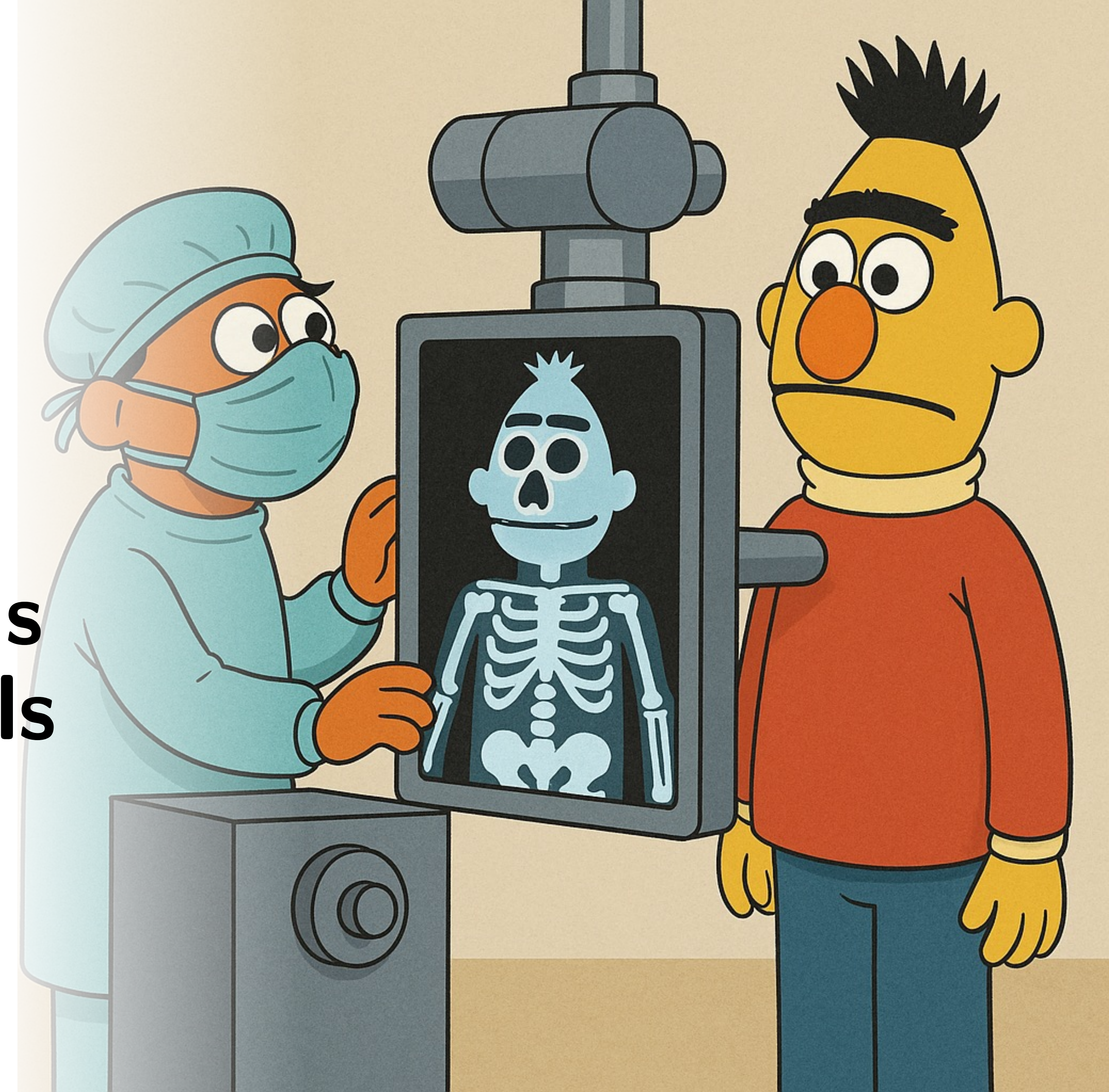


Diagnostic methods for language models

Alex Kindel
Assistant Professor of Sociology
Sciences Po médialab

SICSS Lake Como | 9 July 2025

Outline for today

- What is measurement?
- Social measurement and machine learning
- Measuring cultural associations in text
  - Word embeddings and cosine similarity
  - Mean cosine similarity
  - Anisotropy
  - GCD
  - Measuring geometric inconsistency in MCS models
- Keyword lists?

What is measurement?

What is measurement?

- Informally, a synonym for data collection procedures
- Fundamentally, the connection between “theory” and “data” – procedures identifying operations on numbers with theoretical relations of interest

Fundamental problem of measurement

- Assigning numbers to items in a set so that some **operation** on these numbers is isomorphic to a relation of interest
  1. Representation or isomorphism
     - What operations reflect real relations in the set?
  2. Uniqueness or invariance
     - What operations reflect our arbitrary choice of measurement procedure?

Validation and measurement

- **Validation** has a lot of parts:
  - Was it performed correctly?
  - Does it generalize to other settings/cases/iterations?
  - Does it predict relevant external criteria?
  - *Does it measure the specific kind of variation it is meant to measure?*
- Validation means evaluating how well a proposed measure represents what it is **supposed to represent** (Cronbach & Meehl 1955)
  - This becomes a lot harder when we don't have an unambiguous reference measure, for example if the target relation is **latent**.

Social measurement and machine learning

How do we measure social life?

1. participating in it
   - cultivating interpersonal relationships so that complex, difficult, long-term, or otherwise hard-to-survey social processes become visible
2. asking structured, general questions to samples of populations
   - opinion surveys; likely voter polls; educational tests; longitudinal surveys; ...
3. observing social interactions using the Internet
   - social media posts; clickstream data; traditional media archives

Machine learning

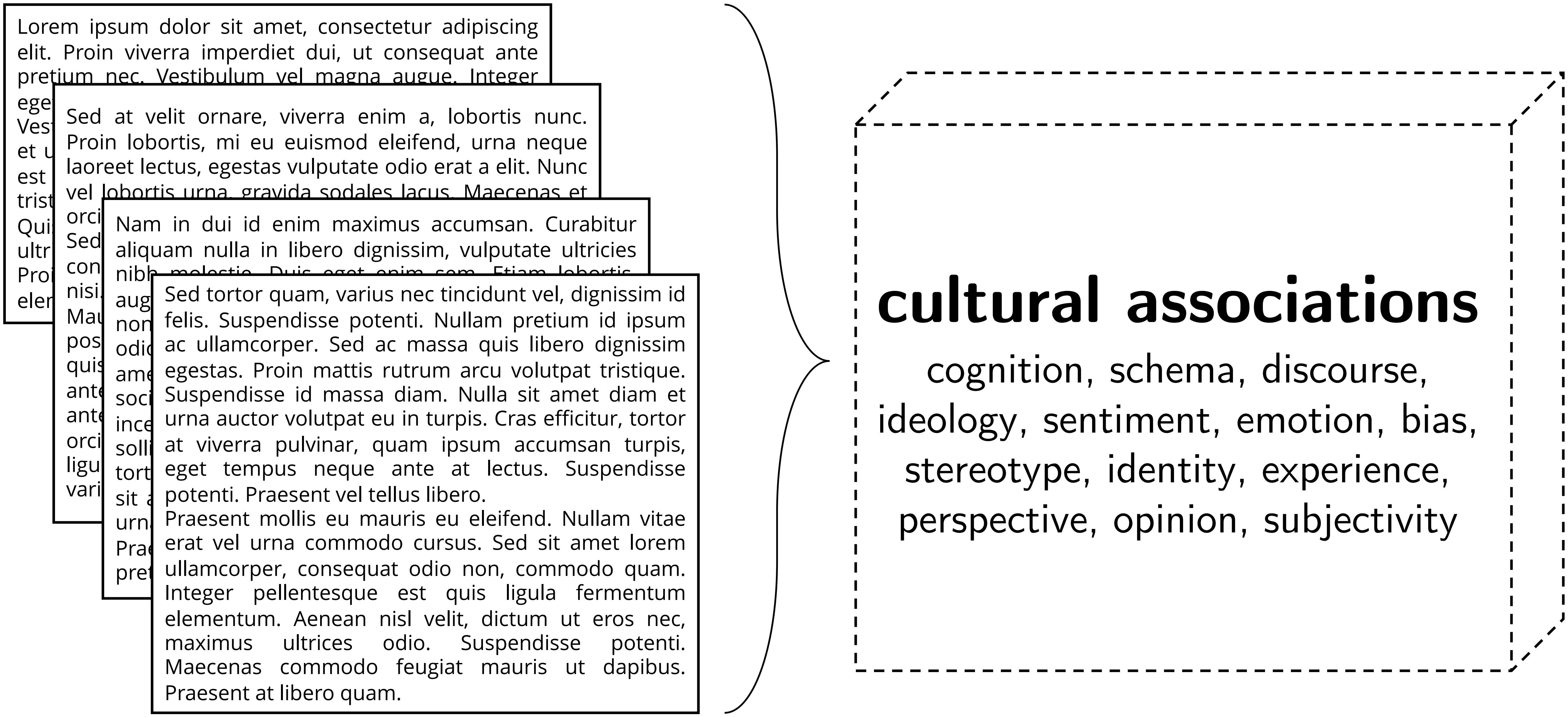
- A subfield of computer science that builds prediction machines
- Enables fitting much more general functions to much more complex data than previously possible in traditional QSS
    - Hope in social science is that we can use the **latent representations** learned by such models to study social phenomena
- But validation in ML focuses entirely on **criterion validity**: goal is to produce a model that is the best at every "benchmark" (Raji et al. 2021)
    - "Artificial general intelligence"
    - In practice most social science applications of ML try to convert their research question into a benchmarking task setup

