

# The geometry of diet: using projections to quantify the similarity between sets of dietary patterns

Beatrix Jones\*, joint work with Karen Mumme, Shiyu Fan and John Thompson

27 November 2025

## Creation of Empirical Dietary Patterns (Typically)

- ▶ Food Frequency questionnaire or food diary/recall followed by categorization into food groups.
- ▶ PCA/EFA dimension selection. Work with correlation matrix.
- ▶ Usually 2-5 components
- ▶ Varimax rotation of loadings on the correlation scale
- ▶ Use correlations with original food groups to interpret the axes.

# Example: REACH study n=371, p=56



**Dietary patterns analyse the whole diet.**  
**Understanding who is eating what can assist in health promotion**

## What dietary patterns are older New Zealand adults following?



### Mediterranean

↑ beta-carotene, folate and vitamin E\*

salad, cruciferous (broccoli, cabbage, silver beet etc) and other vegetables (pumpkin, corn etc), avocados, olives, alliums (onions, garlic), nuts, seeds, fish, berries, salad dressings, eggs, cheese, tomatoes, other fruits (feijoa, kiwifruit etc)



Female

or



University educated

or



Physically active



### Western

↑ energy intake\*

processed fish and meat, sauces, condiments, cakes, biscuits, puddings, meat pies, hot chips, confectionery, chocolate, cheese, beer, vegetable oils, salad dressings, sweetened cereal



Male

or



Secondary educated

or



Living with others



Higher alcohol intake



### Prudent

↑ carbohydrate and fibre\*

dried, fresh and frozen legumes, soy-based foods, whole grain, carrots and spices



Physically active

or



Lower (or no) alcohol intake

\*adjusted R<sup>2</sup>>0.25



Mumme K, Conlon C, von Hurst P, Jones B, Stonehouse W, Heath A-LM, Coad J, Haskell-Ramsay C, de Seymour J, Beck K.  
Dietary patterns, their nutrients, and associations with socio-demographic and lifestyle factors in older New Zealand adults.  
Nutr 2020;12(11):3425. doi: 10.3390/nu12113425.



## Example: REACH study comparisons

“Dietary pattern and education associations . . . found in the current study are consistent with other studies in older adults. ‘Mixed’, ‘fat and meat’, ‘Western’, and ‘traditional’ dietary patterns have frequently been associated with a lower education [10,12,14,16], while dietary patterns comprising more healthy food groups, such as ‘vegetable based’, ‘fruit and milk’, ‘plant-based’, or ‘healthy’, are frequently associated with a higher education [9,10,11,14,15,17], although some exceptions have been reported.”

Can we be more quantitative than this? Yes. Caveat—we need the other study to use the same variables.

## Comparing individual dietary patterns

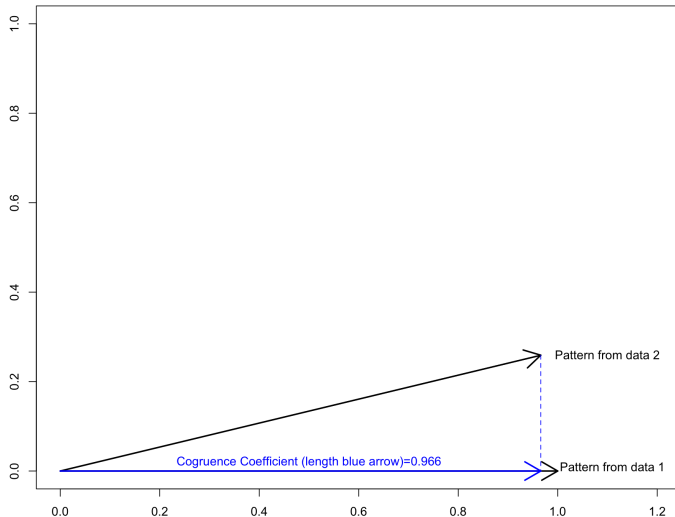
$$\varphi(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^p x_i y_i}{\sqrt{\sum_{i=1}^p x_i^2 \sum_{i=1}^p y_i^2}}$$

“Cosine similarity”, “Tucker’s congruence coefficient,” “congruence coefficient”

Like the correlation without centering. . . if we shift all values of  $\mathbf{x}$  by a constant, the coefficient changes.

