of cardinality k in W that are most similar to themselves with respect to $\text{Sim}_W(\cdot, \cdot)$. A geometrically consistent metric satisfies $\mathcal{I}_{q,k}(\text{Sim}_W(\cdot, \cdot)) = 1$ for any choice of q and k up to the ambient dimensionality defined by W. It is particularly worrisome if the expected value of $\mathcal{I}_{q,k}$ is decreasing as k increases; adding more relevant keywords to an analysis should not make it less consistent. The index of inconsistency is closely related to the regularity of the k-simplex induced by X. A helpful heuristic discrepancy measure is the Euclidean condition number of X, which measures the elongation of its hyperelliptical projection (*8*).

A canonical metric for multidimensional similarity. To construct a geometrically consistent multidimensional similarity metric, define $W(k)$ as the set of row subspaces spanned by k-subsets of the rows of W. Denote the row space of W as $\mathcal{R}(W)$. For all subspaces $A \in W(k)$ consider the corresponding orthogonal projector $\mathcal{P}(A) = A(A^T A)^{-1} A^T$. Then, equip $W(k)$ with the symmetric bilinear form $\text{Sim}_{\text{CCA; k}} : W(k) \times W(k) \to \mathbb{R}$ corresponding to the Frobenius inner product of the orthogonal projections (*9–11*):

$$\text{Sim}_{\text{CCA; k}}(W_a, W_b) = \langle \mathcal{P}(W_a), \mathcal{P}(W_b) \rangle_F = \text{tr}(\mathcal{P}(W_a)^T \mathcal{P}(W_b)).$$

$\text{Sim}_{\text{CCA; k}}$ defines an inner product over the space of orthogonal projections onto $\mathcal{R}(W)$ (*12*). It is always non-negative because the projection matrices are positive semi-definite. When the comparison is one-to-one it is exactly equivalent to the squared cosine similarity between the two vectors. Unlike mean cosine similarity, this quantity has the desired consistency properties with respect to the dimensionality of the comparison. $\text{Sim}_{\text{CCA; k}}(A, B)$ is constrained to lie between 0, indicating the subspaces share no common direction, and $p = \min(\text{rank}(A), \text{rank}(B))$, indicating the subspaces are isotropically aligned. The metric can be scaled to lie between [0, 1] if we divide it by $\sqrt{\text{Sim}_{\text{CCA; k}}(W_a, W_a)\text{Sim}_{\text{CCA; k}}(W_b, W_b)}$, equivalent to the geometric mean of the lengths of the underlying word lists. The dimensionwise alignment is equivalent to the vector of singular values obtained by taking the singular value decomposition of the projection matrix product; each singular value corresponds exactly to the cosine of the angle between the ith pair of singular

vectors. I refer to the total metric as the *canonical subspace metric* and its component quantities as the *canonical congruences*.

The metric is very closely related to canonical correlation analysis, but there are two critical modifications. First, the analysis is carried out without recentering the subspaces, so it is not a local correlation measure. We avoid recentering because we are working with a vector space, so the subsets of word vectors we have selected into the analysis already share a fixed point at the origin. If we subtract their local means, the overall difference in frequency of use between the two keyword lists distorts the resulting affine metric.[2] Second, we carry out the analysis between row subspaces of the same vector space, rather than between the column spaces of two matrices with rows corresponding to the same units. Rather than studying two sets of measures on the same set of observations, we are studying two different sets of observations with respect to the same set of measures, i.e. the variation we have estimated by the word embedding algorithm. The [0, 1] rescaled quantity has been discussed in many areas of multivariate analysis; in psychometrics it is called Tucker's congruence coefficient (*14*, *15*), and in the French school of data analysis it is more often called the RV coefficient (*16*, *17*).[3]

Why does WEAT fail to measure multidimensional semantic association? To show that score components in WEAT do not have the desired consistency property, it is helpful to review the design of the test. WEAT is defined by two operations on the input word vector space W. First, the researchers selects disjoint keyword lists of length k that appear in the vocabulary of W. This step is typically performed using preset IAT keyword lists. Given a keyword list, WEAT selects the row vector in W with a label corresponding to this word. As a running example, Table S1 lists the keywords for a WEAT measure comparing the differential association between "male" or "female" words and "pleasant" or "unpleasant" words.

Second, W is equipped with a bivariate association metric (cosine similarity) that quantifies the amount of association between all k^2 pairs of word vectors across two keyword lists. The WEAT measure is then computed by taking the arithmetic mean of the $k \times k$ matrix of cosine similarities and comparing these scores to the other three mean cosine similarities. Formally, let A, B, C, D be the subspaces of the row

[2]This can also be motivated from the perspective of the sparse high-dimensional word association measure approximated by the word embedding matrix. The word vectors lie on the surface of a high-dimensional convex body in $\mathbb{R}^n$ that we have mapped onto a hyperelliptical cross-section in $\mathbb{R}^k$; this surface captures most of the variation in position in the larger space (*13*). However, we must not forget that we are ignoring the remaining $n - k$ dimensions. Imagine then that the word embedding matrix W is missing $n - k$ columns of zeroes. It is safe to omit these columns when we take the unidimensional cosine because this does not affect the calculation of the angular metric once we have constructed the low-rank approximation. But, if we recenter subspaces of W without remembering the remaining dimensions, the estimated centroids will be very far away from the value we should have used, which is generally quite close to the zero vector.

[3]Kornblith and colleagues (*18*) propose using a kernelized version of this quantity to measure similarity between neural network layers. This measure could be used if researchers were using word lists with cardinality larger than rank(W). In practice, this is not necessary for word association problems of the scale targeted by WEAT-like models.

space of W induced by four mutually disjoint keyword lists of equal length $k < p$.[4] Then the WEAT score with respect to $W_{A,B;C,D}$ is the grand mean difference in cosine similarities across $\{A, B\}$ and $\{C, D\}$: $\text{WEAT}(A, B, C, D) = 1/k^2 \sum_i^k \sum_j^k (\cos(A_i, C_j) - \cos(A_i, D_j) + \cos(B_i, D_h) - \cos(B_i, C_j)$.

Due to the symmetry in the test it is sufficient to characterize the behavior of $\frac{1}{k^2} \sum_i^k \sum_j^k \cos(X_i, Y_i)$ for any two disjoint and arbitrarily ordered word vector subsets $X, Y \subset W$ of size k. Denote the subspaces of $\mathcal{R}(W)$ spanned by X, Y as W_X, W_Y. Denote the power sets of X, Y as $\text{Pow}(X), \text{Pow}(Y)$. Define the similarity metric $\text{Sim}_q(W_X, W_Y)$ between the subspaces corresponding to every q-subset in $\text{Pow}(X), \text{Pow}(Y)$, where $q \in [1, k_z - 1]$. As we increase k to the ambient dimensionality of W, the intersection of the orthogonal complements of W_X and W_Y eventually vanishes; that is, regardless of the choice of words, the two subspaces will approach one-another until they coincide exactly once we hit the number of dimensions in the vector space. If $\text{Sim}_q(W_X, W_Y)$ is geometrically consistent with respect to W, then this must also be true everywhere on the lattice of similarity metrics defined by $\text{Pow}(X), \text{Pow}(Y)$ as q increases. Thus $\text{Sim}_q(W_X, W_Y)$ must be nondecreasing as we increase q. Reciprocally, if some values of $\text{Sim}_q(W_X, W_Y)$ are not nondecreasing in k, then it cannot be geometrically consistent with respect to W.

Figure 1 demonstrates that the MCS estimator $\text{Sim}^*(I, J) := (1/k^2) \sum_{q=i} \sum_{q=j} \cos([W]_i^k, [W]_j^k)$ is not geometrically consistent for $q \geq 3, k \geq 4$. To show this, we must consider the lattice of all possible analyses involving subsets of words in a keyword list A. Each column of heatmaps depicts the lattice of canonical subspace metrics (left) mean cosine similarities (center panel) between every subset of the keyword list at the top of the figure. Each cell of the heatmap corresponds to the metric over the corresponding q-subset. A reasonable heuristic predictor of inconsistency is the 2-norm condition number of the cosine similarity matrix (see Fig. 1; right panel); the lower this value, the more likely it is that a subspace of dimension q will be confused for one of its counterparts.

Figure 2 displays the canonical subspace metric (Fig. 2; left panel) and the corresponding MCS metric (Fig. 2; center panel) over the lattice of comparisons between subsets of the keyword lists. In the right panel, the full distribution of metrics corresponding to each heatmap is displayed; the dots indicate (in vertical order) the maximum value, expectation, and minimum value of the metric over q. In the right panel only, the mean of the *squared* cosine similarities is used to show the equivalence with the canonical subspace metric for the unidimensional problem. The expectation of the MCS metric over all possible sub-analyses of equivalent dimension is decreasing in the input dimension, while the expected canonical subspace metric increases monotonically in the number of dimensions as desired. As the number of cosine similarity estimates corresponding to the size of the analysis increases, the association between any pair of words already in the

[4]For brevity, I only consider the case of equal word list lengths. In general, multidimensional similarity cannot have dimensionality exceeding the minimum cardinality of the input keyword lists.

list remains fixed with respect to the new words. This causes MCS to converge to the expected coplanar angle in the input vector space as we increase k. In other words, MCS is a particular estimator of the expected cosine similarity between *any two vectors* in X and Y, and it does not measure what or how much *all of the vectors* have in common.

Figure 3 displays the canonical congruences for the four subspace pairs compared to the three possible expected distributions of minimal common subspace alignment in $W(k)$. The corresponding WEAT score component for each metric is shown for comparison (MCS: dotted line; CCA: dashed line). Any comparison implies three reference distributions: the distribution obtained by randomizing both keyword lists, and the two distributions obtained by randomizing one list while holding the other fixed. There is no necessary relationship between these distributions, so there are many different answers to the question of whether an association is larger or smaller than what we would expect. One particularly interesting comparison is the difference between the estimate's relation to the two-way null and its relation to one or both of the one-way nulls. The association can be simultaneously smaller than what we would expect by a two-way random draw and larger than what we would expect by either one-way random draw, or vice-versa. There is no necessary relationship between the rank index of the estimated congruences and their positions with respect to their reference distributions; for example, only the eighth canonical congruence in the {male, unpleasant} comparison is outside all three 95% prediction intervals.

Implications for word association measurement. WEAT-like measures are widely used in the social sciences to measure a wide range of cultural processes observable in text that traditionally were measured by human raters (*19*, *20*). The key validity evidence for WEAT in this context is its agreement with human word association ratings (i.e., the IAT and similar psychological test-based methods). However, the convergent validity of the measure is weak evidence if it is not geometrically consistent. MCS does not satisfy this criterion. In practice, the canonical subspace metric is always a more consistent estimator of the targeted variation than MCS, and facilitates interpreting variation in the scale of observed associations in applied research. A general conclusion is that researchers should not use mean cosine similarity as a measure of multidimensional semantic association in applied statistical text analysis.

It is worth emphasizing that this result *cannot* be interpreted as evidence that stereotypes and biases do not exist. The canonical subspace metric surfaces word associations between gendered identity words and sentiment words with magnitude in excess of what we would expect from totally random selections of words. The more fundamental problem is that it is not clear why we should define (e.g.) stereotypical association for Black names with respect to white names, or biases against women with respect to men (*21*). The results in this paper challenge the notion that there are unambiguously categorically opposed sets of words that would

justify this analytic approach. In some contexts male/female and white/Black are talked about as if they are opposites, but this is only one limited perspective on the vast spectrum of similarities, differences, and ambiguities indexed by gender and racial identity (*22*). Reliance on keyword lists as a priori representations of "concepts" ("attributes", "identities", "schemas", etc.) renders this measurement approach dependent on the reader's intuitive acceptance of the concept label, rather than evidence that the keywords specifically pick out this concept. It is logically inconsistent to presume the meaning of words in advance of seeing their contexts if we are trying to learn word meanings by observing the contextual use of language (*24*).

References

1. A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases. *Science*. **356**, 183–186 (2017).

2. A. G. Greenwald, D. E. McGhee, J. L. Schwartz, Measuring individual differences in implicit cognition: The implicit association test. *Journal of personality and social psychology*. **74**, 1464 (1998).

3. H. Hotelling, Relations between two sets of variates. *Biometrika*. **28**, 321–377 (1936).

4. B. A. Nosek, M. R. Banaji, A. G. Greenwald, Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, research, and practice*. **6**, 101 (2002).

5. J. Pennington, R. Socher, C. D. Manning, "GloVe: Global vectors for word representation" in *Proceedings of the 2014 conference on empirical methods in natural language processing EMNLP* (2014), pp. 1532–1543.

6. P. L. Rodriguez, A. Spirling, B. M. Stewart, E. M. Wirsching, Multilanguage word embeddings for social scientists: Estimation, inference and validation resources for 157 languages (2023) (available at https://alcembeddings.org/).

7. A. Tversky, Features of similarity. *Psychological review*. **84**, 327 (1977).

8. G. H. Golub, C. F. Van Loan, *Matrix computations* (Johns Hopkins University Press, 2013).

9. W. Krzanowski, Between-groups comparison of principal components. *Journal of the American Statistical Association*. **74**, 703–707 (1979).

10. C. D. Meyer, *Matrix analysis and applied linear algebra* (SIAM, 2023).

11. I. Borg, P. J. Groenen, *Modern multidimensional scaling: Theory and applications* (Springer Science & Business Media, 2005).

12. R. A. Horn, C. R. Johnson, *Matrix analysis* (Cambridge University Press, 2012).

13. A. Dvoretzky, "Some results on convex bodies and banach spaces" in *Proceedings of the international symposium on linear spaces* (Jerusalem Academic Press, 1961), pp. 123–160.

14. L. R. Tucker, "A method for synthesis of factor analysis studies." (Educational Testing Service, 1951).

15. B. Korth, L. R. Tucker, The distribution of chance congruence coefficients from simulated data. *Psychometrika*. **40**, 361–372 (1975).

16. Y. Escoufier, Le traitement des variables vectorielles. *Biometrics*, 751–760 (1973).

17. P. Robert, Y. Escoufier, A unifying tool for linear multivariate statistical methods: The rv-coefficient. *Journal of the Royal Statistical Society Series C: Applied Statistics*. **25**, 257–265 (1976).

18. S. Kornblith, M. Norouzi, H. Lee, G. Hinton, "Similarity of neural network representations revisited" in *Proceedings of the 36th international conference on machine learning* (2019), pp. 3519–3529.

19. A. C. Kozlowski, M. Taddy, J. A. Evans, The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*. **84**, 905–949 (2019).