Validation beyond benchmarking

- Today I'll focus on characterizing the representation and invariance properties of **a simple language model-based measure of association**.

Measuring cultural associations in text

“Cosine similarity and keyword selection in multidimensional semantic analysis.”
(R&R, *Political Analysis*)



Lorem ipsum dolor sit amet, consectetur adipiscing elit. Proin viverra imperdiet dui, ut consequat ante pretium nec. Vestibulum vel magna augue. Integer
ege
Vest
et u
est
trist
Qui
ultr
Proi
elen

Sed at velit ornare, viverra enim a, lobortis nunc. Proin lobortis, mi eu euismod eleifend, urna neque laoreet lectus, egestas vulputate odio erat a elit. Nunc vel lobortis urna, gravida sodales lacus. Maecenas et
orci
Sed
con
nisi.
Mau
pos
quis
ante
ante
orci
ligu
vari

Nam in dui id enim maximus accumsan. Curabitur aliquam nulla in libero dignissim, vulputate ultricies nibh molestie. Duis eget enim sem. Etiam lobortis
aug
non
odio
ame
soci
ince
solli
tort
sit a
urna
Prae
pret

Sed tortor quam, varius nec tincidunt vel, dignissim id felis. Suspendisse potenti. Nullam pretium id ipsum ac ullamcorper. Sed ac massa quis libero dignissim egestas. Proin mattis rutrum arcu volutpat tristique. Suspendisse id massa diam. Nulla sit amet diam et urna auctor volutpat eu in turpis. Cras efficitur, tortor at viverra pulvinar, quam ipsum accumsan turpis, eget tempus neque ante at lectus. Suspendisse potenti. Praesent vel tellus libero.

Praesent mollis eu mauris eu eleifend. Nullam vitae erat vel urna commodo cursus. Sed sit amet lorem ullamcorper, consequat odio non, commodo quam. Integer pellentesque est quis ligula fermentum elementum. Aenean nisl velit, dictum ut eros nec, maximus ultrices odio. Suspendisse potenti. Maecenas commodo feugiat mauris ut dapibus. Praesent at libero quam.

cultural associations

cognition, schema, discourse, ideology, sentiment, emotion, bias, stereotype, identity, experience, perspective, opinion, subjectivity

Word embedding association tests (WEAT)

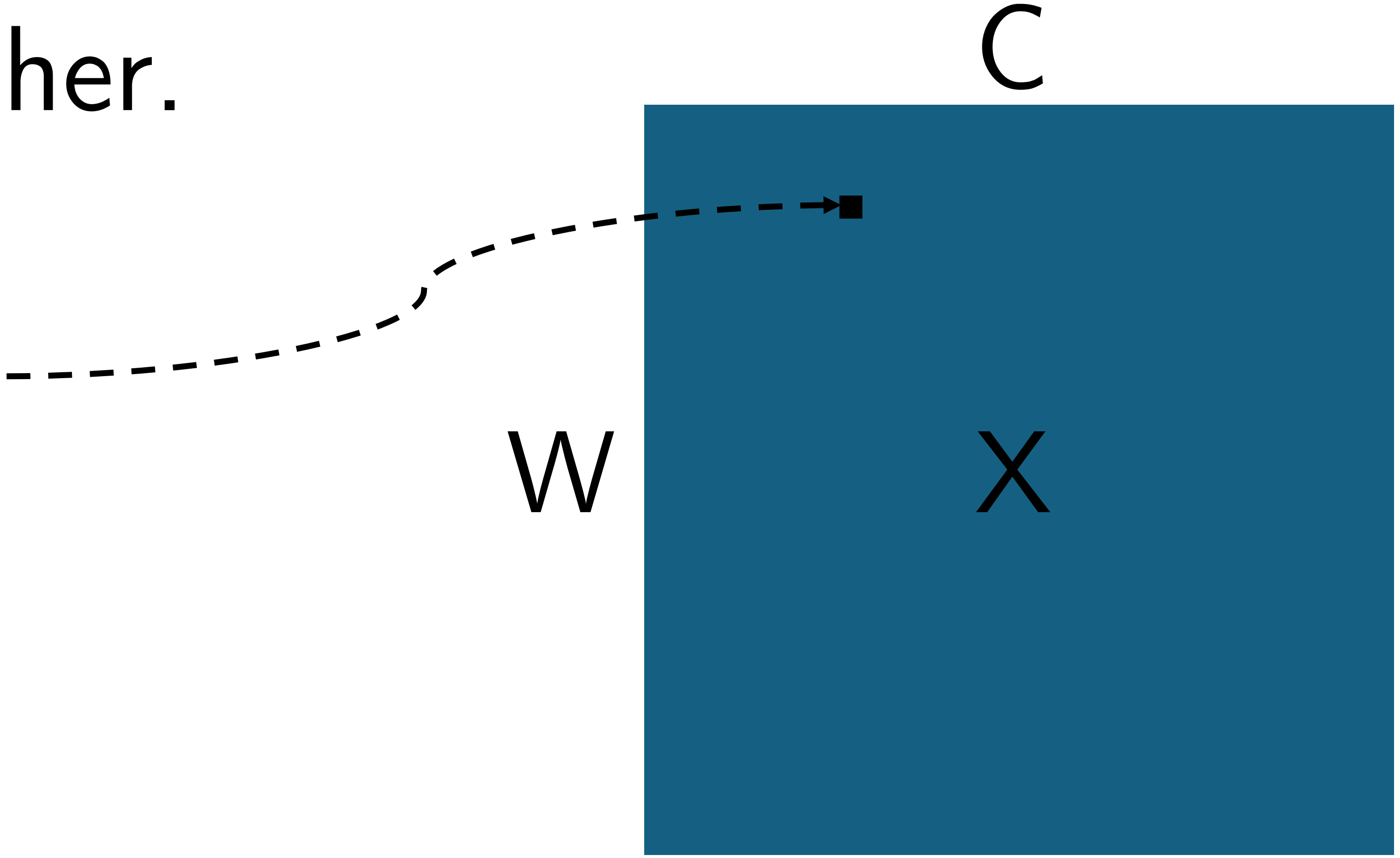
- “Semantics derived automatically from language corpora contain human-like biases” (Caliskan, Bryson and Narayanan 2017)
  - Aims to “replicate” findings from the Implicit Association Test in social psychology (an experimental test of differential word association) using **word embeddings**
  - To do this, authors propose a measure of association between concepts based on paired lists of keywords for each concept, then correlate this measure with experimental IAT results (it’s high!)