# Cosine similarity: Geometry.

Projection of one 1D pattern onto another (based on same survey instrument).



Cosine of the angle between the directions.

# Multivariate Cosine Similarity

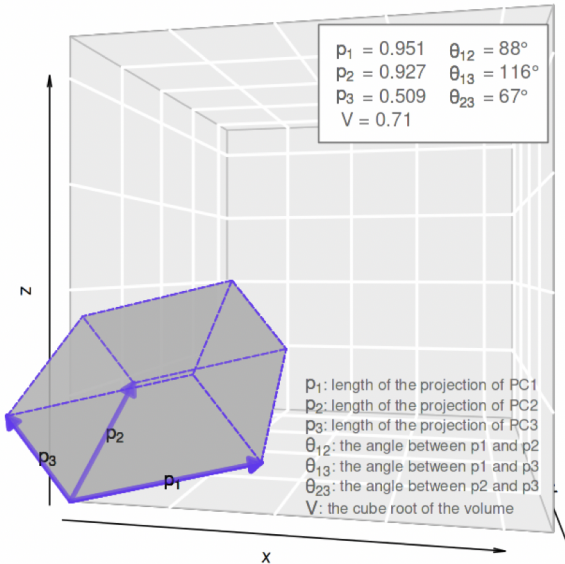
We have two, 3-D dietary pattern spaces (eg one from men, one from women)

1. Project a unit cube from one coordinate system onto the other coordinate system.
2. How much did it shrink?
3. Take the cube root to make it more comparable to the 1-D quantities.

Can extend to any number of dimensions

# Geometry

## Projection in 3D





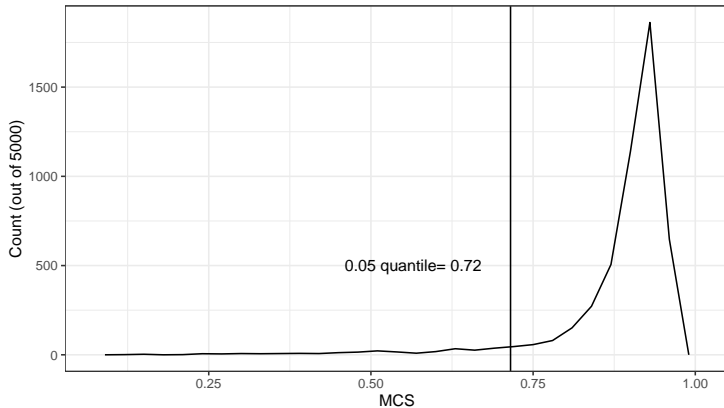
## Volume of the projection

Let  $\beta_{1,1:k}$  be the vectors of coefficients of the PC's/factors for the population,  $\beta_{2,1:k}$  the vectors of coefficients of the PC's/factors for the sample. Both orthonormal sets of vectors.

$$MCS = \sqrt[k]{|\det(\beta_{1,1:k}^T \beta_{2,1:k})|}$$

- ▶ Invariant to rotation/reflection
- ▶ Need to start with an orthonormal basis (for EFA can find with Gram-Schmidt, if necessary)

# Bootstrap interval for MCS between first 3 sample and population PCs



But what about...

