Investigating paired differences for data sets with special structures after PCA

John Castura
Compusense Inc.
Canada

Paula Varela
Tormod Næs
Nofima AS
Norway

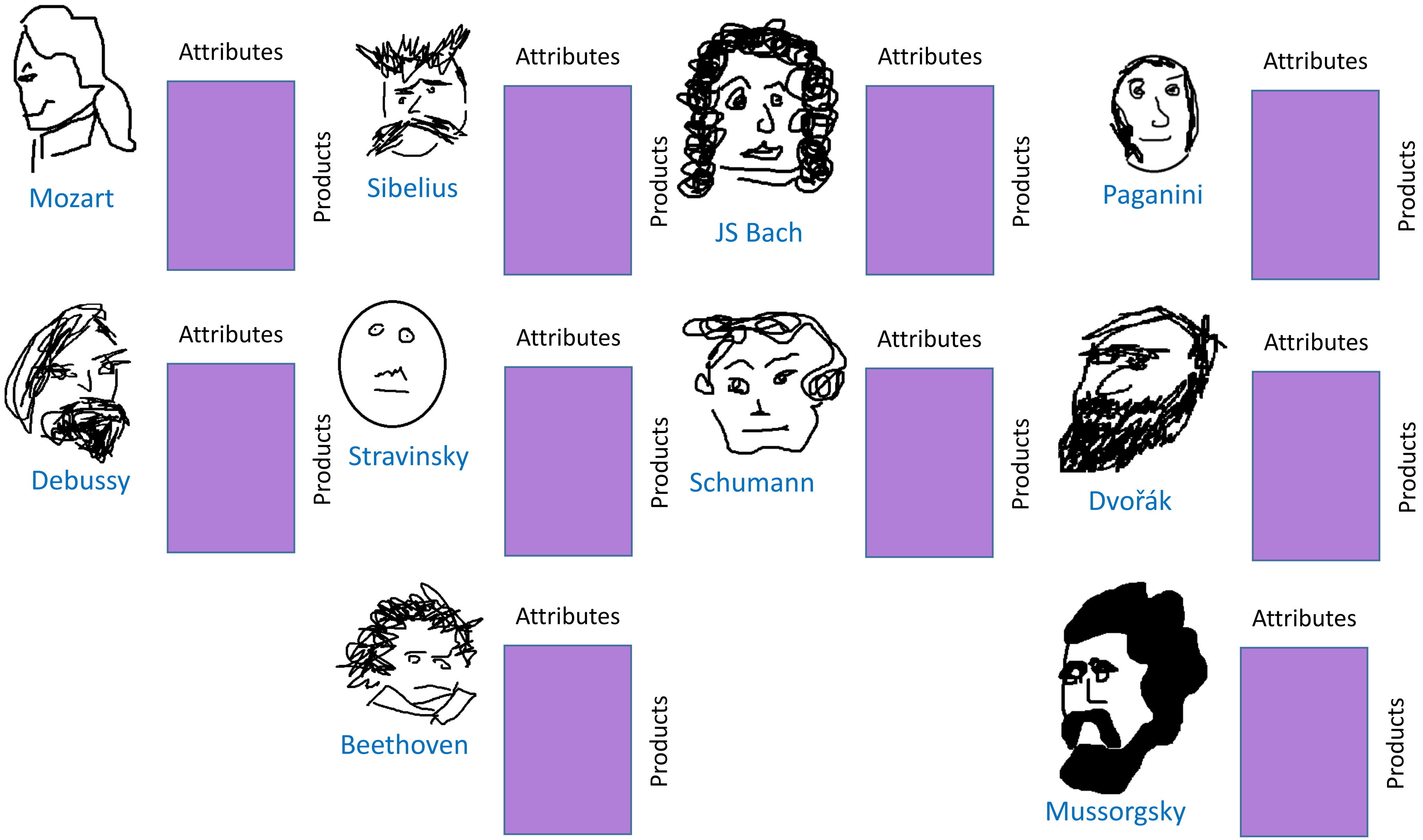
Pangborn 2023
15th Pangborn Sensory Science Symposium
Meeting new challenges in a changing world
20–24 August 2023
Nantes
France