Word embeddings

- A **geometric model**: assign each word to a **position in multidimensional space** so that words with more similar meanings are **closer together in direction.**

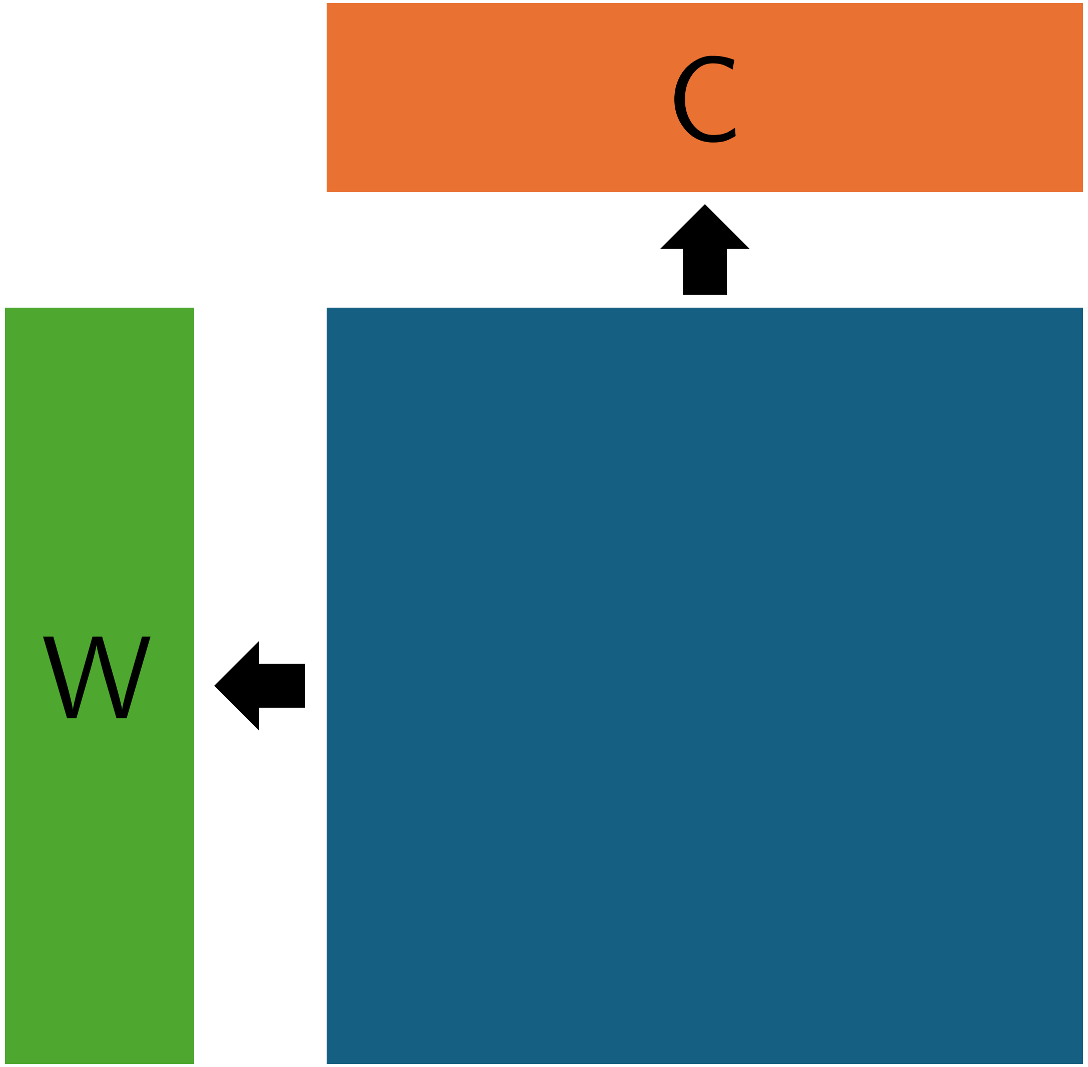
Word co-occurrence measures quantify how often pairs of words occur near each other.

$$X_{ij} = f(w,\ c)$$

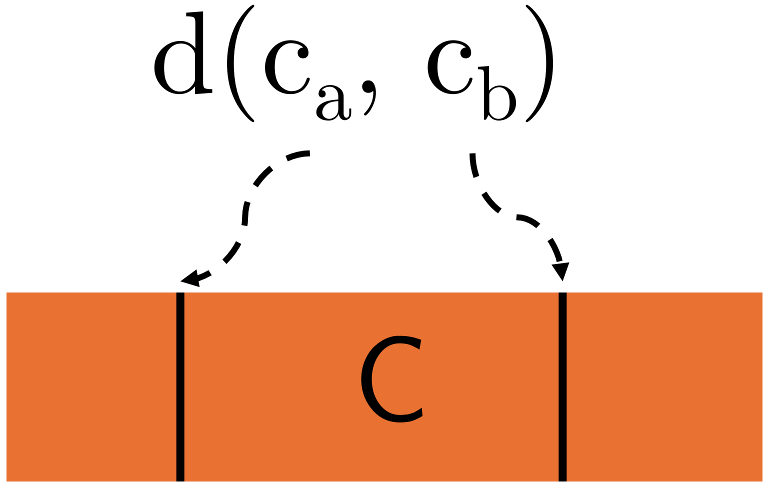


$$\log \frac{p(w,c)}{p(w)p(c)}$$

Word embeddings efficiently estimate word co-occurrence measures.



Word association functions compare the co-occurrence distributions of word pairs



Cosine similarity

Comparing directions

- We use a measure of angular proximity: **cosine similarity**
- This is just the **Pearson product-moment correlation coefficient**!
  - The only difference is that cosine similarity is **not recentered.**

dot product

$$\theta_{ab} = \frac{a \cdot b}{||a||\ ||b||}$$

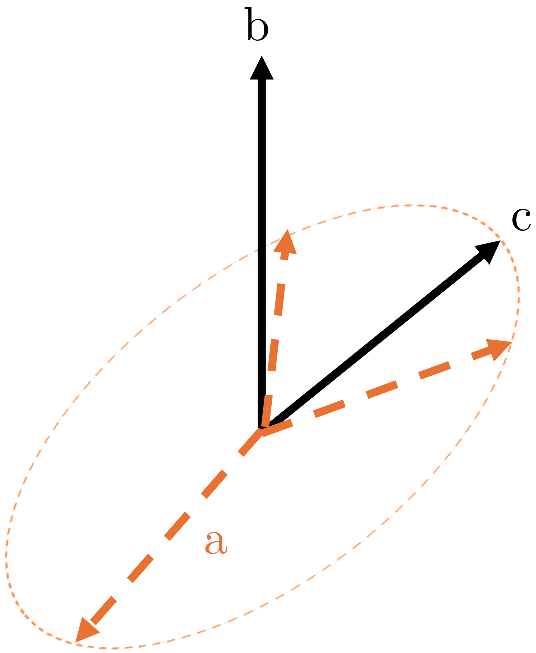
cosine similarity

norm product

(counterfactual independence measure)