Average congruence after procrustes rotation (“Pro”)  $\beta_{1\cdot}^*$  = rotated version of  $\beta_{1,1:k}$  to best match  $\beta_{2,1:k}$

$$\sum_{i=1}^k (\beta_{1i}^{*\top} \beta_{2i})^2 / k$$

Krzanowski Index, normalized to have maximum 1 (“Krz”)

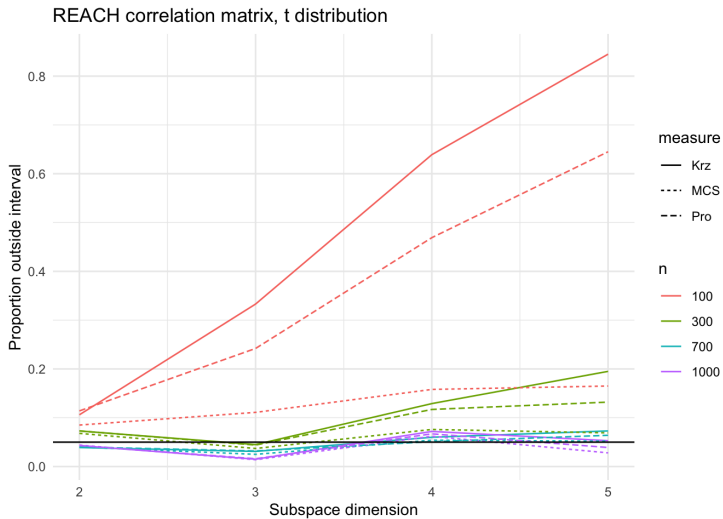
$$\frac{1}{k} \sum_{i=1}^k \sum_{j=1}^k (\beta_{1i}^{\top} \beta_{2j})^2$$

Could be used in the same way... and how do intervals perform relative to nominal coverage?

# Simulation

- ▶ Use correlation matrix from REACH as population value.
- ▶ Simulate samples using this matrix and the t-5 distribution.
- ▶ Sample  $n = 100, 300, 700, 1000$
- ▶ Look at subspaces of dimension 2, 3, 4, 5
- ▶ Nominally 95% intervals; for 1000 replications, in what proportion does the constructed interval exclude the true MCS between population and sample?

# Results

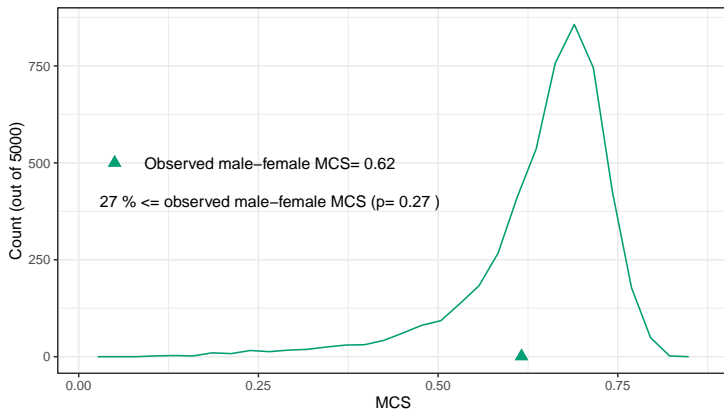


## Things to note:

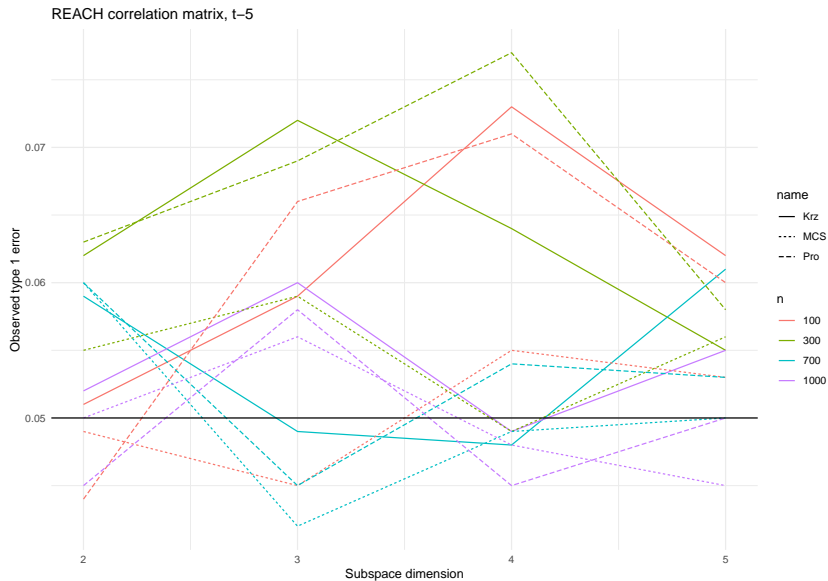
- ▶ MCS much better at higher subspace dimension and small-to-moderate sample size.
- ▶ Based on other examples, issue not subspace dimension but how distinctive consecutive eigenvalues are.