20. L. K. Nelson, Leveraging the alignment between machine learning and intersectionality: Using word

embeddings to measure intersectional experiences of the nineteenth century us south. *Poetics.* **88**, 101539 (2021).

21. S. S. Johfre, J. Freese, Reconsidering the reference category. *Sociological Methodology.* **51**, 253–269 (2021).
22. J. Butler, *Gender trouble: Feminism and the subversion of identity* (Routledge, 1990).
23. A. Hobbs, *A chosen exile: A history of racial passing in american life* (Harvard University Press, 2014).
24. L. Wittgenstein, *Philosophical investigations* (Blackwell Publishing, 1953).

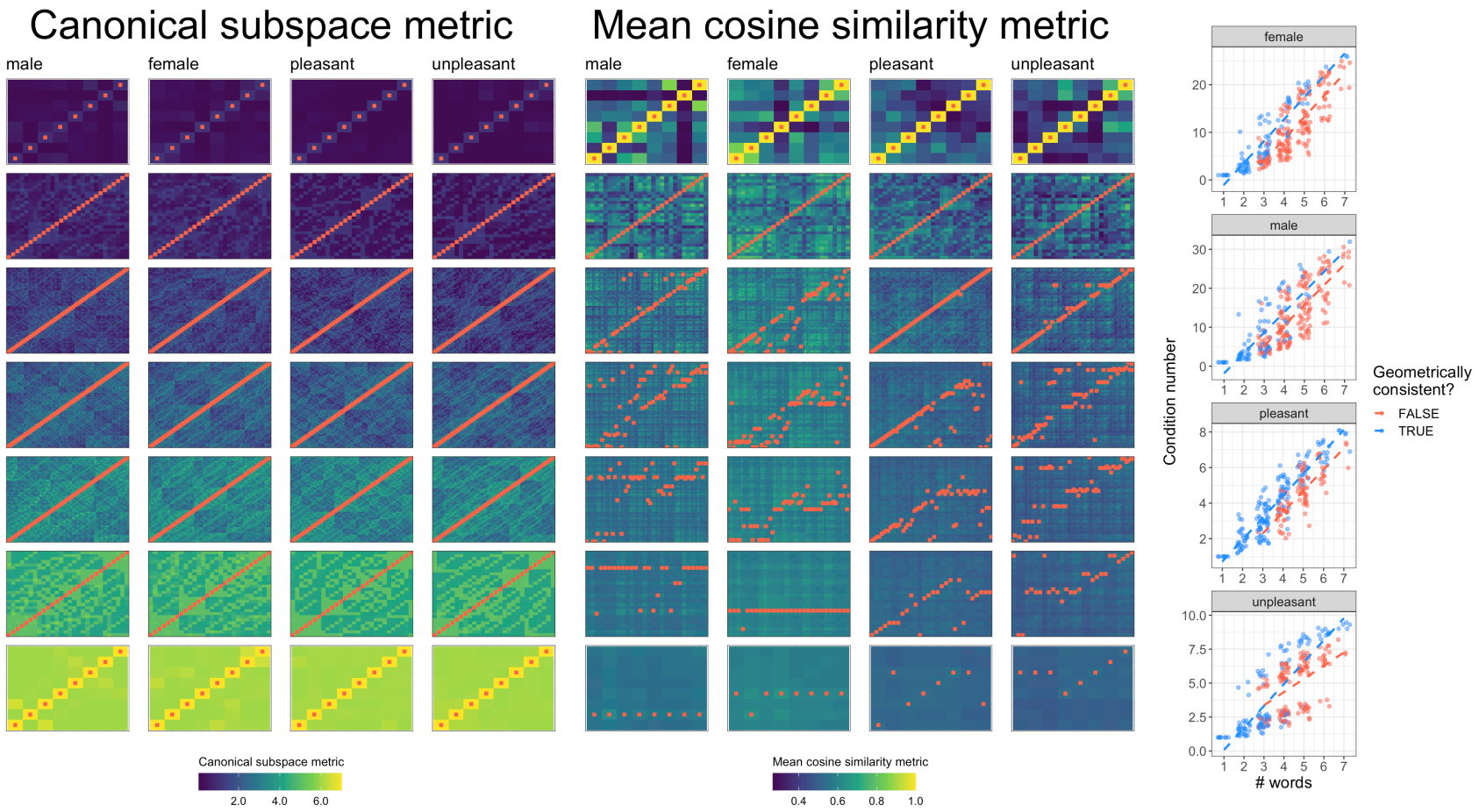
Figures

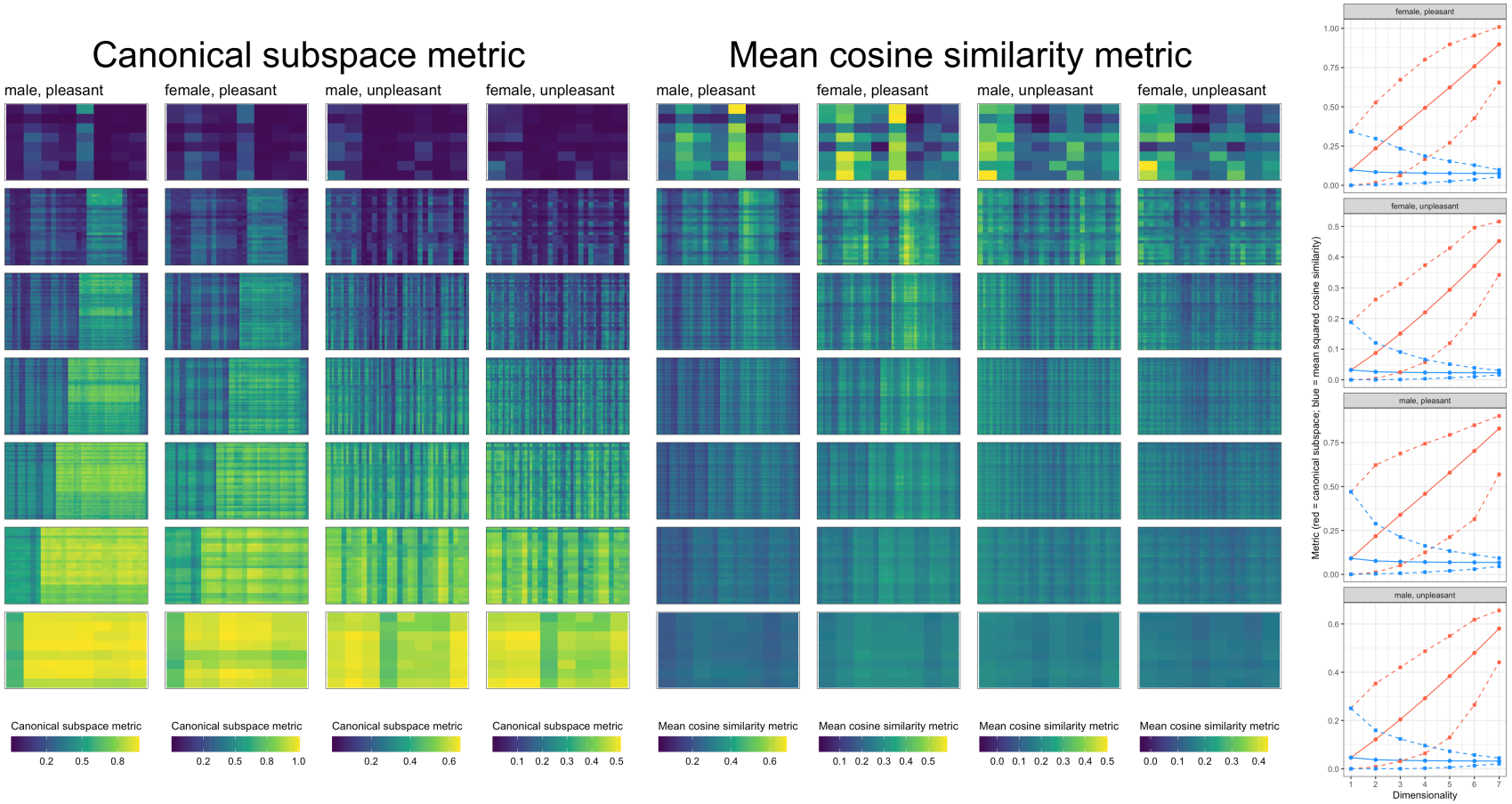
Captions

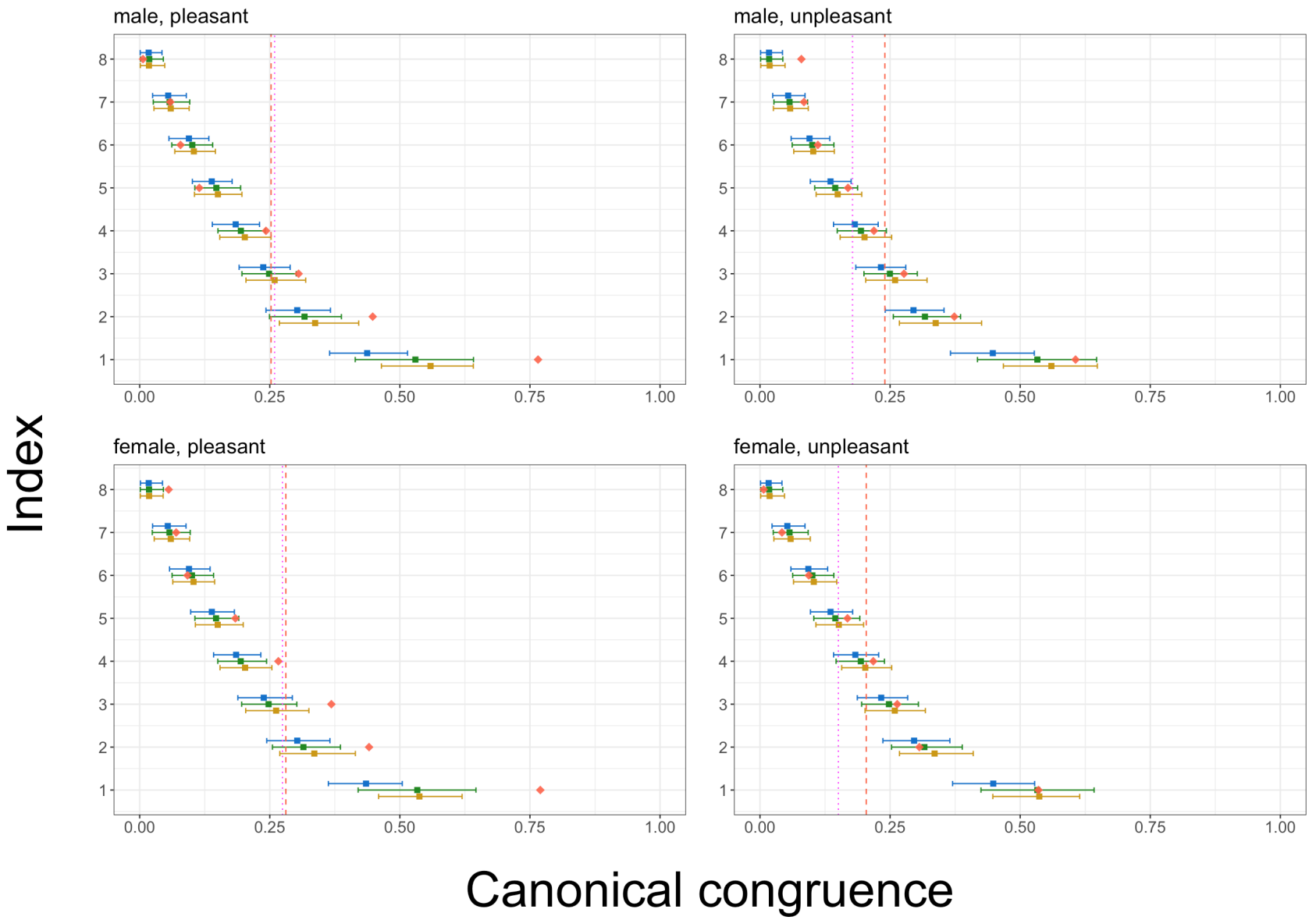
Figure 1. Multidimensional similarity metrics within subspaces. Each column forms a lattice over the set of comparisons between subspaces. Each row increases the subspace dimensionality by one (top row is 1, bottom row is 7). Red dots indicate the row/column maximum (note the matrices are symmetric). A geometrically consistent similarity measure for multidimensional word association problems must locate all of the maxima on the diagonal. The canonical subspace metric satisfies this criterion; the mean cosine similarity metric does not (heatmaps, left panel). Geometric inconsistency is partially predictable from the hyperelliptical elongation (condition number) of the k-simplex described by each cosine similarity matrix (scatterplots, right panel).

Figure 2. Multidimensional similarity metrics between subspaces. Cell colors indicate more commonality between subsets (heatmaps, left panel). The solid red (canonical subspace) and blue (mean squared cosine similarity) lines indicate their respective expectations. The dashed lines indicate the maximum and minimum values at each dimensionality. The canonical subspace metric increases as we gain more information about the common semantics of the input word. When the word association problem involves only one-to-one word pairings, the canonical subspace metric is equivalent to the *squared* cosine similarity. The mean cosine similarity decreases toward the global mean as more words are added to the analysis, whereas the canonical subspace metric increases.

Figure 3. Canonical congruences (red diamonds) with 95% prediction intervals for the one- and two-way null distributions. Blue intervals correspond to the one-way null distributions holding the pleasant/unpleasant lists constant while randomizing the male/female lists. Yellow intervals correspond to the one-way null distributions holding the male/female lists constant while randomizing the pleasant/unpleasant lists. Green intervals randomize both lists. The mean cosine similarity metric for each comparison is plotted as the pink dotted line; the canonical subspace metric is plotted as the red dashed line.