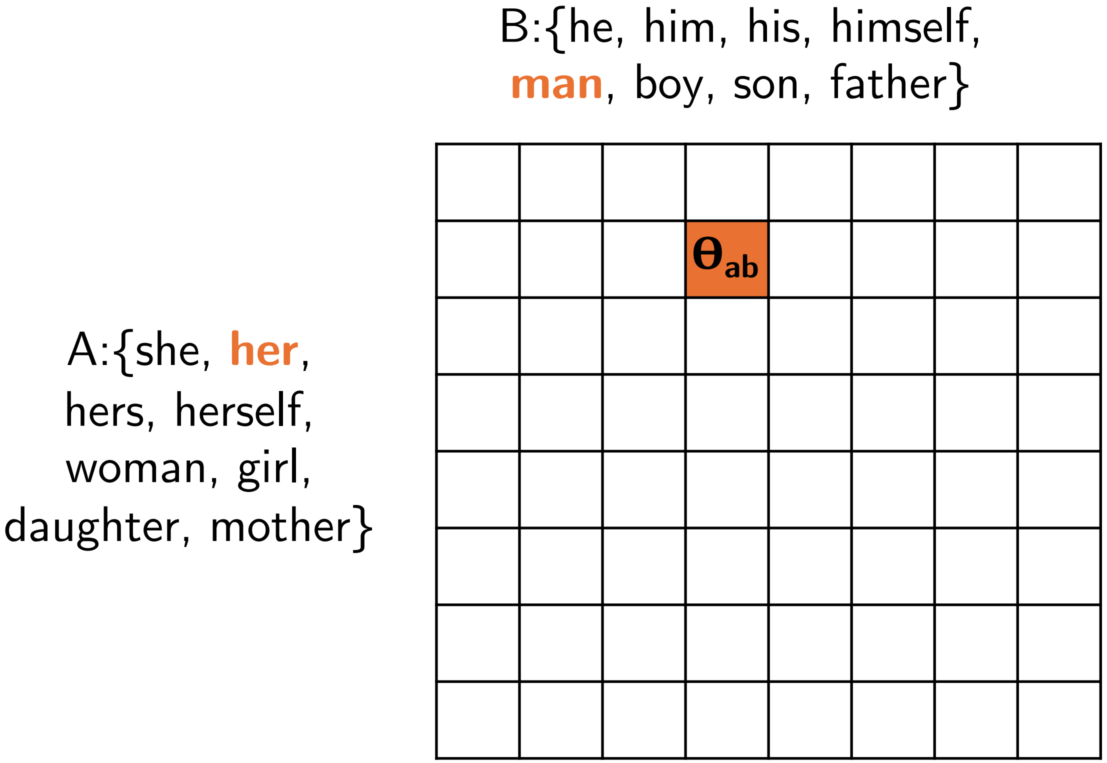
Range restriction

- For two isotropically distributed random vectors, cosine similarity **varies between -1 and 1**, and the expected value is 0.
- However, if we have more than two random vectors, and *some of the pairwise cosines are known*, the remaining cosines that can be constructed from this set are in general not supported on [-1, 1].
- This stems from the fact that **matrices of cosine similarities are positive (semi-)definite.**



The **attainable range** of cosine similarity depends on the input vectors.

Cross-cosine matrix



B:{he, him, his, himself, **man**, boy, son, father}

A:{she, **her**, hers, herself, woman, girl, daughter, mother}

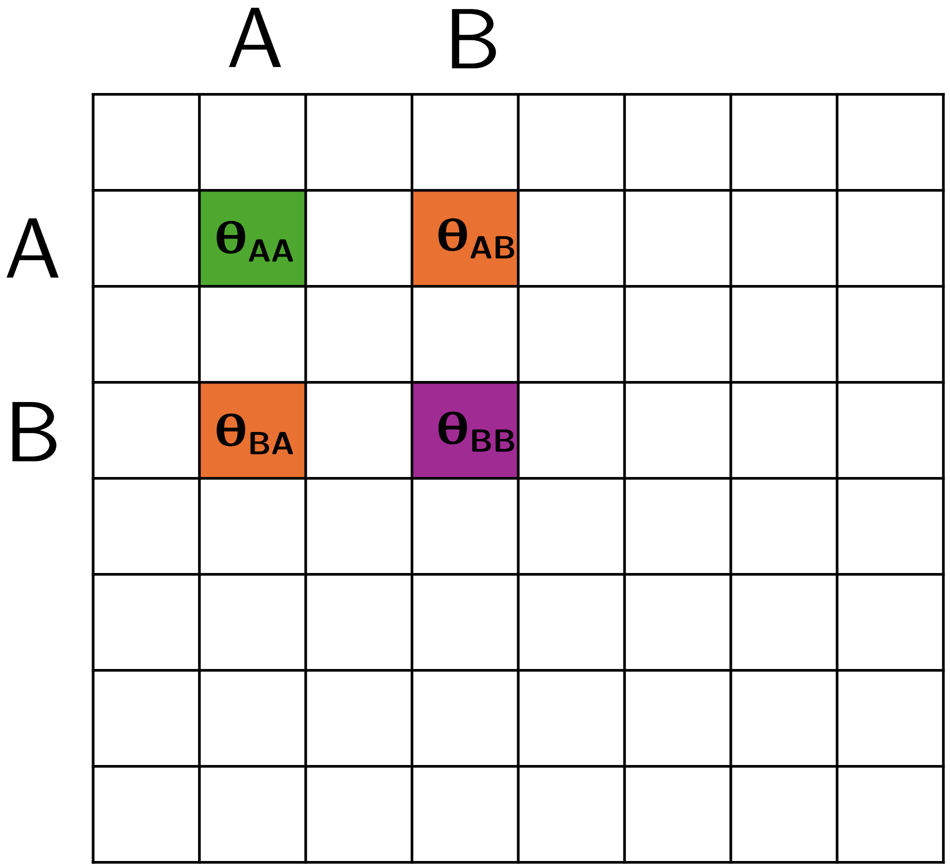
Cosine matrix

A:{she, her, hers, **herself**, woman, girl, daughter, mother}

A:{she, **her**, hers, herself, woman, girl, daughter, mother}

1							
	1		$\boldsymbol{\theta}_{\mathbf{aa}}$				
		1					
			1				
				1			
					1		
						1	
							1

Global cosine matrix



$$\Theta_{XX} = \begin{bmatrix} \Theta_{AA} & \Theta_{AB} \\ \Theta_{AB}^{T} & \Theta_{BB} \end{bmatrix}$$

Sylvester's criterion

- A symmetric matrix M is positive definite if and only if all of its principal minors (determinants of upper-left blocks) are positive.
  - **Positive definite** implies det(M) $>$ 0.
- Since A (B) is drawn **before** we make any comparison, the entries of Θ_{XX} are constrained by our choice of Θ_{AA} (Θ_{BB}):
  - By the Schur complement theorem, $\det(\Theta_{\mathrm{XX}}) = \det(\Theta_{\mathrm{AA}})\det(\Theta_{\mathrm{BB}} - {\Theta_{\mathrm{AB}}}^T {\Theta_{\mathrm{AA}}}^{-1}\Theta_{\mathrm{AB}}) > 0$.
  - $\det(\Theta_{\mathrm{AA}}) > 0$ since Θ_{AA} is also SPD, so $\det(\Theta_{\mathrm{BB}} - {\Theta_{\mathrm{AB}}}^T {\Theta_{\mathrm{AA}}}^{-1}\Theta_{\mathrm{AB}}) > 0$.

$$\Theta_{AB}^T \Theta_{AA}^{-1} \Theta_{AB} \preccurlyeq \Theta_{BB}$$

Multidimensional range restriction

$$\Theta_{BB}^{-1/2}\Theta_{AB}^{T}\Theta_{AA}^{-1}\Theta_{AB}\Theta_{BB}^{-1/2} \preccurlyeq I_k$$

Mean cosine similarity

Problem setup

- Let A and B be two keyword lists.
    - Assume A and B are disjoint (no words in common) and deduplicated (no repeated words).
    - Assume that $|A|, |B| < d$.
- These assumptions guarantee that the vectors in $\{A, B\}$ are linearly independent and their cosine matrices Θ_{AA} and Θ_{BB} are positive definite.

Mean cosine similarity (“MCS”)

$$d_{\mathrm{MCS}}(A, B) = \sum_{i}^{k} \sum_{j}^{k} \eta_{ij} \cos(A_i, B_j) = \mathrm{mean}(\boldsymbol{\Theta}_{\mathbf{AB}})$$

Expected cosine similarity between any pair of words in A and B *when the vectors are isotropically distributed*.