A typical application of **principal component analysis** in sensory evaluation

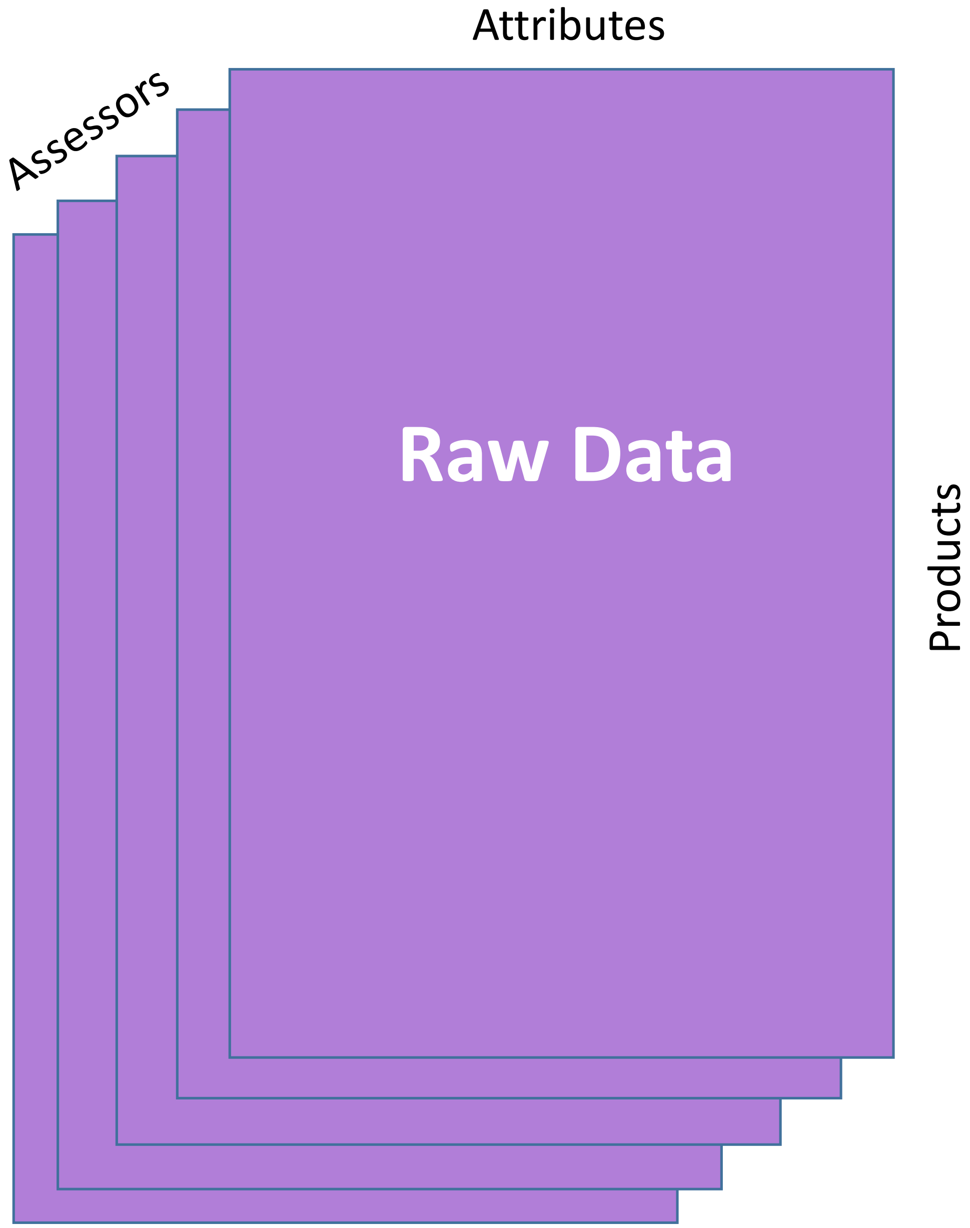
Panel of Trained Sensory Assessors