Canonical subspace metric
Mean cosine similarity metric
male
female
pleasant
unpleasant
Canonical subspace metric
2.0
4.0
6.0
Mean cosine similarity metric
0.4
0.6
0.8
1.0
Condition number
words
Geometrically consistent?
FALSE
TRUE


Canonical subspace metric
Mean cosine similarity metric
male, pleasant
female, pleasant
male, unpleasant
female, unpleasant
male, pleasant
female, pleasant
male, unpleasant
female, unpleasant
Canonical subspace metric
0.2 0.5 0.8
Canonical subspace metric
0.2 0.5 0.8 1.0
Canonical subspace metric
0.2 0.4 0.6
Canonical subspace metric
0.1 0.2 0.3 0.4 0.5
Mean cosine similarity metric
0.2 0.4 0.6
Mean cosine similarity metric
0.1 0.2 0.3 0.4 0.5
Mean cosine similarity metric
0.0 0.1 0.2 0.3 0.4 0.5
Mean cosine similarity metric
0.0 0.1 0.2 0.3 0.4
female, pleasant
female, unpleasant
male, pleasant
male, unpleasant
Metric (red = canonical subspace; blue = mean squared cosine similarity)
Dimensionality


male, pleasant
male, unpleasant
female, pleasant
female, unpleasant
Index
Canonical congruence
0.00
0.25
0.50
0.75
1.00
1
2
3
4
5
6
7
8

Tables

Captions

Table 1. Reanalysis of WEAT tests comparing mean cosine similarity metric (WEAT_{MCS}) to canonical subspace metric (WEAT_{CCA}). The test score is $(\text{Sim}(A, C) + \text{Sim}(B, D)) - (\text{Sim}(B, C) + \text{Sim}(A, D))$. The values of the similarity metric components corresponding to each of the keyword list comparisons is displayed in the eight columns on the right. The ratio of the two test scores ("Ratio") tends to be larger than 1, so Sim_{MCS} tends to overstate the difference in association. Sim_{MCS} also does not consistently estimate the distribution of order statistics for the multidimensional word association problem: the median Spearman rank correlation coefficient (ρ) between the mean cross-similarities is 0.4, and the ordering is correct only for one test (WEAT10). The two metrics exhibit opposite behaviors as the minimum cardinality of the input keyword lists (N) increases: Sim_{MCS} tends to shrink, while Sim_{CCA} tends to grow.

							Mean cosine similarity (MCS)				Canonical subspace metric (CCA)			
	Keyword lists	N	WEAT_{MCS}	WEAT_{CCA}	ρ	Ratio	A:C	A:D	B:D	B:C	A:C	A:D	B:D	B:C
WEAT 1	A: flowers B: insects C: pleasant D: unpleasant	25	0.041	0.002	0.4	18.854	0.040	0.014	0.034	0.020	0.131	0.121	0.118	0.126
WEAT 2	A: instruments B: weapons C: pleasant D: unpleasant	25	0.070	0.005	0.4	14.160	0.041	-0.003	0.066	0.040	0.111	0.115	0.132	0.123
WEAT 3	A: white names B: Black names C: pleasant D: unpleasant	25	0.023	0.003	-0.4	6.966	0.085	0.071	-0.023	-0.032	0.184	0.165	0.173	0.188
WEAT 4	A: white names B: Black names C: pleasant D: unpleasant	17	0.013	0.009	0.0	1.368	0.072	0.065	-0.012	-0.018	0.108	0.102	0.110	0.107
WEAT 5	A: white names B: Black names C: pleasant D: unpleasant	8	0.051	0.041	0.2	1.246	0.132	0.059	-0.020	0.002	0.097	0.063	0.087	0.080
WEAT 6	A: male names B: female names C: career D: family	8	0.138	0.022	0.2	6.265	0.210	0.208	0.236	0.100	0.065	0.069	0.059	0.033
WEAT 7	A: math B: arts C: male D: female	8	0.022	0.005	0.8	4.334	0.099	0.078	0.243	0.243	0.072	0.074	0.090	0.083
WEAT 8	A: science B: arts C: male D: female	8	0.011	-0.014	0.8	-0.818	0.139	0.104	0.209	0.233	0.045	0.053	0.079	0.085
WEAT 9	A: mental illness B: physical illness C: temporary D: permanent	7	0.128	0.065	0.4	1.963	0.124	0.124	0.216	0.087	0.073	0.056	0.116	0.067
WEAT 10	A: young names B: elderly names C: pleasant D: unpleasant	8	0.053	0.013	1.0	4.095	0.131	0.035	0.028	0.072	0.072	0.037	0.032	0.053

Supplementary Materials for "Geometrically consistent estimation of multidimensional word associations in text corpora"

Alexander T. Kindel

18 September 2023

Contents

Materials and methods

Analytic procedure

The embeddings used in the paper are publicly available GloVe embeddings of Wikipedia (*1*). Table S2 provides the keyword lists. To ensure the word embedding vocabularies support each analysis, I manually

adjust the spelling of some of the words in the original paper (*2*). A summary of these adjustments in each set of embeddings is also provided in Table S2. This is particularly important for the list of Black names derived from the IAT (*3*), because many of the names in this word list appear to be misspellings and/or close mispronunciations of Black names that are more common in the US context. For example, the list includes the common Arabic given name "Aiesha," but a far more common spelling of this name in the US context is "Aisha." The misspellings considerably affect the implementation of WEAT 3 in particular.

I reproduce the results in two alternative publicly available word embedding matrices: the Stanford NLP GloVe embeddings of Wikipedia 2014 and Gigaword 5 text (*4*), and the skip-gram negative sampling (i.e. word2vec) embeddings of Google News text (*5*). Figures S1-S3 rerun the figures in the paper with the word2vec embeddings; Figures S4-S6 rerun the figures with the Stanford NLP GloVe embeddings. Note that the substantive meaning of the cosine similarity metric differs somewhat between the two spaces because the GloVe embeddings use the $W + C$ representation and the word2vec embeddings use only the W matrix.

Computing the canonical subspace metric

The metric can be computed in `R` using the built-in CCA function `cancor` with the centering parameters set to False. In practice, if the canonical congruences are not of direct interest for the analysis (i.e. we only want to quantify the total amount of commonality), then it is not necessary to compute a singular value decomposition. Instead, it is faster to sum all of the values in the matrix obtained by computing the Hadamard (entrywise) product of the two projection matrices.

Supplementary text

On cosine similarity

Cosine similarity in the usual sense is a ratio of the Euclidean dot product between two vectors to the scalar product of their Euclidean norms: $\vec{x} \cdot \vec{y} / ||\vec{x}||_2 ||\vec{y}||_2$. It measures the amount of covariance between the two vectors that is in excess of what would be expected if they varied independently. The square of this quantity can be interpreted as the amount of variation retained when projecting one word vector onto the span of the other.

The widespread use of cosine similarity in the statistical measurement of word associations is attributable to the influential work of Gerard Salton and colleagues on information retrieval in large text document databases (*6*). However, there is a subtle but importance difference in research goals between the two enterprises. Information retrieval is concerned with optimally finding relevant documents, so it does not place very much weight on *how* the search is performed as long as it performs well — that is, we care about

what we retrieve, and how we retrieve it is only interesting to the extent that we find good matches. In contrast, the measurement of word associations in the social sciences is more or less pursuing the opposite goal: what we retrieve is only interesting insofar as we can say we have a measure of the thing we are interested in. Cosine similarity is a satisfactory plug-in heuristic for a range of prediction problems, but once we are interested in measurement, we must be much more precise about the relationship between what we mean and what our metrics say. This paper points out that this precision is particularly important if we are interested in many-to-many comparisons.

A key limitation of this quantity as a tool for data analysis is the moniker "cosine similarity." Such a ratio can be computed between any two vector-valued arguments, so the "cosine" in question can be many different things. Thus the term hides important variation in what it is used to measure, particularly when the arguments are initially derived from the same vector space but are pre-transformed in various ways before it is computed.

Notably, researchers frequently describe cosine similarity as "the dot product of the vectors after they have been normalized to unit length" (*2*). However, this statement is very misleading, because cosine similarity does not define an inner product space over the input word vector space due to the normalizing transformation. Adding estimated cosine similarities does not consistently aggregate information about linear combinations of the component vectors. The cosine similarity space $\{W, \text{Sim}_{\cos}\}$ is a transformation of the usual inner product space defined by the Euclidean dot product $\{W, \cdot\}$ such that every point on the bilinear form is locally weighted by $1/\sqrt{\langle W_a, W_a\rangle \langle W_b, W_b\rangle}$. This space is scale invariant and linear in each argument only under a stringent condition on W; specifically, the Euclidean dot product between two vectors must equal negative one-half their respective norms everywhere in the vector space:

$$\begin{aligned}
\varphi(A+B,C) &= \frac{\sum_i (A_i+B_i)(C_i)}{\sqrt{\sum_i (A_i+B_i)^2}\sqrt{\sum_i (C_i)^2}} \\
&= \frac{\sum_i (A_i+B_i)(C_i)}{Q_\varphi(A+B,C)} \\
&= \frac{\sum_i (A_iC_i+B_iC_i)}{Q_\varphi(A+B,C)} \\
&= \frac{\sum_i A_iC_i}{Q_\varphi(A+B,C)} + \frac{\sum_i B_iC_i}{Q_\varphi(A+B,C)}.
\end{aligned}$$

$$\begin{aligned}
Q_\varphi(A+B,C)^{-1} &= \sqrt{\sum_i (A_i+B_i)^2}\sqrt{\sum_i C_i^2} \\
&= \sqrt{\sum_i (A_i^2+B_i^2+2A_iB_i)}\sqrt{\sum_i C_i^2} \\
&= \sqrt{\sum_i A_i^2+\sum_i B_i^2+2\sum_i A_iB_i}\sqrt{\sum_i C_i^2} \\
&= \sqrt{||A||_2^2+||B||_2^2+2\langle A,B\rangle}||C|| \\
&= g^*(A,B)||C||.
\end{aligned}$$

$$g^*(A,B) = ||A|| = ||B|| \text{ if and only if } \langle A,B\rangle = -\frac{1}{2}||A||^2 = -\frac{1}{2}||B||^2$$

$$\begin{aligned}
\varphi(\lambda A,B) &= \frac{\sum_i (\lambda A_iB_i)}{\sqrt{\sum_i (\lambda A_i)^2}\sqrt{\sum_i B_i^2}} \\
&= \frac{\lambda\sum_i (A_iB_i)}{\sqrt{\sum_i (\lambda A_i)^2}\sqrt{\sum_i B_i^2}} \\
&= \frac{\lambda\sum_i (A_iB_i)}{\sqrt{\sum_i \lambda^2 (A_i)^2}\sqrt{\sum_i B_i^2}} \\
&= \frac{\lambda\sum_i (A_iB_i)}{\lambda\sqrt{\sum_i (A_i)^2}\sqrt{\sum_i B_i^2}} \\
&= \frac{\cancel{\lambda}\sum_i (A_iB_i)}{\cancel{\lambda}\sqrt{\sum_i (A_i)^2}\sqrt{\sum_i B_i^2}}
\end{aligned}$$

A shorthand for this assumption is that W must be *isotropic* for cosine similarity to define an inner product over W. Addition and multiplication over cosine similarities are not defined in the conventional way unless W is isotropic, and it is not possible for W to be isotropic. To see why, define the index of isotropy $\text{Iso}(X)$ to be the dimension of the largest isotropic subspace of X (*8*). To establish that an inner product space is anisotropic with respect to the inner product operator Q_ψ, it suffices to show that $\text{Iso}(X) = 0$, i.e., that X contains no self-orthogonal subspaces of dimension 1. This condition holds trivially for all popular

word embedding models because their codomain is $\mathbb{R}^p$, so the dot product between any vector and itself is strictly greater than zero.

Applying the arithmetic mean to $\text{Sim}_{\cos}$ despite this contradictory premise has an important substantive consequence: adding words to the inducing keyword lists always moves the representation toward generality rather than specificity. As k increases, the mean cosine similarity converges in probability to the expected cosine between two vectors in $\mathbb{R}^p$ irrespective of the specific choice of input subspaces and basis dimensionality. This is the opposite of what we want the metric to do. The number of ways a set can be partitioned is proportional to the cardinality of the set, so as we add more words to a word list, the number of ways that they can be associated increases. Consequently, if the word list is converging to a specific concept, we should gain *more information* about what we mean by adding to the word list, and not less.

One interpretation of the mean cosine similarity is that it is an estimator of the expected association between any two words in the input keyword lists. Unfortunately, this mean is only consistent for the intended quantity when one of the arguments is fixed (i.e., one of the keyword lists contains only one word). If that condition holds, then the underlying angles can be interpreted as the central angles that parameterize the subgroup of orthogonal rotations from the fixed vector to all of the vectors in the other list. However, the Euclidean arithmetic mean cosine is a biased estimator of the mean Euclidean *rotation*. The bias is proportional to the difference between the number of vectors in the analysis and the square root of the determinant of the arithmetic sum of the corresponding rotation matrices. See (*9*) for more detailed discussion.

There are parametric significance tests for dependent correlation and congruence coefficients based on the Fisher Z-transformation (*10*). In the paper, I prefer to compare each metrics to the 95% prediction interval obtained by randomizing one or both of the input word lists. There are two reasons for this. First, I think the comparison to randomization provides a more intuitive way of understanding uncertainty in the word association problem than the distributional result. For example, the prediction intervals could be discretely interpolated toward the observed metric by randomizing only a subset of the words in either list; this could be used to ascertain the leave-one-out sensitivity of the comparison. Second, because the input metric is intrinsically multidimensional, our uncertainty about it is also hypervolumetric; treating our uncertainty about the canonical subspace metric as a scalar interval obscures this fact.

Varieties of WEAT-like measurement models

In practice, researchers sometimes design WEAT-like analyses with two modified versions of the MCS metric. First, the original metric takes the grand mean of the matrix of cosine similarities between A

and B, so each word vector A occurs in k comparisons. Researchers sometimes employ a **paired estimator** $\text{WEAT}_1(A, B)$ that uses only a one-to-one comparison between the subspaces, so that each word vector occurs in just one angle. Second, the original measure constructs the matrix of cosine similarities with respect to each vector in A and B separately. Researchers sometimes use the **centroid estimator** $\text{WEAT}_k(A, B)$ that substitutes the centroid $\mathbf{C}(B)$ for B so that the score reflects the mean cosine similarity of every vector in A with $\mathbf{C}(B)$. In this section I denote the original estimator $\text{WEAT}_0(A, B)$ to reduce ambiguity. I discuss the relationship between all of these modeling decisions and the notion of "frequency bias" in much greater detail below.

Both methods introduce additional drawbacks. In particular, the centroid estimator $\text{WEAT}_k(A, B)$ is a poor representation of the targeted subspace in high dimensions. Locally, for all subspaces A the centroid for B does not induce a separable neighborhood with respect to A. Define the local neighborhood of $\mathbf{C}(B)$ to be the rank-ordered vector of cosine similarities between $\mathbf{C}(B)$ and $A \cup B$, and define $\text{Sep}(A|\mathbf{C}(B)) = \inf_i \text{rank}(\cos \mathbf{C}(B), A_i)$ to be the smallest rank statistic of this neighborhood corresponding to a vector in A. The distribution of $\text{Sep}(A|\mathbf{C}(B))$ is parameterized by k: as we add more word vectors to the centroid, it gets closer to A and further from B! Globally, the centroid is not especially close to its subspace in high-dimensional vector spaces because it does not lie on the hyperelliptical manifold defined by the input vector space. The number of multidimensional subspaces that are closer to $\mathbf{C}(B))$ than B is $\frac{(n_B-1)!}{k!(n_B-1-k)!}$, where n_B is the number of word vectors $W_x \in W$ satisfying $\cos(\mathbf{C}(B)), W_i) > \sup_i \text{rank} \cos(\mathbf{C}(B)), B_i)$. For example, if $k = 2$ and there are two vectors in closer alignment with the centroid than the less associated of the two subspace vectors, there is one subspace closer to the centroid than B. This number grows very quickly in n_B and k.