Cross-cosine matrix

B:{he, him, his, himself,
man, boy, son, father}

A:{she, **her**,
hers, herself,
woman, girl,
daughter, mother}

			θ_{ab}				

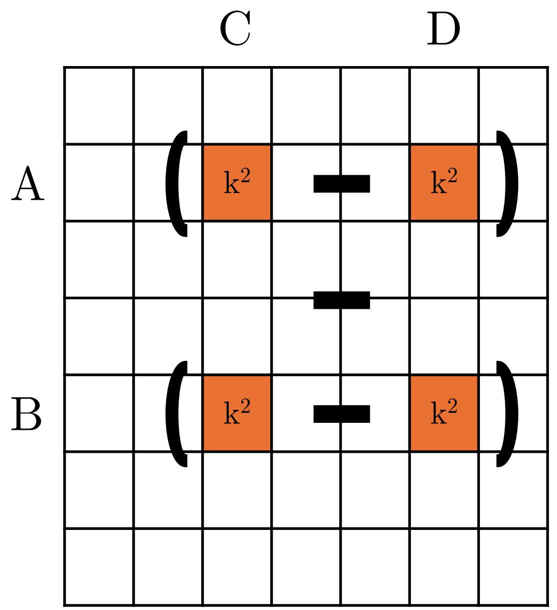
$$d_{\mathrm{MCS}}(A, B) = \mathrm{mean}(\theta_{AB})$$

Word Embedding Association Test (WEAT)

$$\mathrm{WEAT(A,B;C,D)} = \begin{array}{c}(\mathrm{MCS(A,\ C)} - \mathrm{MCS(A,\ D)}) \\ - \\ (\mathrm{MCS(B,\ C)} - \mathrm{MCS(B,\ D)})\end{array}$$

$$\eta_{ij} = 1/\mathrm{k}\ (\mathrm{or}\ 1/\mathrm{k}^2)$$

Two-way difference in association between pairs of "opposing" concepts; e.g. $\{$***Masculinity***, ***Femininity***; ***Good***, ***Bad***$\}$



(Caliskan, Bryson and Narayanan. "Semantics derived automatically from language corpora contain human-like biases." (*Science*, 2017)

The MCS model family

Extension 1: Covariates

e.g. author, year, cos(x, y), n(a)

$$\eta_{ij} = \mathrm{p}(\cos(\mathrm{A_i},\, \mathrm{C_j})|\mathrm{X})$$

Extension 2: semantic “axes”

$$\cos(\bar{\mathrm{A}} - \bar{\mathrm{B}},\, \bar{\mathrm{C}} - \bar{\mathrm{D}})$$

$$\eta_{ij} = \frac{||\mathrm{A_i}||\ ||\mathrm{C_j}||}{||\bar{\mathrm{A}} - \bar{\mathrm{B}}||\ ||\bar{\mathrm{C}} - \bar{\mathrm{D}}||}$$

Kozlowski, Taddy and Evans, “The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings.” (*American Sociological Review*, 2019)

$$\mathrm{mean}(\texttt{word association}) = \textit{cultural association}$$

Under what conditions can we say that this identity holds?

Anisotropy

Insight: Keywords are highly interrelated.

- Keyword selection induces sets of words that are **much more closely related to one-another than we would expect** on average.
- If we draw sets of 8 words uniformly at random, we don't get this:
  - apple, banana, cherry, raspberry, blueberry, watermelon, guava, grape
  - car, bus, train, bicycle, airplane, motorcycle, skateboard, scooter
- We'd like a **diagnostic measure** for how unusually related lists like this are in terms of their vector representations.

An anisotropy diagnostic

- A set of word vectors A is **isotropic** if and only if its cosine matrix Θ_{AA} is equal to the identity matrix.
  - This is a slightly weaker condition than requiring the **covariance matrix** to be the identity matrix; we just ignore the norms of the vectors here.
- For each keyword list A, we would like to measure how far the observed cosine matrix Θ_{AA} is from the order k identity matrix I_k.

Measuring distance on SPD(k)

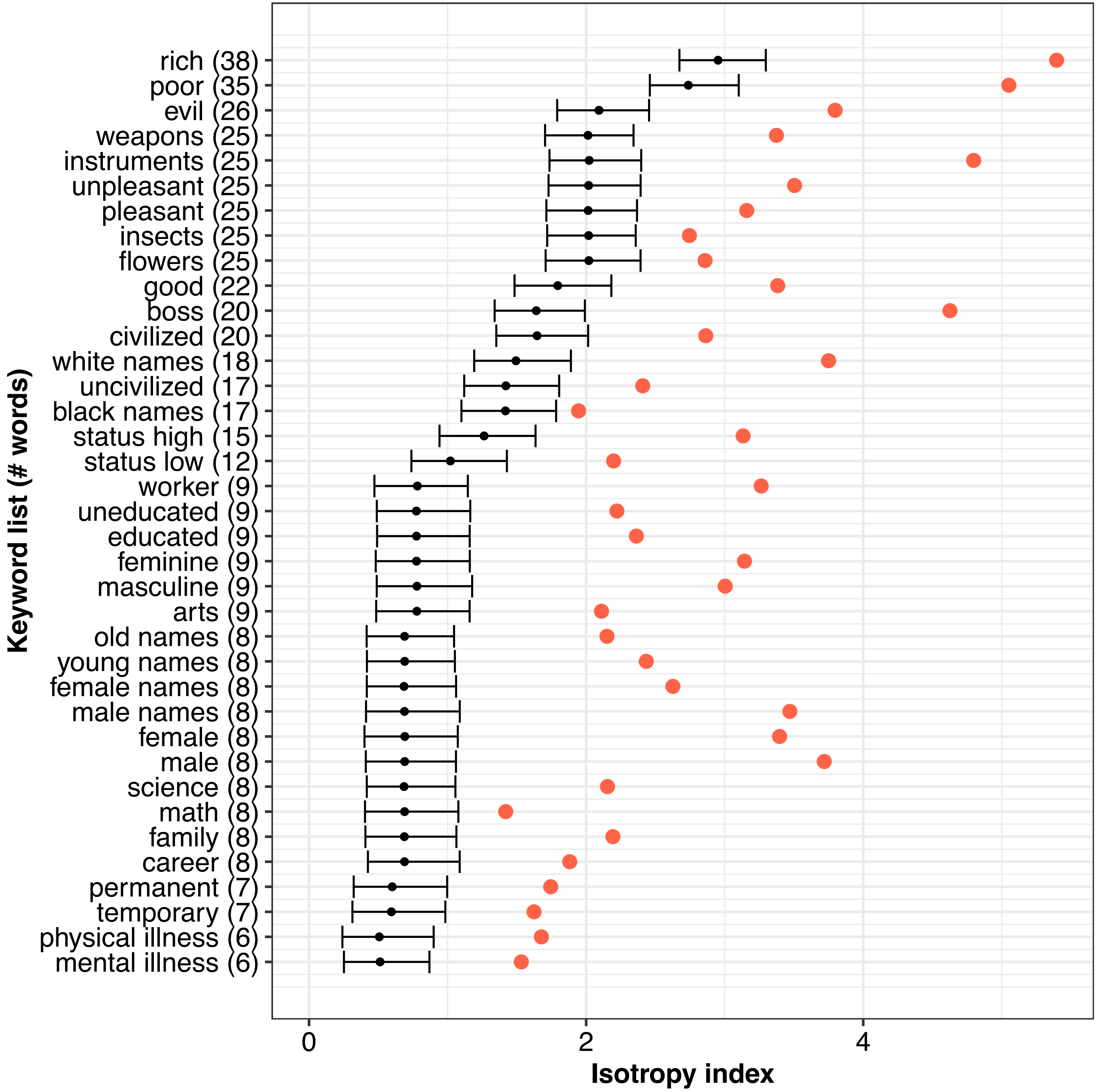
- The set of order k symmetric positive definite (SPD) matrices is an incomplete metric space under the Euclidean inner product Tr(AB).
  - Direct paths between two SPD matrices A and B in this metric space pass through non-SPD matrices.
- We instead use the Riemannian metric:
  - $\delta^2(A, B) = ||\mathrm{Log}(A^{-1/2}BA^{-1/2})||_F^2$
  - $A^{-1/2}$ is the inverse of the matrix square root of A
  - Log(X) is the matrix logarithm of X
- SPD(k) denotes the manifold of all order k covariance matrices under $\delta^2(A, B)$.