Illustration credit: J.C. Castura

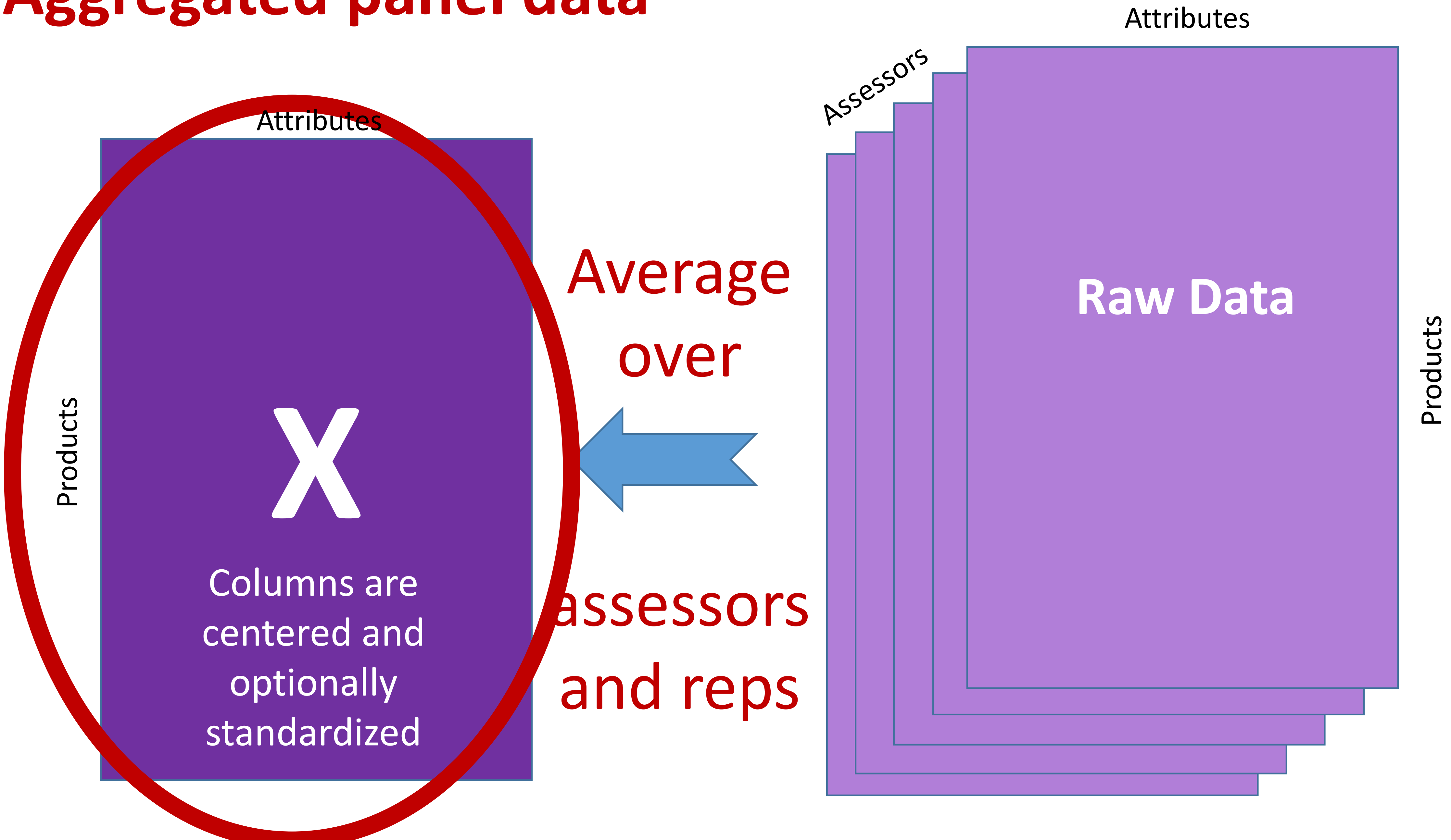
Data from a trained sensory panel



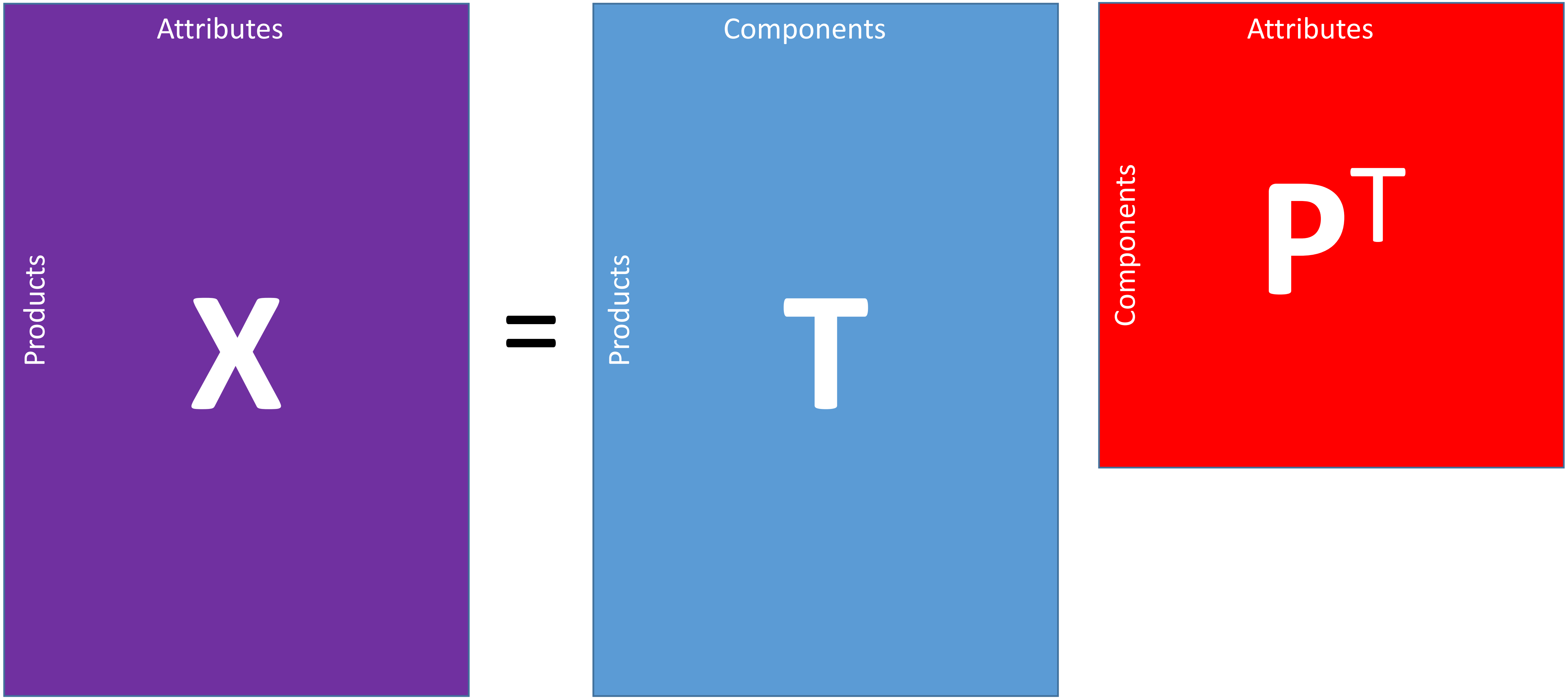