Numerically, the paired estimator $\text{WEAT}_1(A, B)$ often fares somewhat better than the typical metric, because it reduces the impact of overcounting high-similarity word pairs on the analysis. However, this comes at a steep cost: this estimator is not invariant to permutations of the input keyword lists. Thus we will get a different answer if we use (say) {he, him, his, himself} or {he, him, himself, his}. This renders the analysis entirely dependent on the researcher's choice of pairings.

On IAT keyword lists

The scientific provenance of the WEAT keyword lists reveals some conceptual ambiguity in the relation between word lists and their labels. The lists are derived from keyword lists used in the IAT (*3*). The keyword lists for "pleasantness" were taken from a study of word association ratings elicited from college students in the mid-1980s (*16*). These words were in turn derived from a book of word frequencies authored

in the 1940s (*17*). They narrowed down the initial list by having students rate each word using a pleasant-to-unpleasant Likert item for each word. Other word lists were taken from an earlier study of college students published in 1969 (*18*). In this study, the procedure is reversed: students are given the category word only, and are asked to write down related words for 30 seconds. Consequently, in some cases we have gone from category label to keyword list, and in other cases we have gone from keyword list to category label. This raises the question why the label is treated as a category or concept to which the other words belong, and not as another word in the list.

Frequency bias in cosine similarity regression

Previously, some researchers have observed that WEAT-like measures are predictable from the underlying word frequency information. This is often called *frequency bias*. This section characterizes the frequency bias phenomenon in a wide range of regression model specifications using the normalization weight function (i.e. the scalar product of the Euclidean norms of the input vectors). In general, frequency bias is a problematic way of describing the distortion in the estimator. For most applications the corpus frequency has already been used to estimate the word embeddings, so predicting the MCS metric from the corpus frequency is in some sense double-dipping into the data. More to the point, the problem is not really that the measure is *biased* by frequency per se, but that it is not a consistent estimator for the multidimensional quantity it aims to estimate. The frequency-predictable distortion in regression models employing the MCS metric as a variable is better understood as a prevalent symptom of this more fundamental validity issue.

I begin with the simplest case (one cosine similarity) and add analytic complexity progressively to explore the class of cosine similarity regressions (CSR). For the most part researchers have been using the more complex models; the binary case is useful for understanding where the model begins to distort. I discuss three subclasses of the model family in more depth:

1. The **centroid** cosine similarity regression comparing estimated word vectors to Euclidean centroids of sets of vectors;
2. The **overlapping** cosine similarity regression, in which the same vector appears in more than one cosine in the outcome; and
3. The **multivariate** cosine similarity regression incorporating at least one cosine-valued independent variable, potentially implying additional overlap structures and subspace isotropy assumptions.

Each of these design choices results in a measurement model that makes strong assumptions about the geometric relationship between the vectors that are chosen for analysis, and combining them results in a much more complex pattern of frequency distortion. Although I employ an omitted variable bias perspective

to characterize the econometric properties of these models, I hasten to emphasize that this perspective is a bit misleading, as it is not really possible to "correct" the "bias" without using a different metric.

Binary difference-in-cosines. In the simplest case researchers may wish to compare a set of cosine similarities constructed between two overlapping or analogous sets of word pairs (i.e., three or four sets of word vectors). I first discuss the non-overlapping case involving cosines with no recurring component vectors. The most basic summary quantity is the difference in mean cosine similarity, $\cos(A_i, B_i) = \alpha_0^* + \alpha_1^* X_i + \epsilon_i$, where X_i is a binary variable partitioning the word pair sets. The estimated coefficient α_1^* is interpreted as a potentially significant difference in the mean cosine similarity, indicating a substantively meaningful difference in linguistic association between the two groups.

The binary difference-in-cosines is a limiting case of a general relative similarity model where distances in the inner product space are evaluated relative to its ℓ_2 projection. Estimating this model implies two very strong assumptions about the error distribution of the inner product: its distribution with respect to the overall scale of the vector space, and its distribution in the subspaces on each side of the analysis. This can be seen more transparently by multiplying both sides of the model equation by the normalization weight LNW_i, which describes the scale of the association at every point in the vector space:

$$
\begin{aligned}
\cos(A_i, B_i) &= \alpha_0^* + \alpha_1^* X_i + \epsilon_i \\
\frac{\langle A_i, B_i \rangle}{\text{LNW}_i} &= \alpha_0^* + \alpha_1^* X_i + \epsilon_i \\
\langle A_i, B_i \rangle &= \alpha_0^* \text{LNW}_i + \alpha_1^* X_i \text{LNW}_i + \epsilon_i \text{LNW}_i \\
\langle A_i, B_i \rangle &= \alpha_0^* \text{LNW}_i + \alpha_1^* X_i \text{LNW}_i + \epsilon_i \text{LNW}_i \\
&\quad + \alpha_2^* + \alpha_3^* X_i
\end{aligned}
$$

The constrained model omits the two terms highlighted in red by fixing them to 0: an intercept term α_2^* and a term for the main effect of X; estimating the original model α_3^*. It also implicitly assumes that the error term ϵ_i has scale-dependent error. Most importantly, the interpretation of the α_1^* is revealed to be (under most conditions) an estimate of a linear interaction between the normalization weight function and the input partition.

There are a few interconnected problems with the model as an estimator of α_1^* that can be characterized from an omitted variable bias perspective (*19*). First, the missing intercept term α_2^* forces the line of best fit to pass through the origin. This leads to poor model fit in most practical data analysis settings. When working with word vectors, the distribution of vector norms is bounded away from zero, so this constraint

lies strictly outside the support of the data. Additionally, the inner product is only small/negative when the normalization weight is increasing, meaning the implicit frequency adjustment estimated by cosine similarity regression tends to have the wrong sign. The variance of the model is also likely to be high due to its strong dependence on the observations with large outlying errors on the tails of the normalization weight distribution.

Second, the missing main effect term $\alpha_3^* X_i$ distorts the estimated difference in means by requiring the trends in each group to converge at the intercept. This also forces the groupwise normalization weighting estimates to have the same sign. In practice this can result in a global normalizing adjustment that is nearly orthogonal to the *local* (subspace-specific) conditional distribution of the inner product in ℓ_2. This constraint also implies a conditional difference in means that is maximally distinct when the normalization weight increases. This tends to be the opposite of what we see in semantic vector spaces; we observe a wider range of inner product values when the normalization weight is low, and the scale-conditional difference tends to be driven by the low-norm-weight vectors. In practice, the difference in the marginal distribution of the inner product of a set of points with two focal vectors tends to be close to zero, particularly if the number of vectors in the analysis is small.[1]

A generalized (unconstrained) model of the conditional association, the *local inner product regression*, allows the missing intercept and main effect terms to covary with the inner product:

$$\langle A_i, B_i \rangle = \beta_0 + \beta_1 X_i + \beta_2 \text{LNW}_i + \beta_3 X_i \text{LNW}_i + v_i$$

Note that in this model the focus is no longer on the coefficient on X_i, as in the cosine similarity regression model. In practice there are a large range of values of β_1, but the estimate tends to be high-variance. Instead, the target of inference is the interaction effect β_3, which tells us how different the distance-size relationship is between the two groups. Informally, the model estimates a difference in the total amount of association between two sets of vector pairs dictated by the comparative design X that corrects for the conditional dependence of the inner product distribution on the vector norm distribution due to the design. The coefficients β_2 and $(\beta_2 + \beta_3)$ can be interpreted as an estimate of the mean "similarity" (i.e. the mean distance-size association) in the two groups, and β_3 reflects the difference in the size of this relationship.

A closely related specification adds an interaction with the inverse normalization weight into the original

[1]This provides some intution for why centering the vector space is a very effective way of improving the meaningfulness of cosine-based measures: both of these model constraints are more reasonable when the vector space passes through the origin.

model with cosine similarity as the dependent variable:

$$\cos(A_i, B_i) = \beta_2 + \beta_3 X_i + \beta_0 \frac{1}{\text{LNW}_i} + \beta_1 \frac{X_i}{\text{LNW}_i} + \eta_i$$

The numbering of the coefficients reflects the relationship between this model and the inner product regression coefficients; both equations estimate the same model, although each model imposes a different interpretation on the coefficients that correspond to each other: β_1 and β_3 swap interpretations as the coefficient on the grouping variable X and the coefficient on the group-scale interaction term, while β_0 and β_2 swap interpretations as the intercept and the coefficient on the scaling variable. The key difference in the models is whether the estimation focuses on the normalization weight or its reciprocal, equivalent to deciding whether the analysis should be performed on the original vector space or its canonical cosine-normalized projection.

An advantage of this corrected ratio model is the error term η_i. The error term of the inner product regression is by assumption a function of the normalization weight ($v_i = \epsilon_i \text{LNW}_i$) whereas the corrected model need not make this assumption (*20*). This means that the inner product regression is typically the more heteroskedastic estimator. In general the two models give similar results, so the practical implications of this difference are small, although the fit of one model tends to be better. When the inner product regression is preferable, a heteroskedasticity correction can be employed to compute standard errors. Researchers can also directly examine the change in error variance over the normalization weight distribution using the typical Lagrange multiplier test or a related analysis (*23*).

A natural next question is whether the lower-order terms for the vector norms that make up the local normalization weight should be included in the model. Depending on the research design this may introduce exact multicollinearities into the corrected inner product regression. A common issue is that the researcher has chosen sets of word vectors that imply different relative word frequency distributions, leading to the commonly observed "frequency bias" in the measure (*26*). In practice the similarity structure determines how this frequency heterogeneity impacts the estimate of interest. This is because the design of the comparison may lead to word vectors and their norms reoccurring in multiple model terms. I will discuss this issue further as it arises in the context of other models in the cosine similarity regression family.

Some useful decompositions. *Total scale distortion.* The frequency-related distortion between β_1 and α_1^* can be decomposed into terms reflecting the contribution of the omitted intercept and the omitted main effect of the grouping variable, modified by the partial effect on the outcome of, respectively, the normalization

weight function and the group-conditional normalization weight function (*22*):

$$\alpha_1^* = \beta_1 + \beta_2 \left(\frac{cov(X_i, \frac{1}{\text{LNW}_i})}{var(X_i)} \right) + \beta_3 \left(\frac{cov(X_i, \frac{X_i}{\text{LNW}_i})}{var(X_i)} \right)$$

The decomposition shows why it is usually helpful to think of the relationship to frequency in terms of distortion rather than as a bias in the model per se. Even in the simplest case, there are multiple sources of distortion that can cancel out or amplify depending on the particular dataset we happen to be working with. This means that the coefficient α_1^* can be close to β_1 even when the amount of distortion due to each omission is comparatively large. Typically β_2 and the lower-order partial effect are positive, so this situation arises when the signs of β_3 and $cov(X_i, \frac{1}{\text{LNW}_i})$ conflict. The absolute amount of distortion is usually a better measure of total "bias" than the equation above suggests due to this property of the uncorrected ratio model. In practice, because researchers construct more complex word association models, this can be difficult to tease apart analytically.

Local similarity estimation. The inner product regression model factors out a subspace-specific proportion of the inner product that is (by assumption) not predictable from the normalization weight distribution. This quantity is like cosine similarity, but comes with estimate of the amount each group of inner products should be discounted prior to applying a similarity interpretation. To see this, consider the inner product regression model when $X_i = 0$: $\langle A_i, B_i \rangle = \beta_0 + \beta_2 \text{LNW}_i + v_i$. This model implies that up to some error and minus an estimated adjustment β_0, $\langle A_i, B_i \rangle \propto \beta_2 \text{LNW}_i$. ($\beta_1$ and β_3 provide the additive intercept and slope changes for the $X_i = 1$ case.) In other words, by moving the normalization weight function to the other side of the model, the analysis pivots to estimating two different similarity functions over the local inner product subspaces implied by each group of vectors, rather than assuming that these distributions are known in advance.

This partition of the inner product suggests conceptualizing similarity as an operation that discounts the estimated inner product by some group-specific amount and applies the normalizing projection only to the proportion of the variation in the inner product that we have estimated we can attribute to the variation in normalization weights. An observation-wise similarity coefficient can be produced by subtracting the estimated group-specific intercept of the inner product regression from each observation's inner product, then dividing this quantity by the normalization weight for this observation:

$$\text{sim}_{local}(A_i, B_i) = \frac{\langle A_i, B_i \rangle - (\beta_0 + \beta_1 X_i)}{||A_i|| ||B_i||}$$

The difference between this quantity and $\cos(A_i, B_i)$ is a curvilinear function of the normalization weight (recall the dual interpretation of β_0 as the lower-order normalization weight coefficient in the ratio model), and the lower-order coefficient on the normalization weight in the component model β_2 is an estimate of the mean of this distribution. Notice that $\text{sim}_{global}(A_i, B_i) = (\beta_0 + \beta_1 X_i)/||A_i||||B_i||$ is also a similarity coefficient. In other words, another way of writing the cosine similarity when there is a known grouping is in terms of a decomposition into a local similarity component and a global discrepancy component that can be estimated from the data:

$$\cos(A_i, B_i) = \text{sim}_{local}(A_i, B_i) + \delta(A_i, B_i)$$

Assuming that $\beta_0 = \beta_1 = 0$ in the uncorrected cosine similarity regression forces the second similarity term to zero. This amounts to an assumption that all of the similarity between A and B must be related to the grouping variable. The main consequence of this assumption is that $\cos(A_i, B_i)$ is always inflated with respect to the grouping if the ratio is computed in the global basis. Employing the corrected model makes it possible to recover an estimate of $\text{sim}_{local}(A_i, B_i)$ that allows some of the global similarity to be unrelated to the local comparison the model is meant to test.