Finding the nearest isotropic matrix

- $\varphi^* = \text{argmin}_\varphi\ \delta^2(\theta_{AA},\ \varphi I_k)$.
  - The least geodesic squares estimator is just the **geometric mean of the eigenvalues** of $\Theta_{AA,}$ $\varphi^* = \det(\Theta_{AA})^{1/k}$.
- Then project SPD(k) onto Corr(k), the subset of correlation matrices (i.e., the SPD matrices with diagonal entries equal to one).
  - An isometric projection is given by $\pi(A) = \text{diag}(A)^{-1/2}\, A\, \text{diag}(A)^{-1/2}$.
  - For diagonal A, $\pi^{-1}(A)$ is the set of scalar matrices φI_k.
  - So the geodesic distance between Θ_{AA} and I_k along Corr(k) is just equal to $\delta^2(\Theta_{AA},\ \varphi^* I_k)$.
- Call this quantity the **index of isotropy** of A, Iso(A).

Anisotropy in real keyword lists

- 37 keyword lists → 666 multidimensional comparisons
  - Lists range from 6 to 38 words
  - male, female, masculine, feminine, male names, female names, young names, old names, white names, black names, rich, poor, educated, uneducated, civilized, uncivilized, status high, status low, boss, worker, family, career, math, science, arts, mental illness, physical illness, temporary, permanent, weapons, instruments, flowers, insects, pleasant, unpleasant, good, evil
  - Compare Iso(X) to distribution under repeated random sampling of keyword lists of identical cardinality.

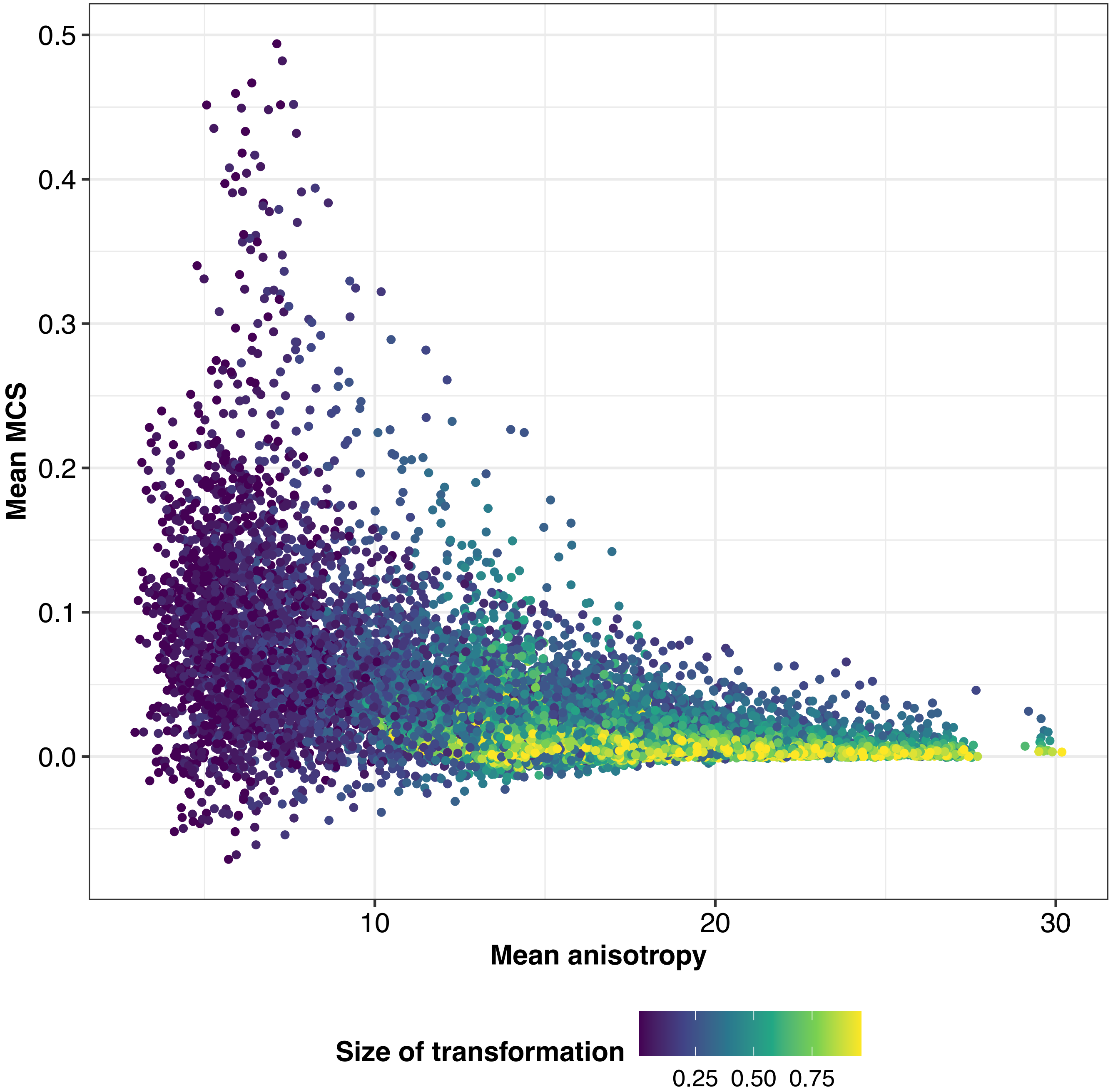


- Black bars cover 99% of isotropy indices for randomly sampled keyword lists of equal cardinality.
- Red dots indicate observed isotropy index for this keyword list.
- Keyword list selection induces unusually anisotropic subsets of word vectors.