Panel data



Aggregated panel data

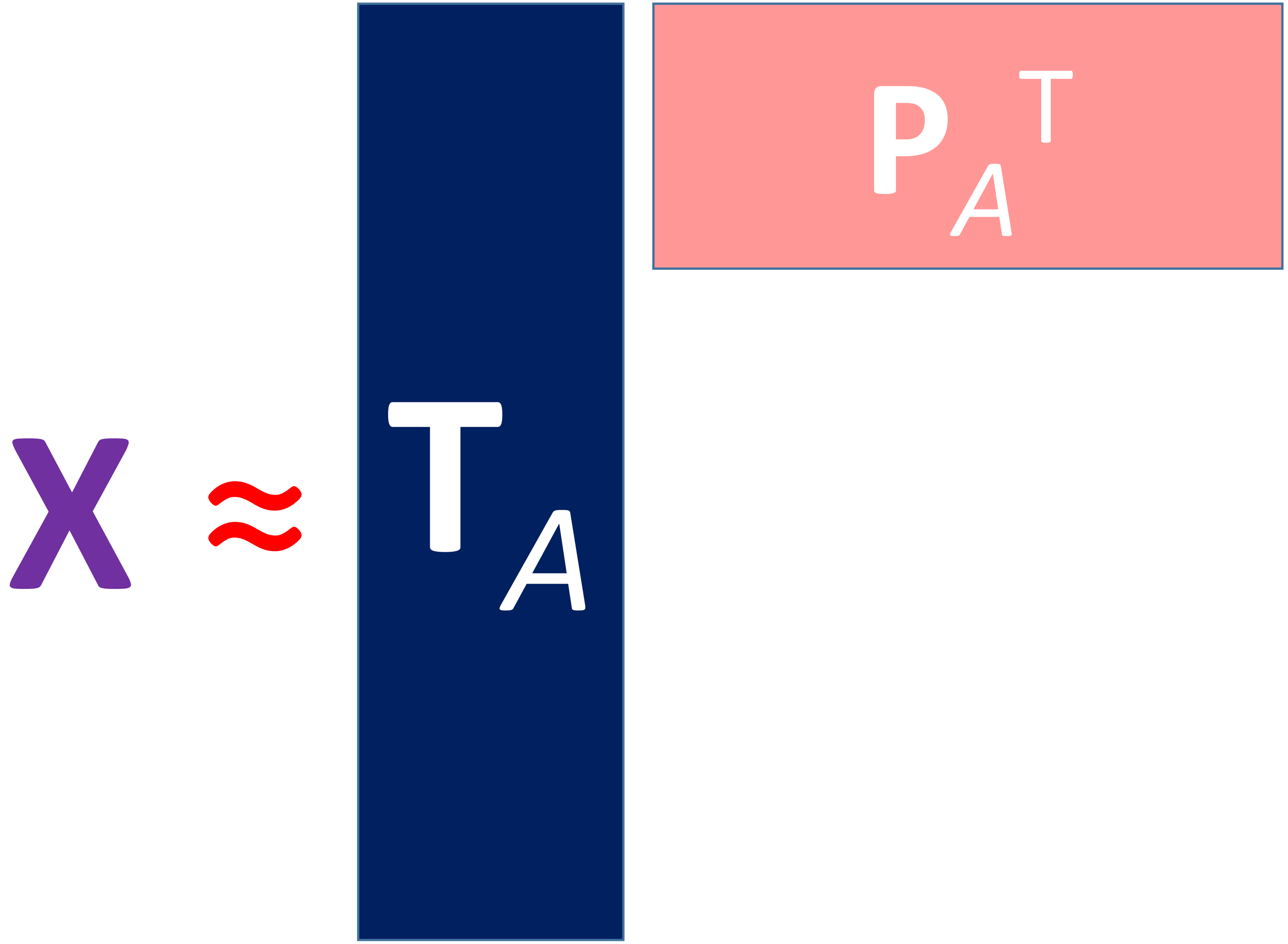


Principal component analysis

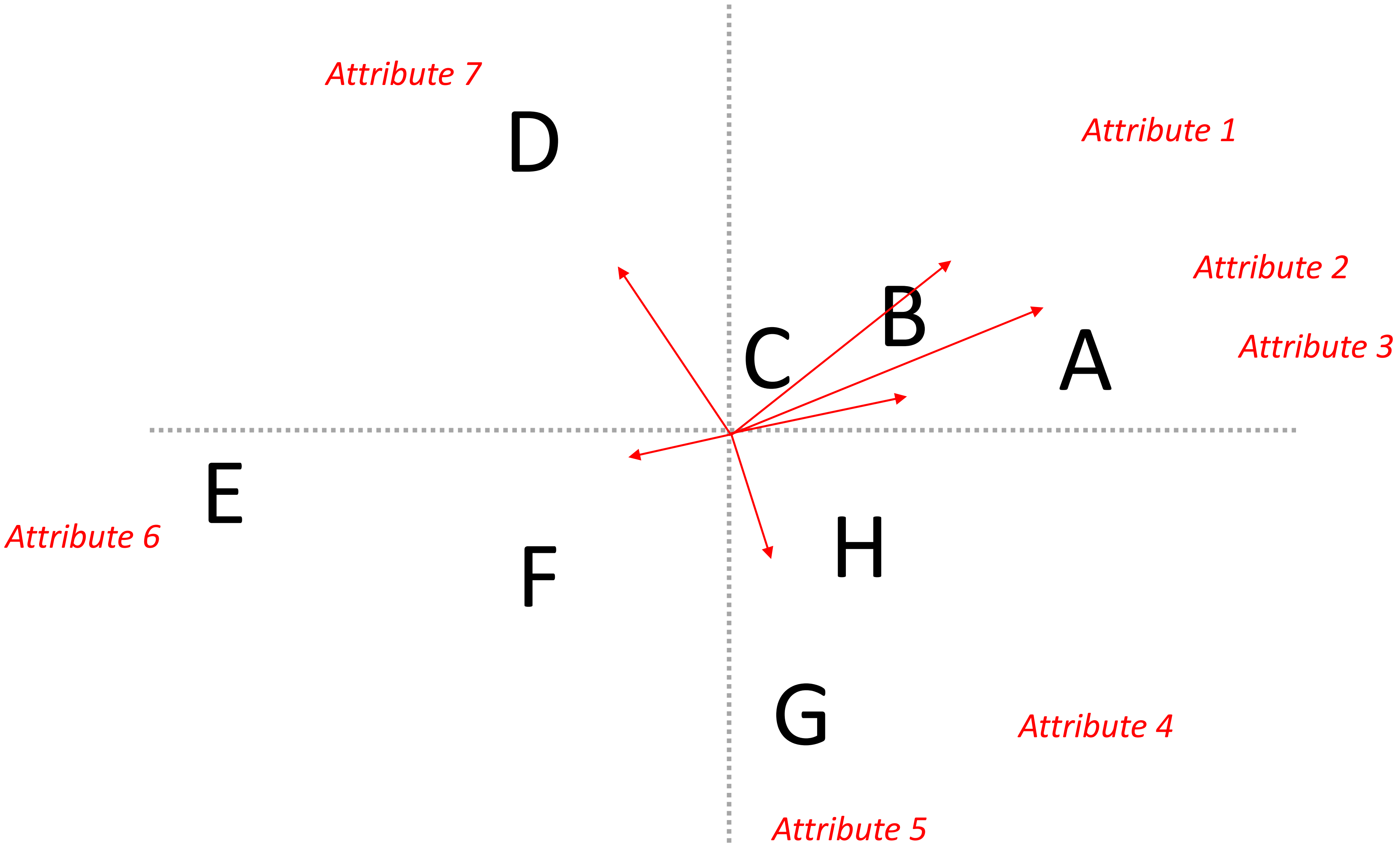