Weight function measurement error. A recurrent feature of this type of analysis is that the local normalization weight function is estimated from the same data as the inner product. One could reasonably object to this procedure. In the context of regression analysis, when the local normalization weight is introduced into the other side of the model, the model fit will usually increase (depending on sample size, etc.). Some of this added explanatory power is due to the fact that the normalization projection errors are correlated.[2] External estimates of word frequency (e.g. the Corpus of Contemporary American English) can be used to show that using external word frequency information results in qualitatively similar but high-

[2]A closely related issue is that word embedding analyses are sensitive to the intrinsic frequency-based selection in the word count sampling procedure. The large number of low-frequency words implies an unobserved set of related words that could have been included, and there is a steep bias-variance tradeoff involved in selecting the minimum frequency window. Both issues motivate considering an external word frequency estimate. Researchers usually use the observed word count above a threshold in a corpus of documents as a plug-in estimate of the probability of observing a word, but this estimate is itself frequency-dependent because there is heightened measurement error in the low frequency region and a sharp discontinuity at the threshold. The threshold is the part of the remaining vector space that is the most affected by the size-biased sampling problem. An implication of this idea is that cutting off the model at a fixed threshold can seriously impact the qualitative interpretation of the model, because this region of terms is also more specific on average. When the research goal is to measure concepts by identifying subspaces of the vector space, it is important to consider how qualitative interpretations of this space are affected by perturbations to this threshold. Additionally, denominator measurement error results in biased estimates, particularly when it is associated with X_i (*29*). This situation also creates difficulties for frequency adjustment because we do not usually have an external *group-specific* frequency estimate. In general researchers should prefer the heteroskedastic estimator because this model expresses an expectation that a similarity measure over word vectors "should" have frequency-dependent projection errors.

variance point estimates of the inner product regression model. Consequently researchers will tend to see a considerable jump in R^2 and changes in the substantive and statistical significance of estimated quantities of interest when any credible estimate of word frequency is added.

Centroid cosine similarity regression. In applied settings, researchers tend to operationalize concepts of interest by using fixed word lists to induce an entire set of related word vector pairs. To represent concepts as discrete mathematical objects, researchers aggregate the set of vectors that results from a word list into a centroid and use the cosine similarity to it as a quantity of interest. To continue our gender-sentiment analysis example from above, we would now be interested in estimating quantities involving (say) $\cos(\Omega(A_{masculine}), B_{unpleasant})$ and $\cos(\Omega(A_{feminine}), B_{pleasant})$, where $\Omega(w)$ indicates the Euclidean centroid of the corresponding vector set. This strategy appears commonly in applied settings; although tempting, in general it tends to intensify the frequency distortion relative to a direct analysis of the component vectors.

Comparisons of this type can be factored into a set of inner products between the j, k selected sets of component vectors in the original vector space. The case with one additional vector, $\cos(A+B, C)$, provides a useful basic case to show how this analytic choice contributes to frequency distortion. First, observe that the normalization weight function implied by adding the vectors includes the pairwise component vector inner products:

$$\begin{aligned}
\cos(A+B, C) &= \frac{\sum_i (A_i + B_i)(C_i)}{\sqrt{\sum_i (A_i + B_i)^2}\sqrt{\sum_i (C_i)^2}} \\
&= \frac{\sum_i (A_i + B_i)(C_i)}{\text{LNW}(A+B, C)} \\
&= \frac{\sum_i (A_i C_i + B_i C_i)}{\text{LNW}(A+B, C)} \\
&= \frac{\sum_i A_i C_i}{\text{LNW}(A+B, C)} + \frac{\sum_i B_i C_i}{\text{LNW}(A+B, C)}.
\end{aligned}$$

$$\begin{aligned}
\text{LNW}(A+B, C) &= \sqrt{\sum_i (A_i^2 + B_i^2 + 2A_i B_i)}\sqrt{\sum_i C_i^2} \\
&= \sqrt{\sum_i A_i^2 + \sum_i B_i^2 + 2\sum_i A_i B_i}\sqrt{\sum_i C_i^2} \\
&= \sqrt{||A||_2^2 + ||B||_2^2 + 2\langle A, B\rangle}||C|| \\
&= g(A_i, B_i)\ ||C_i||
\end{aligned}$$

Note that $g(A_i, B_i)$ is a function of the inner product between A and B in addition to their respective norms. This leads to a different expansion of the uncorrected cosine model into the inner product regression:

$$\begin{aligned}
\cos(A_i + B_i, C_i) &= \alpha_0^* + \alpha_1^* X_i + \epsilon_i \\
\frac{\langle A_i + B_i, C_i \rangle}{\text{LNW}(A + B, C)} &= \alpha_0^* + \alpha_1^* X_i + \epsilon_i \\
\langle A_i + B_i, C_i \rangle &= \alpha_0^* \text{LNW}(A + B, C) + \alpha_1^* X_i \text{LNW}(A + B, C) + \epsilon_i \text{LNW}(A + B, C) \\
\langle A_i + B_i, C_i \rangle &= \alpha_0^* g(A_i, B_i)\ ||C_i|| + \alpha_1^* X_i g(A_i, B_i)\ ||C_i|| + v_i \\
&\quad + \alpha_2^* + \alpha_3^* X_i + \alpha_4^* g(A_i, B_i) + \alpha_5^* ||C_i|| \\
&\quad + \alpha_6^* X_i g_{A,B}(A, B) + \alpha_7^* X_i ||C_i||
\end{aligned}$$

Including an additional vector on one side of the cosine ratio outcome implies a model that omits four additional variables by fixing their coefficients to zero. In addition to the missing intercept and lower-order grouping variable term as in the binary analysis, a key issue is determining whether the components of the normalization weight have distinct (conditional) associations with the transformed inner product.

A particularly common version of this model arises when the collection of cosine similarities is derived from comparisons of centroids composed of j word vectors to members of a set of n word vectors, where $|A| > |B| = 1$. (The paper by Nelson [2021] is an example of this type of model.) Extending to j vectors:

$$\begin{aligned}
\cos(A_1 + ... + A_j, X) = \cos(\Omega, X) &= \frac{\sum_j \sum_i (\Omega_{ji} X_i)}{\sqrt{\sum_i (\sum_j \Omega_{ji})^2} \sqrt{\sum_i X_i^2}} \\
&\left(= \frac{\sum_i (A_1 X_i + ... + A_j X_i)}{\sqrt{\sum_i (A_1 + ... + A_j)^2} \sqrt{\sum_i X_i^2}} \right) \\
\sqrt{\sum_i (\sum_j \Omega_{ji})^2} &= \sqrt{\sum_j ||\Omega_j||_2^2 + 2 \sum_{j,j*} \langle \Omega_j, \Omega_{j*} \rangle}. \\
\text{LNW}(\Omega, X) &= \sqrt{\sum_j ||\Omega_j||_2^2 + 2 \sum_{j,j*} \langle \Omega_j, \Omega_{j*} \rangle} \sqrt{\sum_i X_i^2}
\end{aligned}$$

This cosine similarity is a weighted average of the inner products between the target vector X and each of the component vectors Ω_j. The normalization weight is a function of the squared norms of each of the component vectors and their pairwise inner products with each other. Observe two facts about this weight. First, in each inner product the familiar normalization weight $||\Omega_j||||\Omega_{j*}||$ reappears again. Second, adding the jth word vector to Ω implies adding $j - 1$ more comparisons to the weight function. Pre-composing the word vectors in A to facilitate the comparison to each X_i thus implies computing all of the pairwise inner products

$\langle A_i, A_j \rangle$, but without accounting for the uncertainty associated with this measure. The corresponding inner product regression is as follows:

$$\begin{aligned}
\cos(\Omega_i, X_i) &= \gamma_0^* + \gamma_1^* D_i + \epsilon_i \\
\langle \Omega_i, X_i \rangle &= \gamma_0 \text{LNW}(\Omega_i, X_i) + \gamma_1 D_i \text{LNW}(\Omega_i, X_i) + v_i + \gamma_2 + \gamma_3 D_i \\
\sum_j \langle \Omega_{ji}, X_i \rangle &= \gamma_0 \sqrt{\sum_j ||\Omega_j||_2^2 + 2 \sum_{j,j*} \langle \Omega_j, \Omega_{j*} \rangle} \sqrt{\sum_i X_i^2} \\
&\quad + \gamma_1 D_i \sqrt{\sum_j ||\Omega_j||_2^2 + 2 \sum_{j,j*} \langle \Omega_j, \Omega_{j*} \rangle} \sqrt{\sum_i X_i^2} + v_i + \gamma_2 + \gamma_3 D_i
\end{aligned}$$

The left hand side of the model decomposes into a sum of the component inner products. The implicit model for a given comparison can be constructed by moving the $j-1$ alternative comparisons to the right hand side of the regression equation:

$$\begin{aligned}
\langle \Omega_{ji}, X_i \rangle &= \gamma_0 \sqrt{\sum_j ||\Omega_j||_2^2 + 2 \sum_{j,j*} \langle \Omega_j, \Omega_{j*} \rangle} \sqrt{\sum_i X_i^2} \\
&\quad + \gamma_1 D_i \sqrt{\sum_j ||\Omega_j||_2^2 + 2 \sum_{j,j*} \langle \Omega_j, \Omega_{j*} \rangle} \sqrt{\sum_i X_i^2} \\
&\quad + \gamma_2 + \gamma_3 D_i + v_i \\
&\quad + \xi_1 \langle \Omega_{1,i}, X_i \rangle + \cdots + \xi_{j-1} \left\langle \Omega_{(j-1),i}, X_i \right\rangle
\end{aligned}$$

The inner product regression and the constrained/uncorrected model alike assume that the linked inner product coefficients are all equal; this is equivalent to assuming that the corresponding subspace is totally isotropic. Additionally, the zero conditional mean error assumption in this model also requires the errors of these distances to be mutually uncorrelated. This is a very stringent set of assumptions.

Note that the frequency distortion in cosine similarity induced by pooling the A vectors and constructing normalization weights in this way is not addressed by correcting the normalization weight adjustment. The reason for this is that the grouping variable D_i and the centroid norm $||\Omega^{X_i}||$ are exactly collinear, implying the comparison as constructed cannot differentiate whether the coefficient γ_3 reflects the grouping of vectors X_i corresponding to Ω^{X_i} or the scale of the subspace they inhabit. This relates to the question of whether a model of similarity should allow $||A||$ and $||B||$ to have separable associations with the inner product. I will discuss this issue in depth in the context of a more complex bivariate model.

Centroid cosines are anisotropically weighted estimates of the subspace mean cosine similarity. How should researchers think about the isotropy assumption implicit in comparing quantities like $\cos(\Omega, X_i)$?

One way to break down this model is to think of it as a specific *estimate* of the subspace-specific (i.e. conditional) mean cosine similarity $\mathbb{E}[\cos(\Omega_j, X_i)|J = j]$ implied by the word vectors that make up the composed vector, or alternatively $\mathbb{E}[\langle \Omega_j, X_i \rangle | LNW_{ij}, J = j]$. This simple comparison clarifies what it would take for the centroid model to be no more frequency distorted than the simple mode:

$$\cos\left(\frac{1}{N}\sum_j \Omega_{ji}, X_i\right) = \frac{1}{N}\sum_j \cos(\Omega_{ji}, X_i)$$

$$\frac{\left\langle \frac{1}{N}\sum_j \Omega_{ji}, X_i \right\rangle}{||N^{-1}\sum_j \Omega_{ji}||||X_i||} = \frac{1}{N}\sum_j \frac{\langle \Omega_{ji}, X_i \rangle}{||\Omega_{ji}||||X_i||}$$

$$\frac{1}{N}\sum_j \frac{\langle \Omega_{ji}, X_i \rangle}{||N^{-1}\sum_j \Omega_{ji}||||X_i||} = \frac{1}{N}\sum_j \frac{\langle \Omega_{ji}, X_i \rangle}{||\Omega_{ji}||||X_i||}$$

$$\frac{1}{N}\sum_j \frac{\langle \Omega_{ji}, X_i \rangle}{||N^{-1}\sum_j \Omega_{ji}||} = \frac{1}{N}\sum_j \frac{\langle \Omega_{ji}, X_i \rangle}{||\Omega_{ji}||}$$

$$\frac{1}{N}\sum_j \langle \Omega_{ji}, X_i \rangle = \frac{1}{N}\sum_j \frac{\langle \Omega_{ji}, X_i \rangle \, ||N^{-1}\sum_j \Omega_{ji}||}{||\Omega_{ji}||}$$

The centroid cosine similarity (to a unit reference vector) is a *weighted* mean of the component inner products. For the centroid to describe the subspace associations without distortion, the weights $\psi_{ij} = \frac{||N^{-1}\sum_j \Omega_{ji}||}{||\Omega_{ji}||}$ must be distributed with mean one; that is, the centroid must lie in an approximately isotropic subspace. By construction this assumption is not met in word embedding analysis. Because the scale distortion is complex, a large number of situations may result in approximately zero net bias. On average the distortion in the estimator increases with the number of components in the mean vector; mean vectors are mechanically shortened by the number of component vectors, resulting in a lower scale ratio due to the researcher's choice of subspace size (i.e. number of component vectors). The centroid cosine model tends to overstate the mean subspace cosine as a result.

A minimal alternative procedure is to estimate a separate model for each of the j variables in the composed vector, so that the model parameters can vary from comparison to comparison rather than presuming a single fixed parameter across the entire set. This has the advantage of considerably simplifying the interpretation of the normalization weighting procedure. The inclusion of the ξ_j terms is also suggestive of the bivariate cosine regression (i.e. on average, how mutually predictable are the pairwise distances in Ω to X, conditional on frequency?). Alternatively, researchers could estimate a multivariate linear model with the set of inner products as the outcome vector, which combines both approaches.

Bivariate cosine similarity regression. Researchers sometimes estimate a linear relationship between two sets of cosine similarities directly. For example, we might model the cosine similarity between masculine

and pleasant words as a linear function of the similarity between feminine and unpleasant words. This leads to another set of omitted variable constraints:

$$\begin{aligned}
\cos(A_i, B_i) &= \gamma_0^* + \gamma_1^* \cos(C_i, D_i) + \epsilon_i \\
\langle A_i, B_i \rangle &= \gamma_0 \text{LNW}(A_i, B_i) + \gamma_1 \cos(C_i, D_i) \text{LNW}(A_i, B_i) + v_i \\
&\quad + \gamma_2 + \gamma_3 \cos(C_i, D_i) + \gamma_4 \text{LNW}^{-1}(C_i, D_i) + \gamma_5 \langle C_i, D_i \rangle \\
&\quad + \gamma_6 \langle C_i, D_i \rangle \text{LNW}(A_i, B_i) + \gamma_7 \frac{\text{LNW}(A_i, B_i)}{\text{LNW}(C_i, D_i)}
\end{aligned}$$

Informally, this model creates frequency dependence due to the ratio of frequency functions on each side of the model, in addition to the intrinsic distributional shift on the left side alone. In addition to the zeroed γ_2 and γ_3 coefficients, the bivariate cosine similarity regression omits terms for the lower-order components of the right cosine (γ_4, γ_5). Additionally, γ_6 captures the potential interaction of the right inner product with the left normalization weight; intuitively, this is an estimate of how different we think the distance association is at different locations in the left frequency distribution. Analogously, γ_7 assesses whether there is a difference in the scale association over the distribution of left normalization weights. These terms are particularly important to include if the dependent and independent cosines share a component word vector (i.e. if the incorrect specification is of the form $\cos(A_i, B_i) = \gamma_0^* + \gamma_1^* \cos(A_i, C_i) + \epsilon_i$). This occurs often in applied settings. I discuss this model in more detail below.