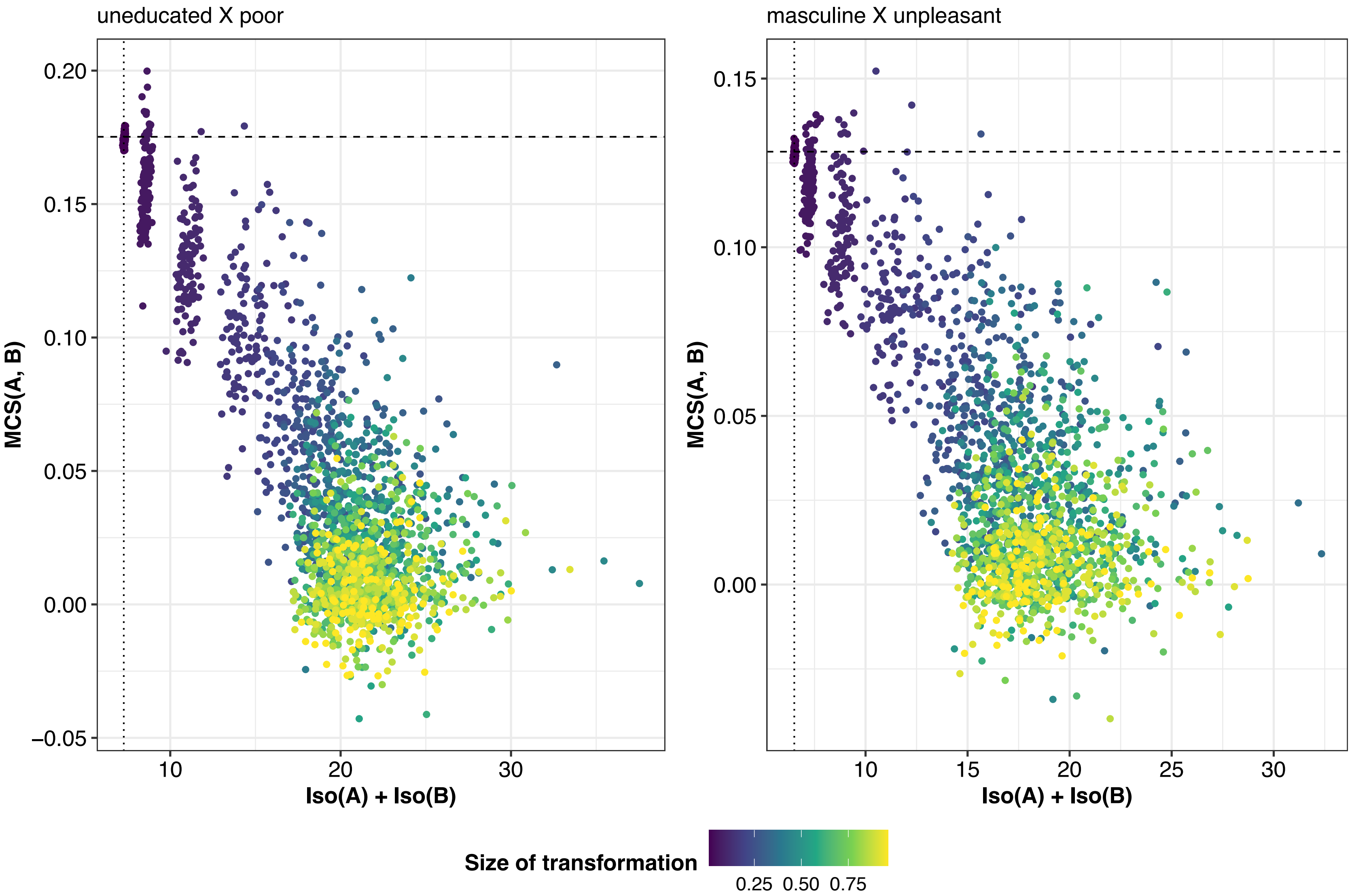
Effects of increasing anisotropy

- We can make A more anisotropic by applying an arbitrary nonsingular transformation M on the right to obtain $A^* = AM$.
- Here we choose M to be of the form $I + X$ where X_{ij} is drawn from a Gaussian distribution with mean zero and variance σ^2, so that σ^2 parameterizes the size of the transformation.

- Increasingly anisotropic transformations of A, B shrink MCS to zero.
- However…



Relatively small changes in anisotropy can make MCS(A, B) **either larger or smaller.**



GCD

An alternative quantity

- Let P_A, P_B be the **orthogonal projection matrices** onto the column space of A and B respectively.
  - GCD(A, B) = cos(P_A, P_B).
- Measures angular separation of the subspaces spanned by A and B.
  - The orthogonal projection matrices are the subset of the positive semidefinite matrices with eigenvalues equal to zero or one.
  - Since Pa and Pb are located on the cone of positive semidefinite matrices, cos(Pa, Pb) is strictly nonnegative; it behaves like a *squared* cosine similarity.
- Why the “generalized” coefficient of determination?
  - If $|A| = 1$ and $|B| = k$ then GCD(A, B) is the R^2 obtained by estimating the linear model $A = BX + e$ using OLS.

Why linear subspaces?

- If the signs of the input vectors were meaningful then we could use this to choose a sign for the square root of GCD.
- But in this setting the signs of the vectors are in general **arbitrary** since we can flip the sign of any vector in W or C by flipping its sign in the other matrix, and this doesn't matter as long as we ignore the second set of embeddings!
  - This is also why we might prefer not to add or subtract vector representations manually.

GCD is anisotropy invariant

- Two ways to see why this is true
  1. Orthogonal projection matrices are invariant to the anisotropy of the inducing vectors.
  2. GCD is the unique choice of MCS weights that is anisotropy invariant.

Orthogonal projection matrices are anisotropy invariant.

- $P_A = A(A^T A)^{-1} A^T$.
- Let M be any nonsingular matrix (i.e., its inverse M’ exists).
- $P_{AM} = AM(M^T A^T AM)^{-1} M^T A^T$
  $= AMM^{-1}(A^T A)^{-1} M^{-T} M^T A^T$
  $= A(A^T A)^{-1} A^T$.

GCD is the unique choice of MCS weights that is anisotropy invariant.

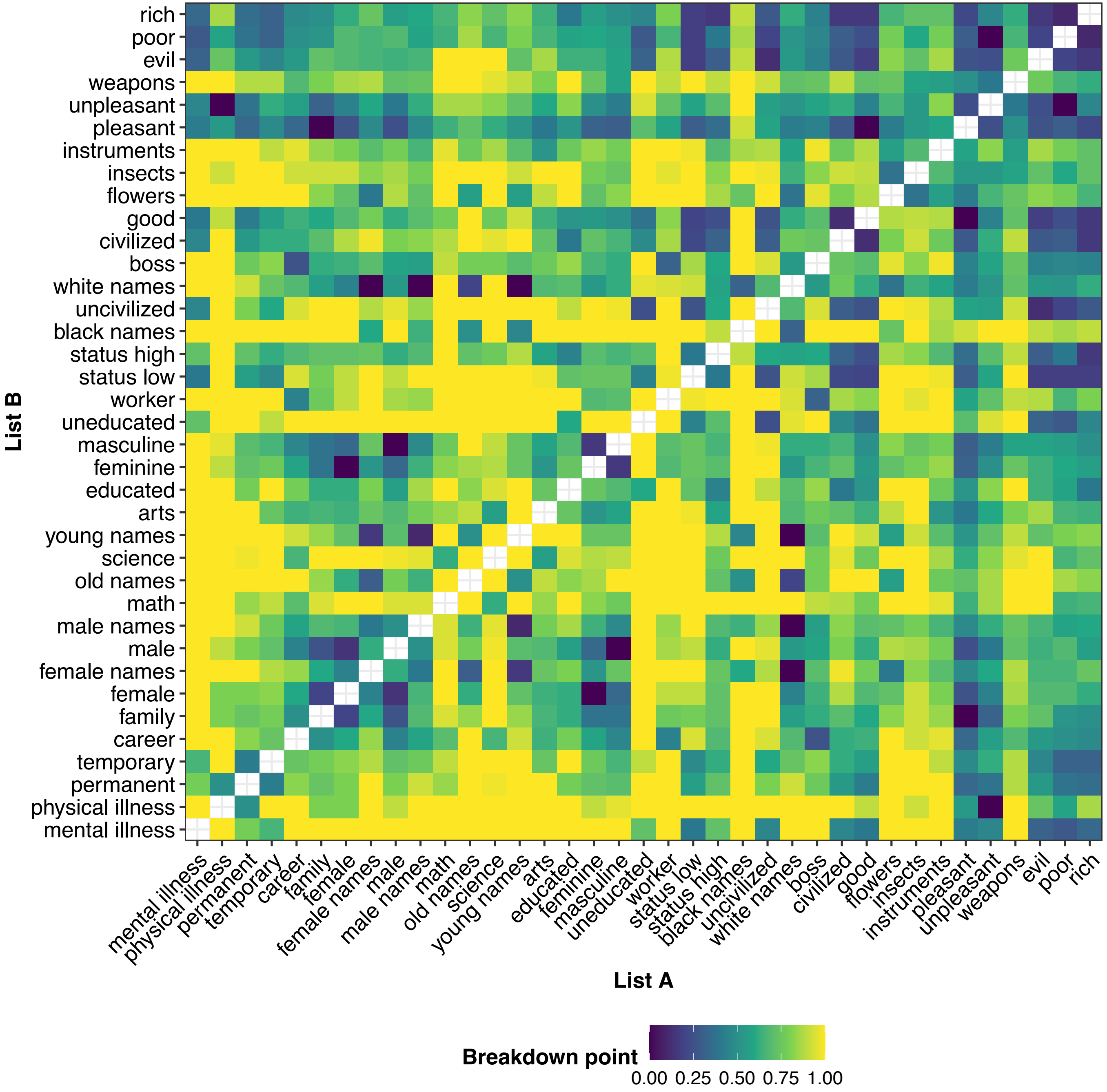
- GCD can be viewed as an MCS quantity for a specific choice of η.
- Unlike MCS, the entries of η can be **negative** for GCD.
  - This allows us to find a linear combination of weights that **cancels out redundant covariance** in the set of pairwise comparisons.
  - If we prefer to view the function as a weighted mean per se, we can instead view the signs as enforcing orientation consistency on the pairwise cosines.