## Example: Permutation test for population differences

REACH– Compare 131 men to 234 women. Observed MCS is 0.62.  
Use as test statistic in a permutation test.



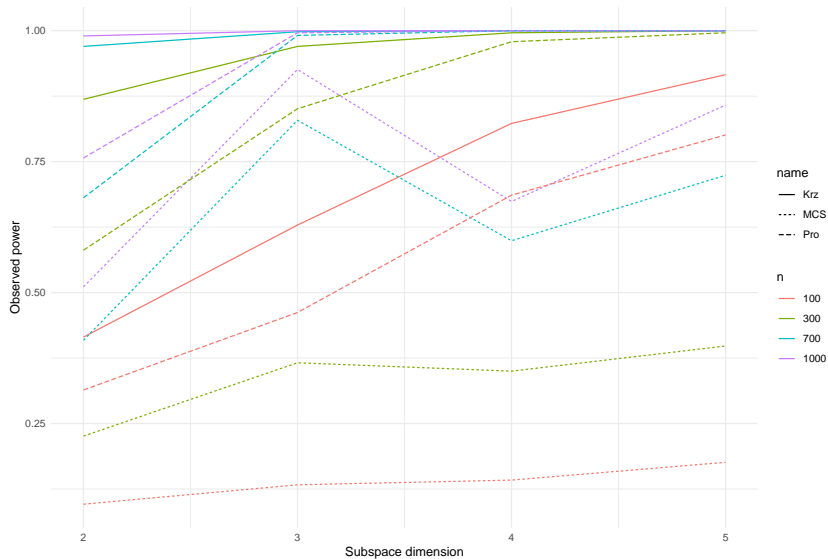
# Type 1 error (vs Krz, Pro)



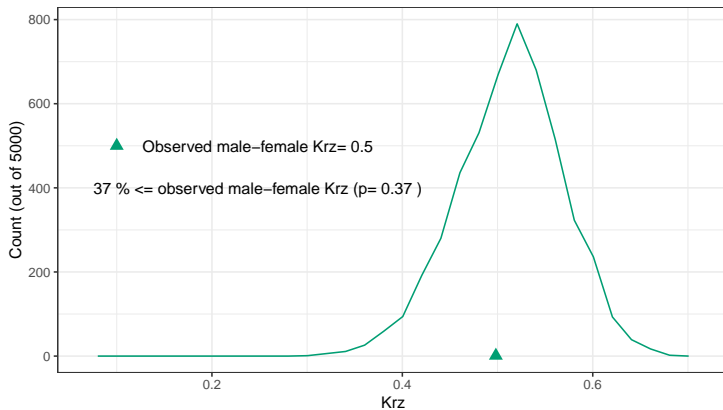


# Power (vs Krz, Pro)

REACH male correlation matrix vs female, t=5



# So what if we redo the test with Krz?



## Other uses

- ▶ Are whole population PCs adequate for subgroups?
- ▶ Permutation test for paired data (longitudinal questions, reproducibility)
- ▶ Could use with supervised dietary patterns (canonical correlation/reduced rank regression/PLS)