Dimension Reduction to *A* PCs

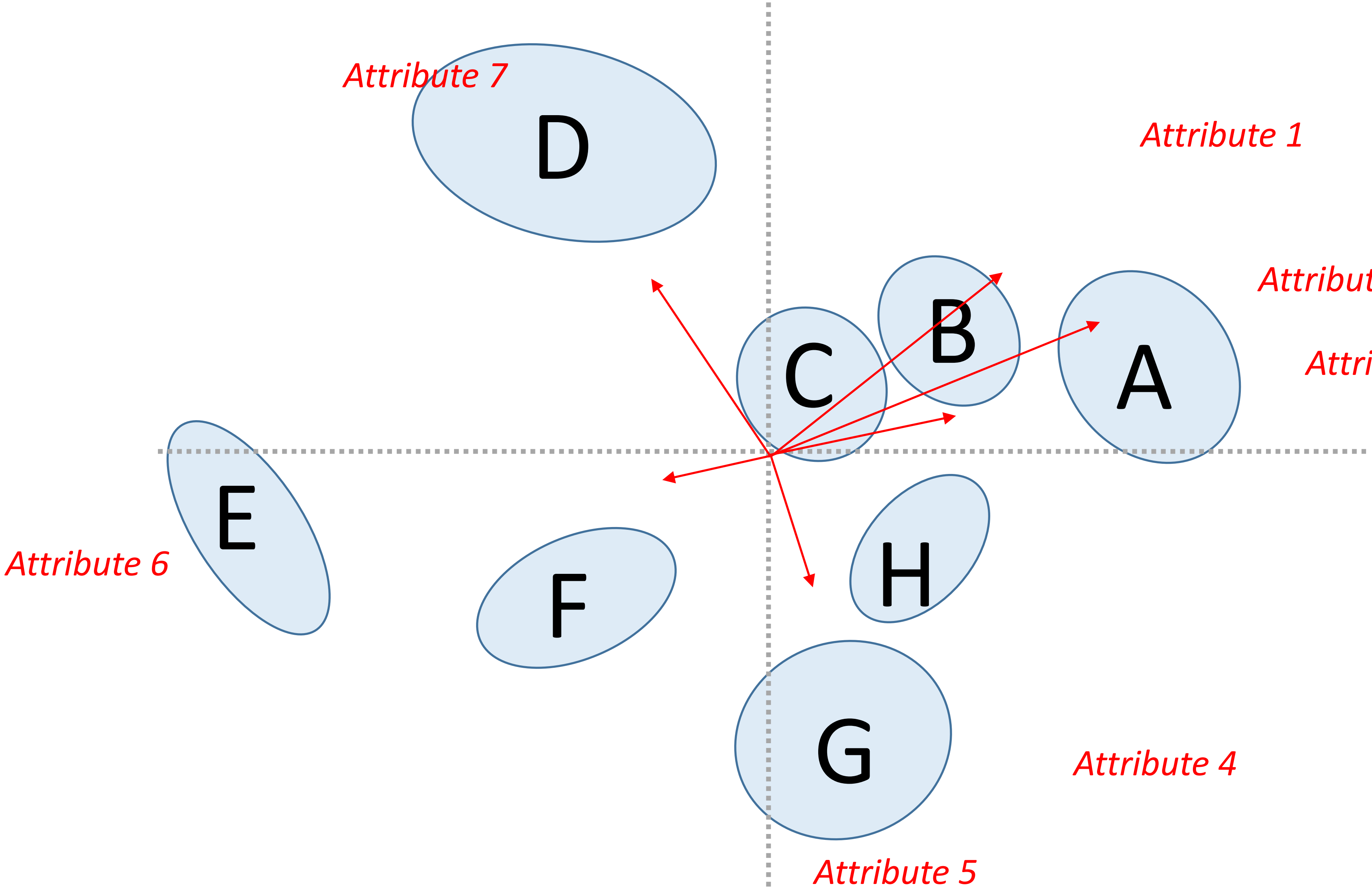
$$\mathbf{X} \approx \mathbf{T}_{A} \mathbf{P}_{A}^{\mathrm{T}}$$