Geometric inconsistency in MCS models

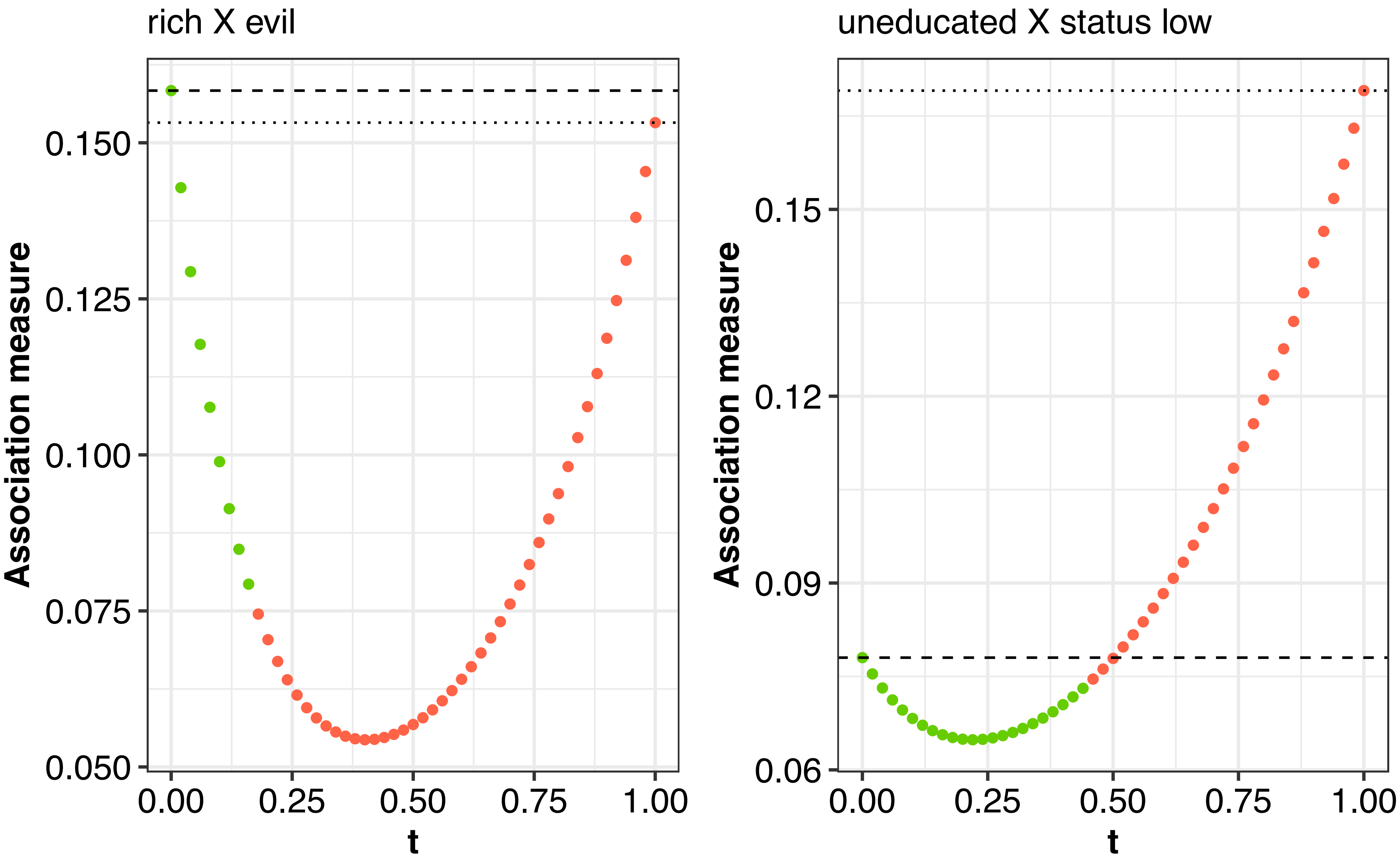
Geodesic interpolation

- To see how GCD relates to MCS, we can “relax” the isotropy assumption by interpolating the observed cosine matrices toward the identity matrix
- The geodesic interpolant between A and B along Corr(k) at time $t \in [0, 1]$ is given by the expression:
  - $\pi(A^{1/2}\mathrm{Exp}[t\ \mathrm{Log}(A^{-1/2}BA^{-1/2})]A^{1/2})$
- We call the maximum value of t for which the global cosine matrix is valid (i.e. remains positive definite) the **breakdown point** for the isotropy assumption.

- If we interpolate Θ_{AA} and Θ_{BB} toward the identity matrix along Corr(k), we can identify the point where the isotropy assumption becomes **geometrically impossible**.
- Only about 1 in 4 comparisons is compatible with the isotropy assumption.



Arithmetic comparisons between MCS (dotted line) and GCD (dashed line) generally underestimate the total **anisotropic distortion** in the former quantity.



$$\mathrm{mean}(\texttt{word association}) = \boldsymbol{cultural\ association}$$

Under what conditions can we say that this identity holds?

Never ☹

MCS does not consistently represent this theoretical quantity unless we assume something **impossible** about the sets of words we chose to represent the concepts, and this **doesn't depend on which MCS model we use**.

GCD is the **unique** MCS method that exactly adjusts for anisotropy in the bilinear system.

Keyword lists?

Could we sample keyword lists randomly?

- GCD partially fixes our selection issue: rather than treat each word vector as an independent draw from the concept, we treat the **set of word vectors** as an $N = 1$ draw from the set of admissible keyword lists, and adjust for the level of within-set association in each list directly when we make a comparison.
    - To make inferences involving the concept (i.e., the *set* of keyword lists we **could have used**) we'd like to increase N.
- To make headway on identifying subspaces with concepts, we would then need to characterize **the set of admissible keyword lists** for a given concept.

You can try this yourself.

What is the concept for the following keyword list?

{apple, orange, cherry, grape, lemon, lime, raspberry}

What about this one?

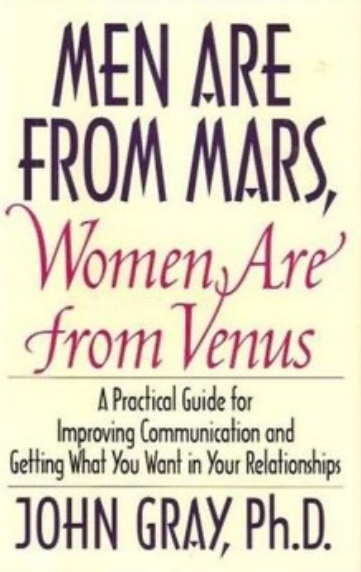
{olive, iris, veronica, lily, daisy, jasmine, heather, rose}

What makes a keyword list a **valid representation** of its target concept?

masculinity

{he, him, his, himself, man, boy, son, father, ...}

{..., uncle, grandfather, men, dude, bloke, infantryman, john, handsome, mustache, mars, mercury, queen, ...}


MEN ARE FROM MARS, Women Are from Venus
A Practical Guide for Improving Communication and Getting What You Want in Your Relationships
JOHN GRAY, Ph.D.

Thank you! ☺

Alex Kindel
Assistant Professor of Sociology
Sciences Po médialab

SICSS Lake Como | 9 July 2025