Bivariate centroid cosine similarity regression. The composed cosines Ω° can be substituted into the bivariate cosine regression to yield a combined model:

$$\begin{aligned}
\cos(\Omega^A, Y_i) &= \gamma_0^* + \gamma_1^* \cos(\Omega^B, X_i) + \epsilon_i \\
\langle \Omega^A, Y_i \rangle &= \gamma_0 \text{LNW}(\Omega^A, Y_i) + \gamma_1 \cos(\Omega^B, X_i) \text{LNW}(\Omega^A, Y_i) + v_i \\
&\quad + \gamma_2 + \gamma_3 \cos(\Omega^B, X_i) + \gamma_4 \text{LNW}^{-1}(\Omega^B, X_i) + \gamma_5 \langle \Omega^B, X_i \rangle \\
&\quad + \gamma_6 \langle \Omega^B, X_i \rangle \text{LNW}(\Omega^A, Y_i) + \gamma_7 \frac{\text{LNW}(\Omega^A, Y_i)}{\text{LNW}(\Omega^B, X_i)}
\end{aligned}$$

Each of the right-hand inner product terms ($\gamma_1, \gamma_3, \gamma_5, \gamma_6$) decomposes further into a sum of componentwise inner products (teal terms below). In the uncorrected model all but γ_1 are zeroed out, but otherwise each of these coefficients is forced to be equal across the marginal and joint inner product and left-hand normalization weight distributions. Additionally, the model incorporates a left-hand centroid, so the isotropy constraint

applies here as well (purple).

$$
\begin{aligned}
\left\langle \Omega_j^A, Y_i \right\rangle = \gamma_0 \text{LNW}(\Omega^A, Y_i) + v_i \\
&+ \gamma_2 + \gamma_4 \text{LNW}^{-1}(\Omega^B, X_i) + \gamma_7 \frac{\text{LNW}(\Omega^A, Y_i)}{\text{LNW}(\Omega^B, X_i)} \\
&+ \gamma_1 \cos(B_1, X_i) \text{LNW}(\Omega^A, Y_i) + \gamma_1 \cos(B_2, X_i) \text{LNW}(\Omega^A, Y_i) + ... \\
&+ \gamma_3 \cos(B_1, X_i) + \gamma_3 \cos(B_2, X_i) + ... \\
&+ \gamma_5 \left\langle B_1, X_i \right\rangle + \gamma_5 \left\langle B_2, X_i \right\rangle + ... \\
&+ \gamma_6 \left\langle B_1, X_i \right\rangle \text{LNW}(\Omega^A, Y_i) + \gamma_6 \left\langle B_2, X_i \right\rangle \text{LNW}(\Omega^A, Y_i) + ... \\
&+ \xi_1 \left\langle \Omega_{1,i}^A, X_i \right\rangle + \cdots + \xi_{j-1} \left\langle \Omega_{(j-1),i}^A, X_i \right\rangle
\end{aligned}
$$

Overlapping cosine similarity regression. Up to this point I have assumed that the underlying word vector sets are non-overlapping, but in practice researchers often construct overlapping comparisons, for example the comparative association of masculine words with pleasant or unpleasant words. The behavior of the model when the cosines on each side of the model share a reference vector ($Y_i = X_i$) is of special interest. This is a common setup when the interest is in comparing how two concepts (represented by a set of vectors or a composed vector) relate to a common set of words. The dependence structure in the local normalization weight function implies that the lower-order interactions $D_i||w_i||$ and $D_i||X_i||$ and the main norm terms $||w_i||$ and $||X_i||$ are of interest. In addition the model assumes that the difference in distances is completely controlled by the distance-scale association and that this association passes through the origin, as in the simple case:

$$
\begin{aligned}
\cos(w_i, X_i) &= \alpha_0^* + \alpha_1^* D_i + \epsilon_i \\
\langle w_i, X_i \rangle &= \alpha_0^* \text{LNW}_i + \alpha_1^* D_i \text{LNW}_i + \epsilon_i \text{LNW}_i \\
\langle w_i, X_i \rangle &= \alpha_0^* \text{LNW}_i + \epsilon_i \text{LNW}_i \\
&\quad + \alpha_1^* D_i ||w_i|| ||X_i|| \\
&\quad + \alpha_2^* + \alpha_3^* D_i
\end{aligned}
$$

A key limitation of the computation of the score is that it does not allow the contribution of the $||w_i||$ term to vary; this makes the normalization weight terms of each individual difficult to interpret on their own, because the lack of variance in the vector set implies we cannot estimate the uncertainty associated with applying the norm weight in each model. The WEAT statistic $S(X, Y, A, B)$ compares the difference in means of the vectors of uncorrected interaction coefficients $\alpha_1^*(i), \alpha_1^*(j)$ across the word sets $X(i)$ and $Y(j)$, resulting in a

nested linear model. The sampling distributions of these coefficients are bilinearly correlated, and each of the word sets generally imply different frequency distributions. This complex dependency structure in the resulting nested model propagates the distortion in the component scores.

Overlapping bivariate centroid cosine similarity regression. Overlap may occur in the multivariate and centroid models separately. For conciseness, I discuss only their joint interaction with overlap and do not comment on the separate cases (overlapping centroid CSR; overlapping multivariate CSR). Continuing the previous example, we may wish to compare two centroids Ω^A and Ω^B to the same set of vectors X. The data for this comparison consist of a set of pairwise overlapping cosines, $\{\cos(\Omega^A, X_i), \cos(\Omega^B, X_i)\}_i$, and the corresponding normalization weights are $\{||\Omega^A||||X_i||, ||\Omega^B||||X_i||\}_i$. The key phenomenon to observe in the model is that the normalization weight factor for the word vector X_i cancels from the scale ratio terms with coefficients γ_1 and γ_7 (orange):

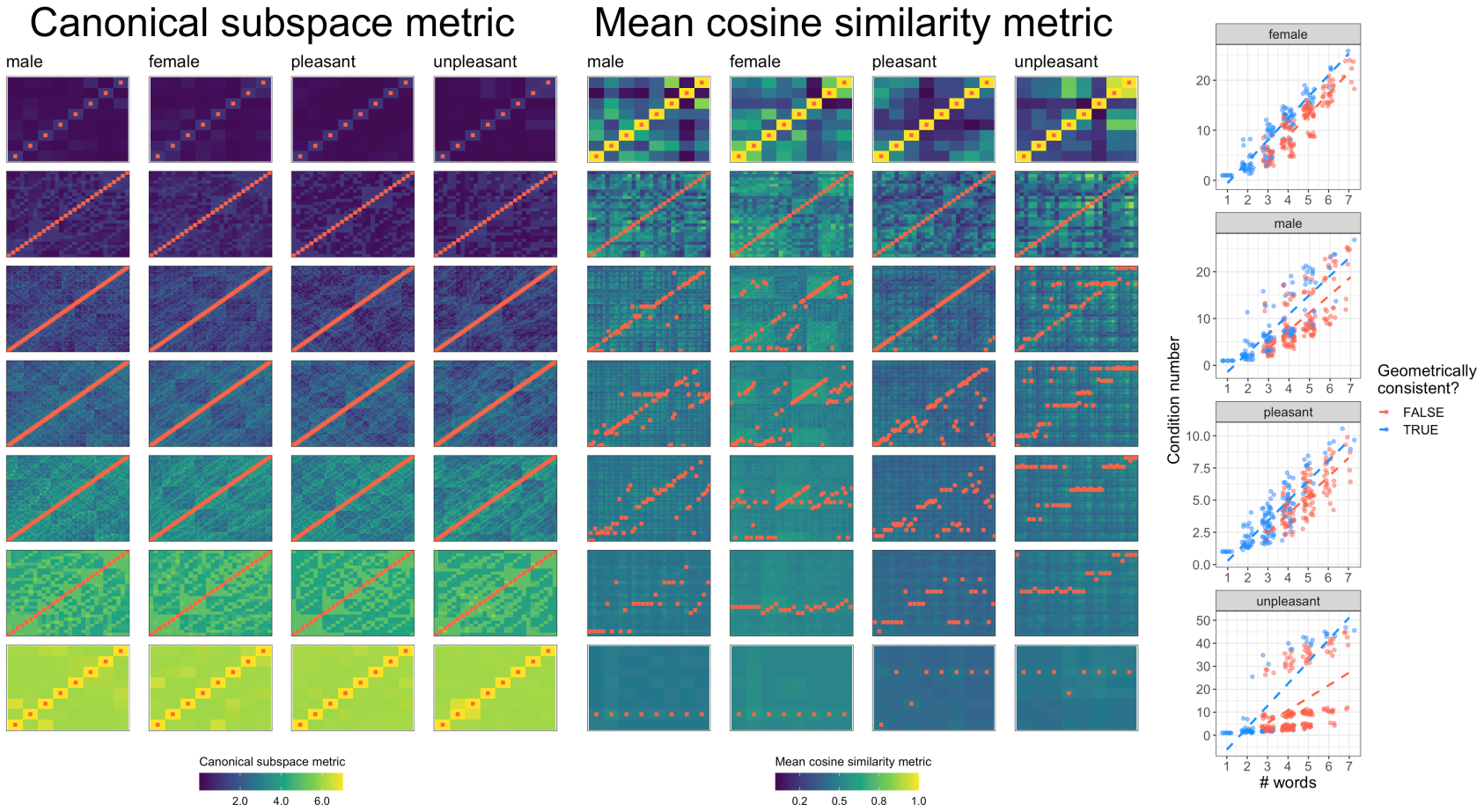
$$
\begin{aligned}
\cos(\Omega^A, X_i) &= \gamma_0^* + \gamma_1^* \cos(\Omega^B, X_i) + \epsilon_i \\
\left\langle \Omega_j^A, X_i \right\rangle &= \gamma_0 \mathrm{LNW}(\Omega^A, X_i) + v_i \\
&\quad + \gamma_2 + \gamma_4 \mathrm{LNW}^{-1}(\Omega^B, X_i) \\
&\quad + \gamma_3 \cos(B_1, X_i) + \gamma_3 \cos(B_2, X_i) + ... \\
&\quad + \gamma_5 \left\langle B_1, X_i \right\rangle + \gamma_5 \left\langle B_2, X_i \right\rangle + ... \\
&\quad + \gamma_6 \left\langle B_1, X_i \right\rangle \mathrm{LNW}(\Omega^A, Y_i) + \gamma_6 \left\langle B_2, X_i \right\rangle \mathrm{LNW}(\Omega^A, Y_i) + ... \\
&\quad + \xi_1 \left\langle \Omega_{1,i}^A, X_i \right\rangle + \cdots + \xi_{j-1} \left\langle \Omega_{(j-1),i}^A, X_i \right\rangle \\
&\quad + \gamma_1 \frac{\left\langle \Omega^B, X_i \right\rangle \sqrt{\sum_j ||\Omega_j^A||_2^2 + 2\sum_{j,j*} \left\langle \Omega_j^A, \Omega_{j*}^A \right\rangle} \cancel{\sqrt{\sum_i X_i^2}}}{\sqrt{\sum_j ||\Omega_j^B||_2^2 + 2\sum_{j,j*} \left\langle \Omega_j^B, \Omega_{j*}^B \right\rangle} \cancel{\sqrt{\sum_i X_i^2}}} \\
&\quad + \gamma_7 \frac{\sqrt{\sum_j ||\Omega_j^A||_2^2 + 2\sum_{j,j*} \left\langle \Omega_j^A, \Omega_{j*}^A \right\rangle} \cancel{\sqrt{\sum_i X_i^2}}}{\sqrt{\sum_j ||\Omega_j^B||_2^2 + 2\sum_{j,j*} \left\langle \Omega_j^B, \Omega_{j*}^B \right\rangle} \cancel{\sqrt{\sum_i X_i^2}}}
\end{aligned}
$$

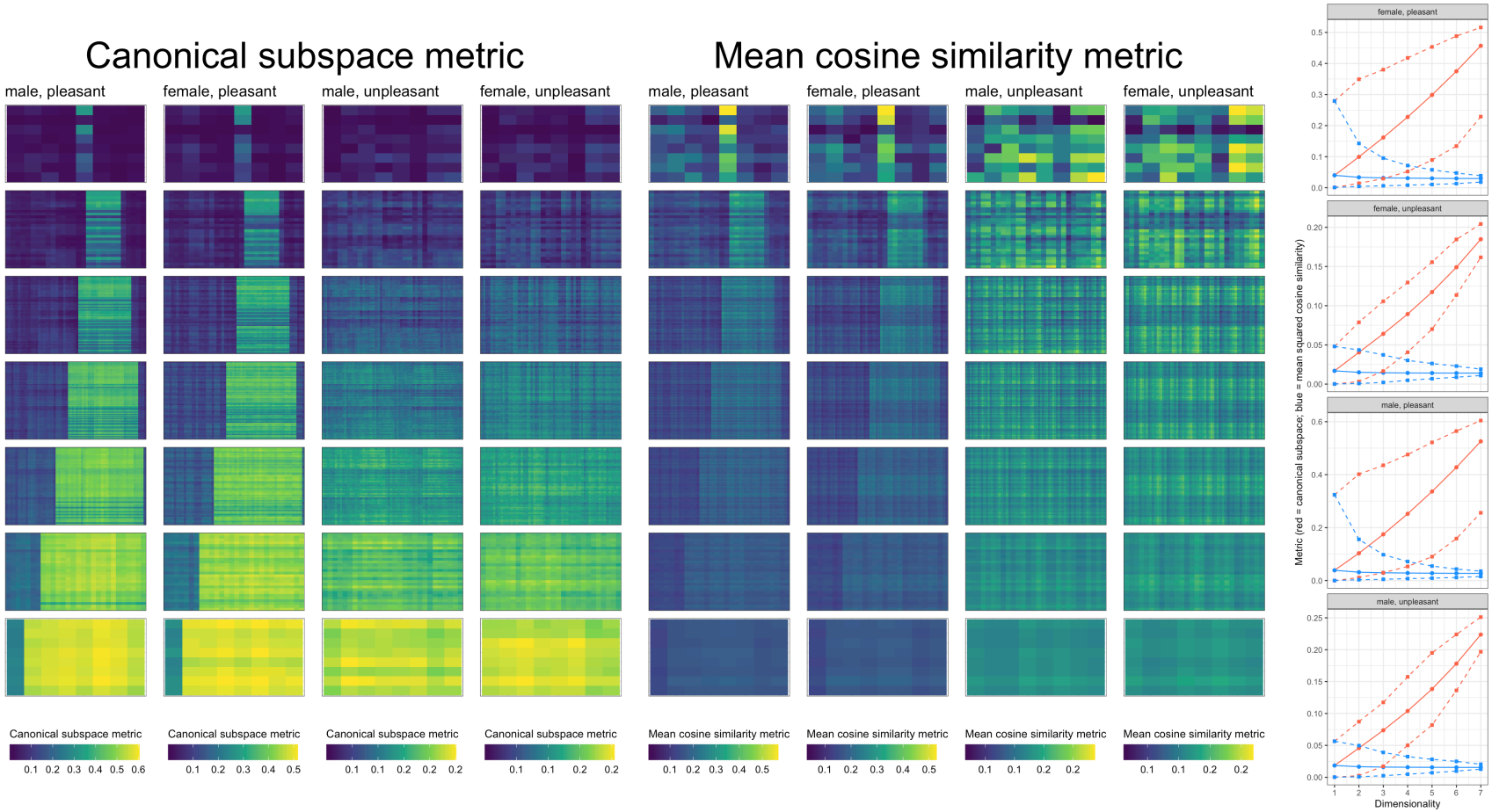
The cancellation induced by this design implies that the linear model fails to identify the desired interaction between the right-hand cosine and the left-hand normalization weights. The term γ_1 averages the three-way interactions of the right-hand cosine with the left-hand normalization weight (γ_0, γ_3); the right-hand inner product with the scale ratio (γ_5, γ_7), *or* the second-order interaction of the right (reciprocal) normalization weight and the interaction of the right inner product and the left normalization weight (γ_4, γ_6). But the interpretation of the variable this coefficient refers to is not the same across the lower-order terms. Some researchers employ a model of this type with an additional correlated cosine similarity regressor sharing the