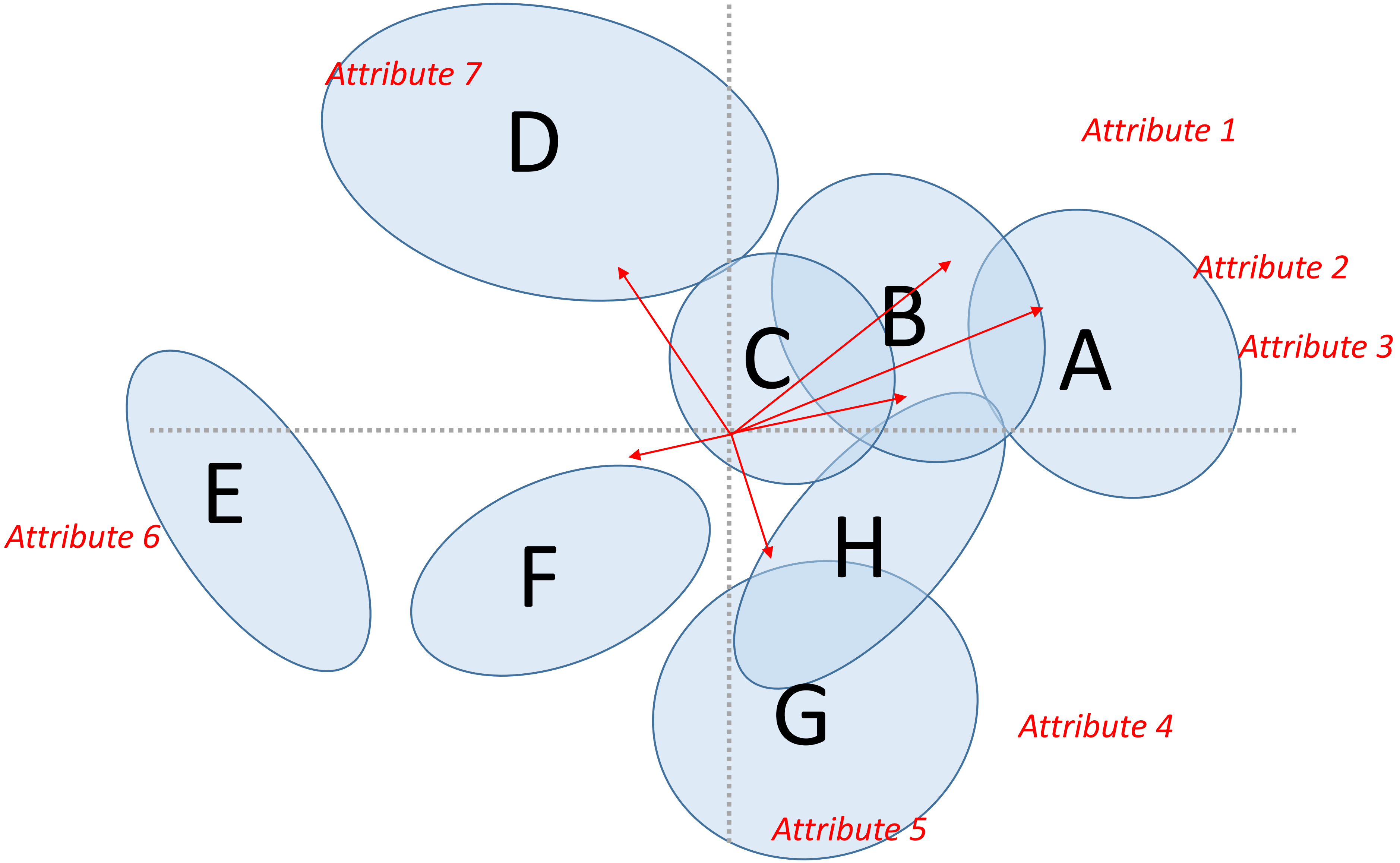
PCA results


Attribute 7
D
Attribute 1
Attribute 2
B
C
A
Attribute 3
E
Attribute 6
F
H
G
Attribute 4
Attribute 5