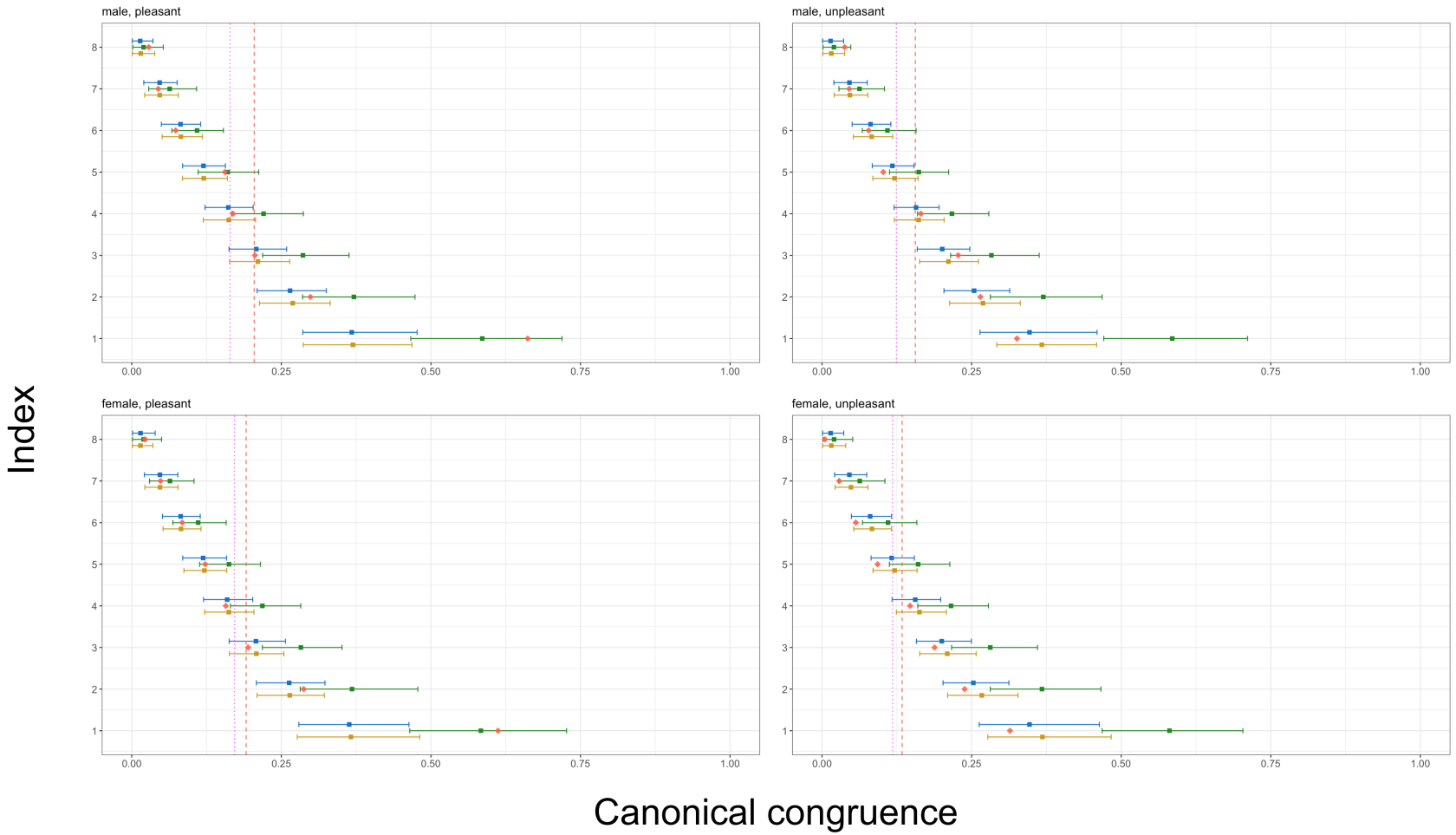
reference vector set X, leading to a trivariate (additive) composed cosine regression.

Figures

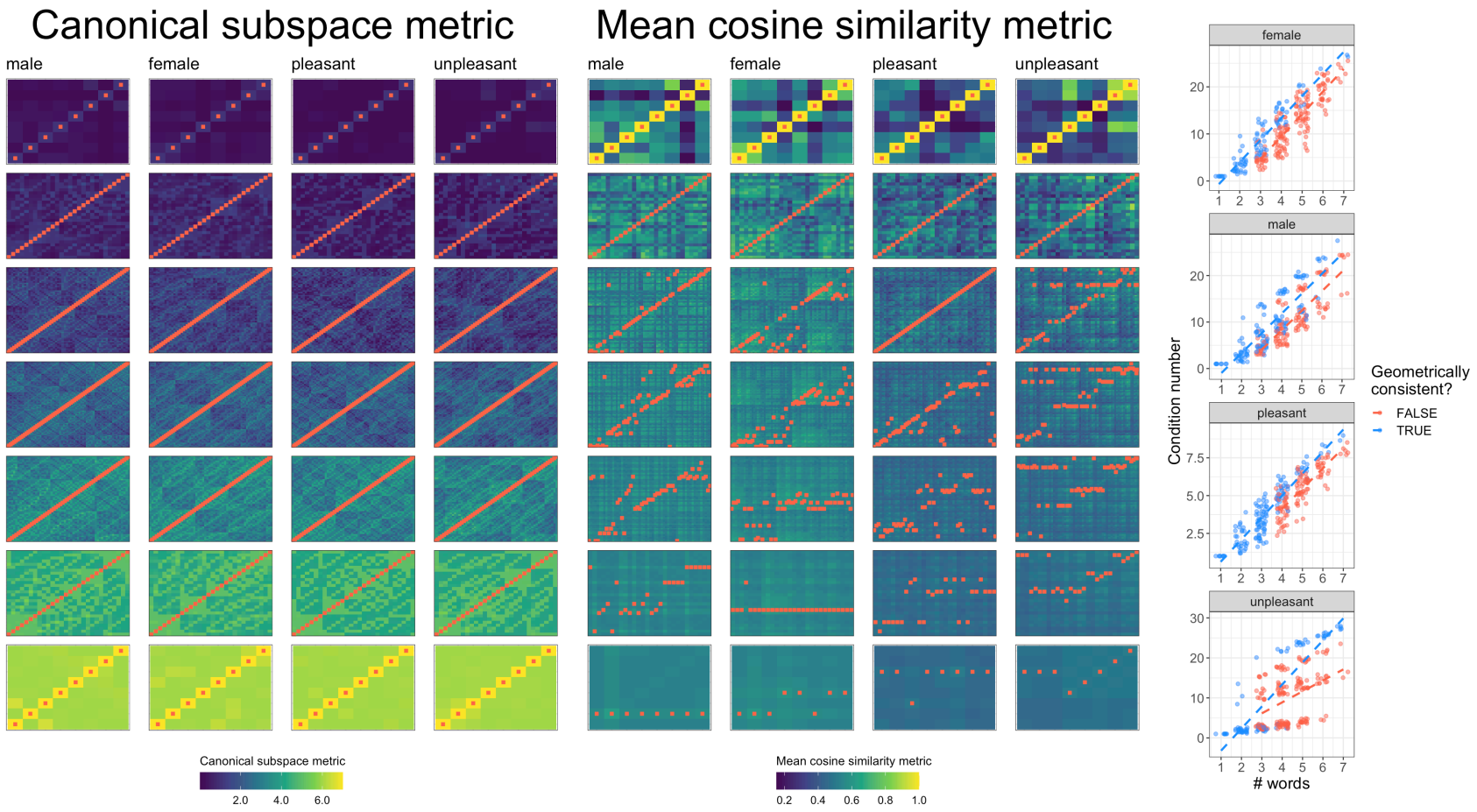
Google News word2vec

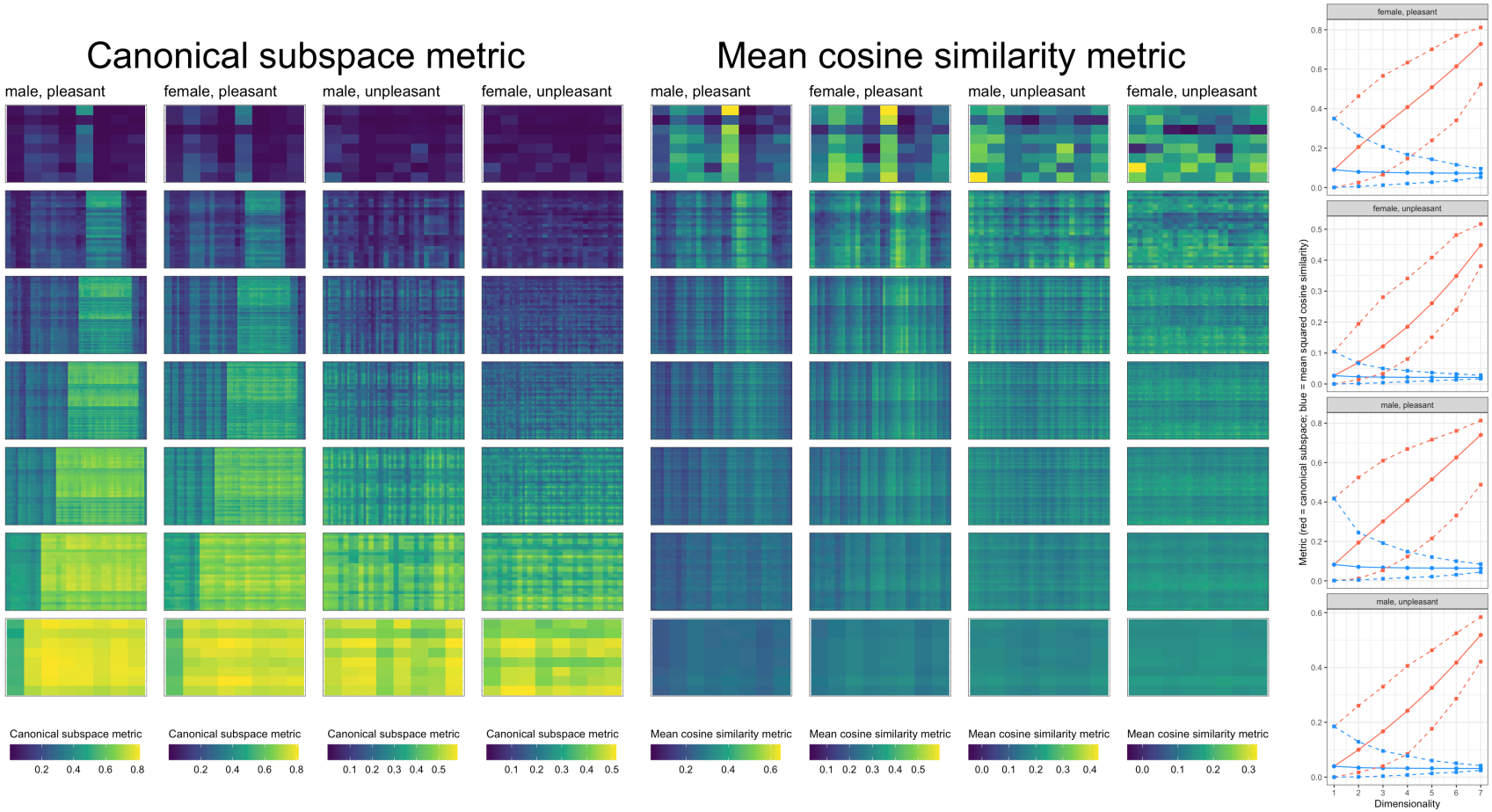

Canonical subspace metric
Mean cosine similarity metric
male
female
pleasant
unpleasant
Canonical subspace metric
2.0
4.0
6.0
Mean cosine similarity metric
0.2
0.5
0.8
1.0
Condition number
words
Geometrically consistent?
FALSE
TRUE

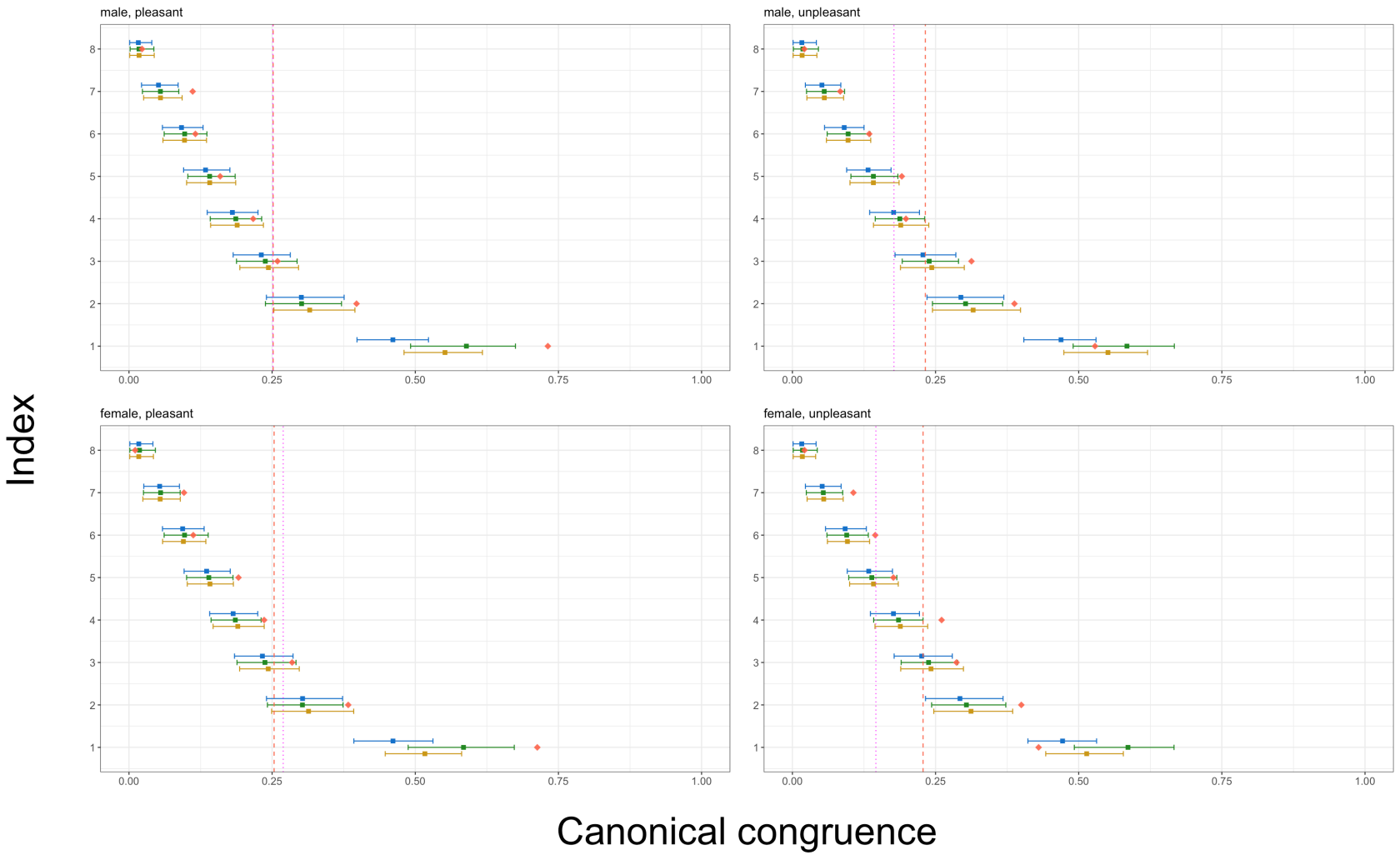

Canonical subspace metric
Mean cosine similarity metric
male, pleasant
female, pleasant
male, unpleasant
female, unpleasant
male, pleasant
female, pleasant
male, unpleasant
female, unpleasant
Canonical subspace metric
0.1 0.2 0.3 0.4 0.5 0.6
Canonical subspace metric
0.1 0.2 0.3 0.4 0.5
Canonical subspace metric
0.1 0.1 0.2 0.2 0.2
Canonical subspace metric
0.1 0.1 0.2 0.2
Mean cosine similarity metric
0.1 0.2 0.3 0.4 0.5
Mean cosine similarity metric
0.1 0.2 0.3 0.4 0.5
Mean cosine similarity metric
0.1 0.1 0.2 0.2
Mean cosine similarity metric
0.1 0.1 0.2 0.2
female, pleasant
female, unpleasant
male, pleasant
male, unpleasant
Metric (red = canonical subspace; blue = mean squared cosine similarity)
Dimensionality


male, pleasant
male, unpleasant
female, pleasant
female, unpleasant
Index
Canonical congruence
0.00
0.25
0.50
0.75
1.00
1
2
3
4
5
6
7
8

Stanford NLP GloVe


Canonical subspace metric
Mean cosine similarity metric
male
female
pleasant
unpleasant
Canonical subspace metric
2.0
4.0
6.0
Mean cosine similarity metric
0.2
0.4
0.6
0.8
1.0
Condition number
words
Geometrically consistent?
FALSE
TRUE


Canonical subspace metric
Mean cosine similarity metric
male, pleasant
female, pleasant
male, unpleasant
female, unpleasant
male, pleasant
female, pleasant
male, unpleasant
female, unpleasant
Canonical subspace metric
0.2 0.4 0.6 0.8
Canonical subspace metric
0.2 0.4 0.6 0.8
Canonical subspace metric
0.1 0.2 0.3 0.4 0.5
Canonical subspace metric
0.1 0.2 0.3 0.4 0.5
Mean cosine similarity metric
0.2 0.4 0.6
Mean cosine similarity metric
0.1 0.2 0.3 0.4 0.5
Mean cosine similarity metric
0.0 0.1 0.2 0.3 0.4
Mean cosine similarity metric
0.0 0.1 0.2 0.3
female, pleasant
female, unpleasant
male, pleasant
male, unpleasant
Metric (red = canonical subspace; blue = mean squared cosine similarity)
Dimensionality
1 2 3 4 5 6 7


male, pleasant
male, unpleasant
female, pleasant
female, unpleasant
Index
Canonical congruence
0.00
0.25
0.50
0.75
1.00
1
2
3
4
5
6
7
8

Tables

Table S1: Gender-sentiment analysis in main paper.

Target concept	W_x
male	male, man, boy, brother, he, him, his, son
female	female, woman, girl, sister, she, her, hers, daughter
pleasant	joy, love, peace, wonderful, pleasure, friend, laughter, happy
unpleasant	agony, terrible, horrible, nasty, evil, war, awful, failure

Table S2: Additional keyword lists. Words are lowercased when this is called for by the input word embeddings; further adjustments listed.

Label	Words	Adjustments
flowers	aster, clover, hyacinth, marigold, poppy, azalea, crocus, iris, orchid, rose, bluebell, daffodil, lilac, pansy, tulip, buttercup, daisy, lily, peony, violet, carnation, gladiolus, magnolia, petunia, zinnia	gladiola → gladiolus
insects	ant, caterpillar, flea, locust, spider, bedbug, centipede, fly, maggot, tarantula, bee, cockroach, gnat, mosquito, termite, beetle, cricket, hornet, moth, wasp, blackfly, dragonfly, horsefly, roach, weevil	
pleasant	caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation	
unpleasant	abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison	
instruments	bagpipe, cello, guitar, lute, trombone, banjo, clarinet, harmonica, mandolin, trumpet, bassoon, drum, harp, oboe, tuba, bell, fiddle, harpsichord, piano, viola, bongo, flute, horn, saxophone, violin	
weapons	arrow, club, gun, missile, spear, ax, dagger, harpoon, pistol, sword, blade, dynamite, hatchet, rifle, tank, bomb, firearm, knife, shotgun, teargas, cannon, grenade, mace, slingshot, whip	axe → ax
career	executive, management, professional, corporation, salary, office, business, career	
family	home, parents, children, family, cousins, marriage, wedding, relatives	
temporary	impermanent, unstable, variable, fleeting, short-term, brief, occasional	
permanent	stable, always, constant, persistent, chronic, prolonged, forever	
math	math, algebra, geometry, calculus, equations, computation, numbers, addition	
arts	poetry, art, dance, literature, novel, symphony, drama, sculpture	
science	science, technology, physics, chemistry, Einstein, NASA, experiment, astronomy	
arts	poetry, art, Shakespeare, dance, literature, novel, symphony, drama	
male	brother, father, uncle, grandfather, son, he, his, him	
female	sister, mother, aunt, grandmother, daughter, she, hers, her	

mental illness	sad, hopeless, gloomy, tearful, miserable, depressed	
physical illness	sick, illness, influenza, disease, virus, cancer	
male names	John, Paul, Mike, Kevin, Steve, Greg, Jeff, Bill	
female names	Amy, Joan, Lisa, Sarah, Diana, Kate, Ann, Donna	
white names	Brad, Brendan, Geoffrey, Greg, Brett, Jay, Matthew, Neil, Todd, Allison, Anne, Carrie, Emily, Jill, Kristen, Meredith, Sarah	deleted Laurie
Black names	Darnell, Hakim, Jermaine, Kareem, Jamal, Leroy, Rasheed, Tremayne, Tyrone, Aisha, Ebony, Keisha, Kenya, Latonya, Latoya, Tamika, Tanisha	deleted Lakisha
young names	Tiffany, Michelle, Cindy, Kristy, Brad, Eric, Joey, Billy	
old names	Ethel, Bernice, Gertrude, Agnes, Cecil, Wilbert, Mortimer, Edgar	
white names	Adam, Harry, Josh, Roger, Alan, Frank, Ian, Justin, Ryan, Andrew, Fred, Jack, Matthew, Stephen, Brad, Greg, Paul, Todd, Brandon, Hank, Jonathan, Peter, Wilbur, Amanda, Courtney, Heather, Melanie, Sara, Katie, Meredith, Shannon, Betsy, Donna, Kristin, Nancy, Stephanie, Ellen, Lauren, Colleen, Emily, Megan, Rachel	deleted Chip, Jed, Crystal, Amber, Peggy, Wendy, Bobbie-Sue, Sue-Ellen
Black names	Alonzo, Jamel, Lerone, Theo, Alphonse, Jerome, Leroy, Rashaan, Torrance, Darnell, Lamar, Lionel, Rashaun, Tyree, Deion, Lamont, Malik, Terrence, Tyrone, Lavon, Marcellus, Terrell, Wardell, Aisha, Nichelle, Shereen, Tamika, Ebony, Latisha, Shaniqua, Jasmine, Latonya, Shanice, Tanisha, Tia, Latoya, Sharice, Yolanda, Lashawn, Malika, Tawanda, Yvette	deleted Percell, Everol, Lashelle, Teretha, Tameisha, Lakisha, Shavonn, Tashika; Rasaan → Rashaan, Terryl → Terrell, Aiesha → Aisha, Temeka → Tamika, Shanise → Shanice, Sharise → Sharice, Lashandra → Lashawn

Supplementary references

1. P. L. Rodriguez, A. Spirling, B. M. Stewart, E. M. Wirsching, Multilanguage word embeddings for social scientists: Estimation, inference and validation resources for 157 languages (2023) (available at https://alcembeddings.org/).

2. A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases. *Science*. **356**, 183–186 (2017).

3. A. G. Greenwald, D. E. McGhee, J. L. Schwartz, Measuring individual differences in implicit cognition: The implicit association test. *Journal of personality and social psychology*. **74**, 1464 (1998).

4. J. Pennington, R. Socher, C. D. Manning, "GloVe: Global vectors for word representation" in *Proceedings of the 2014 conference on empirical methods in natural language processing EMNLP* (2014), pp. 1532–1543.

5. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*. **26** (2013).

6. G. Salton, Automatic text processing: The transformation, analysis, and retrieval of. *Reading: Addison-Wesley*. **169** (1989).

7. P. D. Turney, P. Pantel, From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*. **37**, 141–188 (2010).

8. J. W. Milnor, D. Husemoller, *Symmetric bilinear forms* (Springer, 1973), vol. 5.

9. M. Moakher, Means and averaging in the group of rotations. *SIAM journal on matrix analysis and applications*. **24**, 1–16 (2002).

10. B. Korth, L. R. Tucker, The distribution of chance congruence coefficients from simulated data. *Psychometrika*. **40**, 361–372 (1975).

11. J. H. Steiger, Tests for comparing elements of a correlation matrix. *Psychological bulletin*. **87**, 245 (1980).

12. A. G. Bedeian, A. A. Armenakis, W. A. Randolph, The significance of congruence coefficients: A comment and statistical test. *Journal of Management*. **14**, 559–566 (1988).

13. X.-L. Meng, R. Rosenthal, D. B. Rubin, Comparing correlated correlation coefficients. *Psychological bulletin*. **111**, 172 (1992).

14. C. F. Bond, K. Richardson, Seeing the fisher z-transformation. *psychometrika*. **69**, 291–303 (2004).

15. B. A. Nosek, M. R. Banaji, A. G. Greenwald, Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, research, and practice*. **6**, 101 (2002).

16. F. S. Bellezza, A. G. Greenwald, M. R. Banaji, Words high and low in pleasantness as rated by male and female college students. *Behavior Research Methods, Instruments, & Computers*. **18**, 299–303 (1986).

17. E. L. Thorndike, I. Lorge, The teacher's word book of 30,000 words. (1944).

18. W. F. Battig, W. E. Montague, Category norms of verbal items in 56 categories a replication and extension of the connecticut category norms. *Journal of experimental Psychology*. **80**, 1 (1969).

19. G. Firebaugh, J. P. Gibbs, User's guide to ratio variables. *American Sociological Review*, 713–722 (1985).

20. G. Firebaugh, J. P. Gibbs, Using ratio variables to control for population size. *Sociological Methods & Research*. **15**, 101–117 (1986).

21. R. A. Kronmal, Spurious correlation and the fallacy of the ratio standard revisited. *Journal of the Royal Statistical Society Series A: Statistics in Society*. **156**, 379–392 (1993).

22. R. P. Bartlett, F. Partnoy, The ratio problem. *Available at SSRN 3605606* (2020).

23. T. S. Breusch, A. R. Pagan, A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the econometric society*, 1287–1294 (1979).

24. R. D. Cook, S. Weisberg, Diagnostics for heteroscedasticity in regression. *Biometrika*. **70**, 1–10 (1983).

25. B. Western, D. Bloome, Variance function regressions for studying inequality. *Sociological Methodology*. **39**, 293–326 (2009).

26. M. Antoniak, D. Mimno, Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*. **6**, 107–119 (2018).

27. K. Ethayarajh, D. Duvenaud, G. Hirst, Understanding undesirable word embedding associations. *arXiv preprint arXiv:1908.06361* (2019).

28. A. van Loon, S. Giorgi, R. Willer, J. Eichstaedt, "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* (2022), vol. 16, pp. 1419–1424.

29. M. Casson, Linear regression with error in the deflating variable. *Econometrica: Journal of the Econometric Society*, 751–759 (1973).