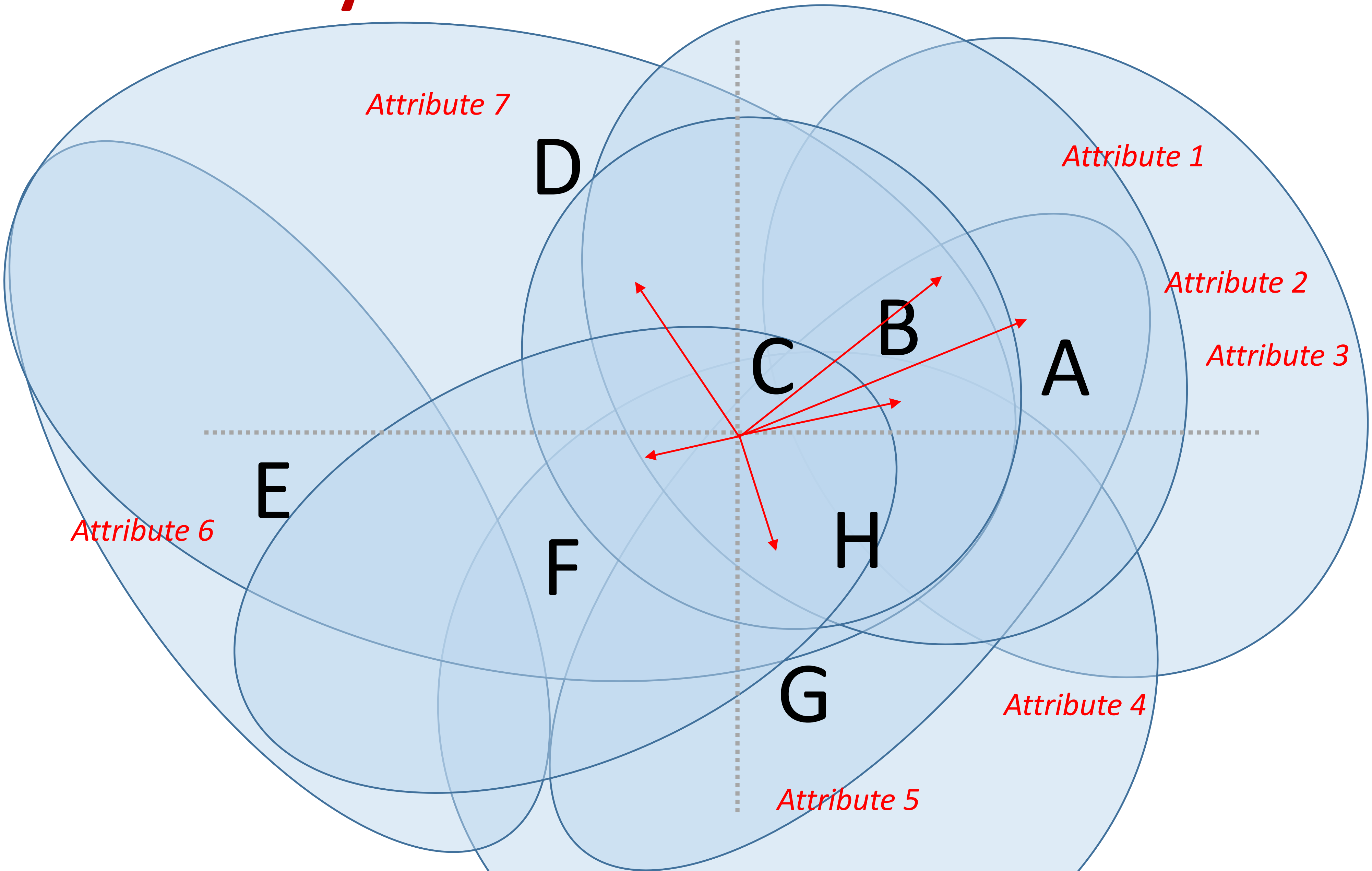
Uncertainty in PCA results



Uncertainty in PCA results



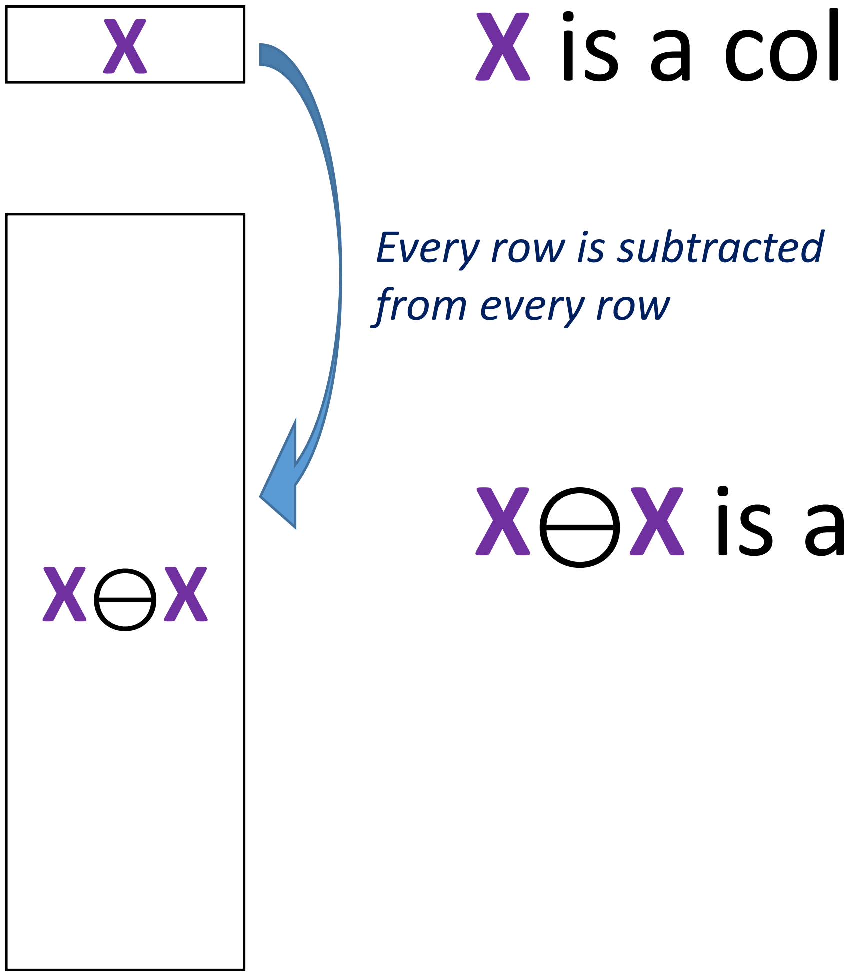
Uncertainty in PCA results



Paired Comparisons after PCA

Castura, J.C., Varela, P., & Næs, T. (2023). Investigating paired comparisons after principal component analysis. *Food Quality and Preference*, 106, 104814.
https://doi.org/10.1016/j.foodqual.2023.104814

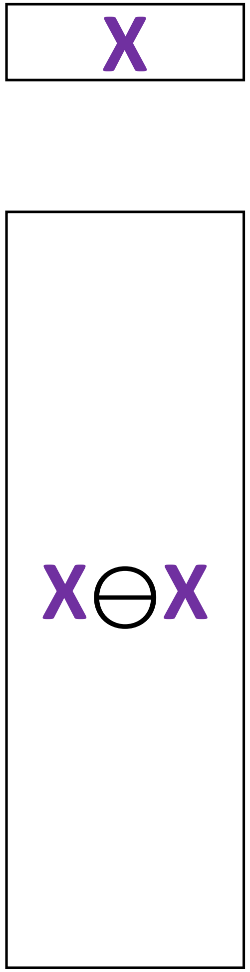
“Crossdiff-unfolding”



X is a column-centered ($J \times M$) matrix

X⊖**X** is a column-centered ($J^2 \times M$) matrix

“Crossdiff-unfolding”



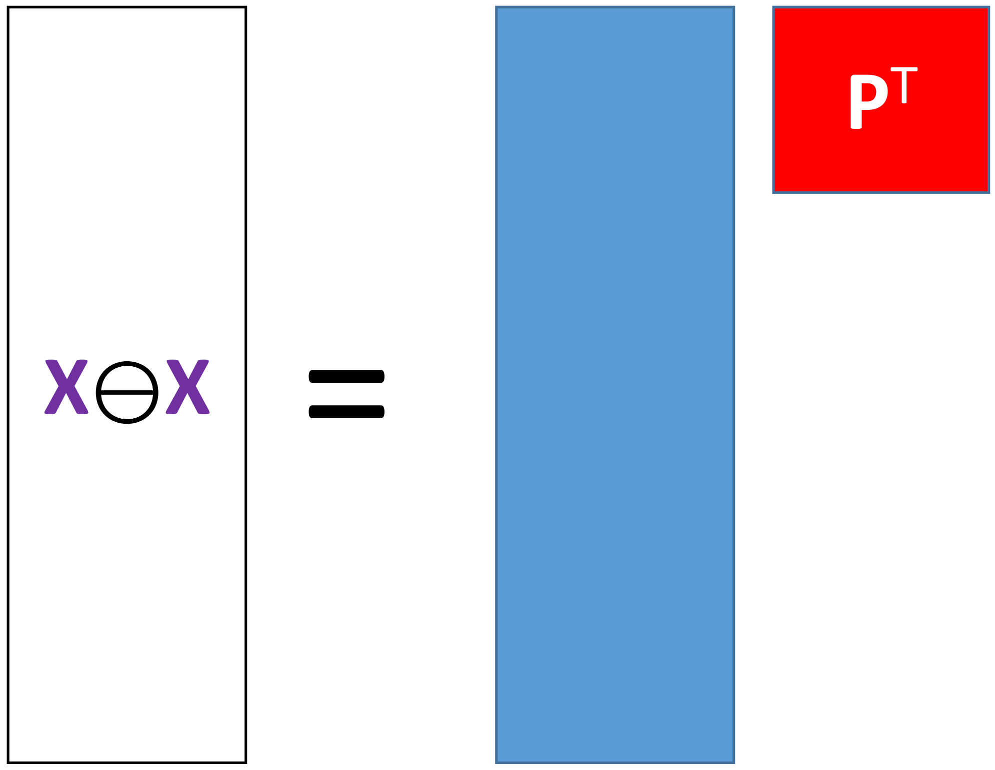
The covariance matrix of **X** and the covariance matrix of **X⊖X** are identical except for a multiplier.

Next, we consider PCA of **X** and PCA of **X⊖X**.

Key relationships

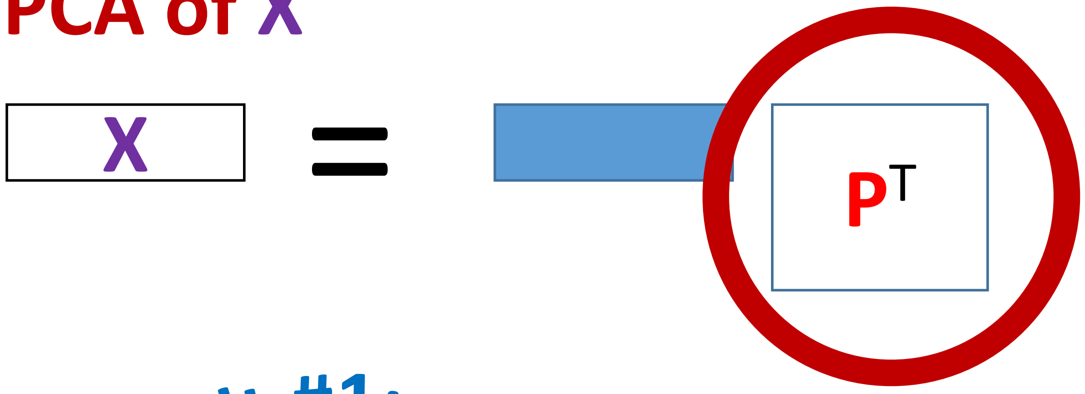
PCA of X

PCA of X⊖X

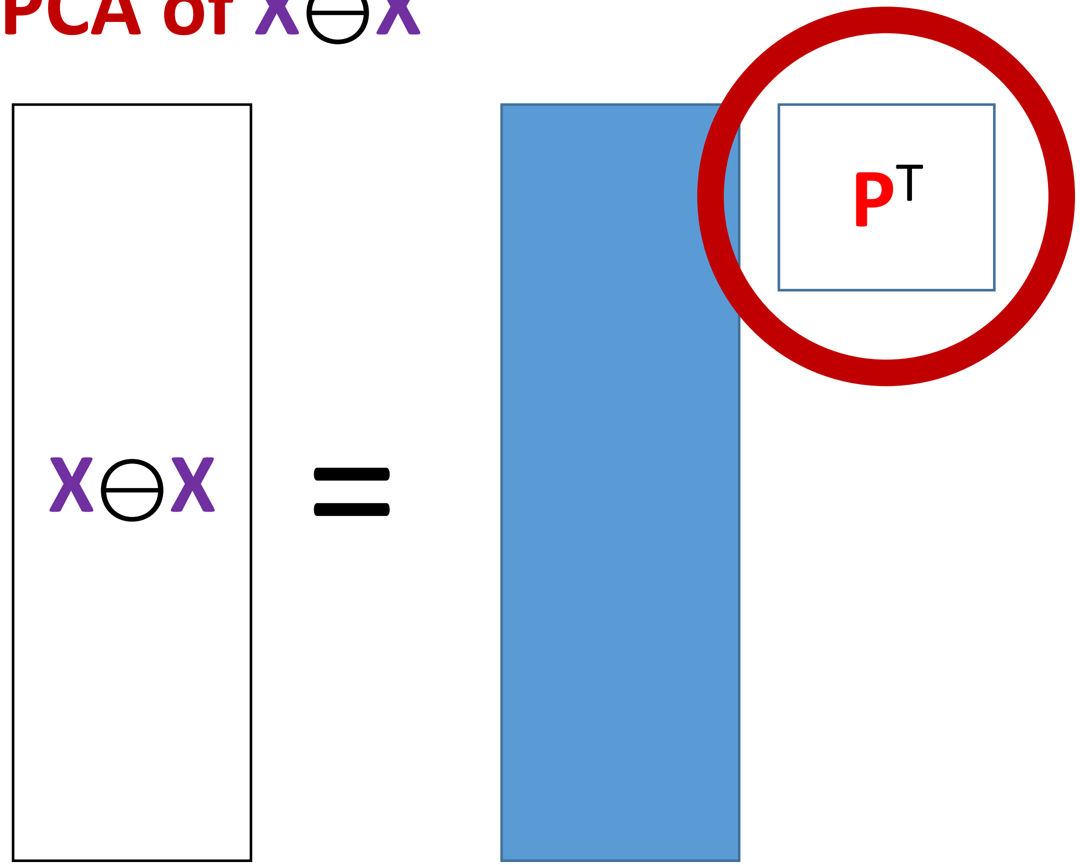


Key relationships

PCA of X



PCA of $X \ominus X$

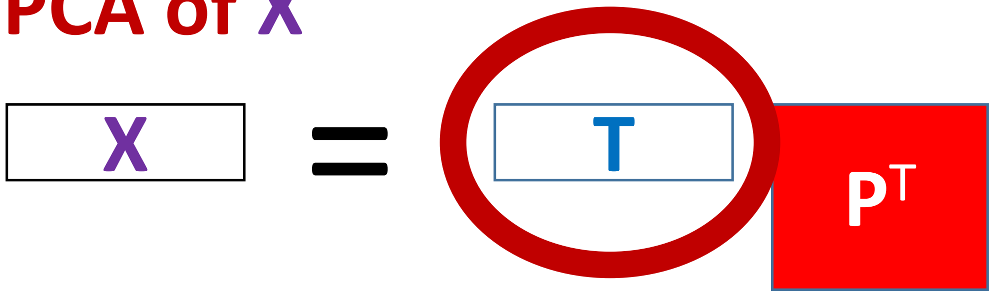


Key result #1:

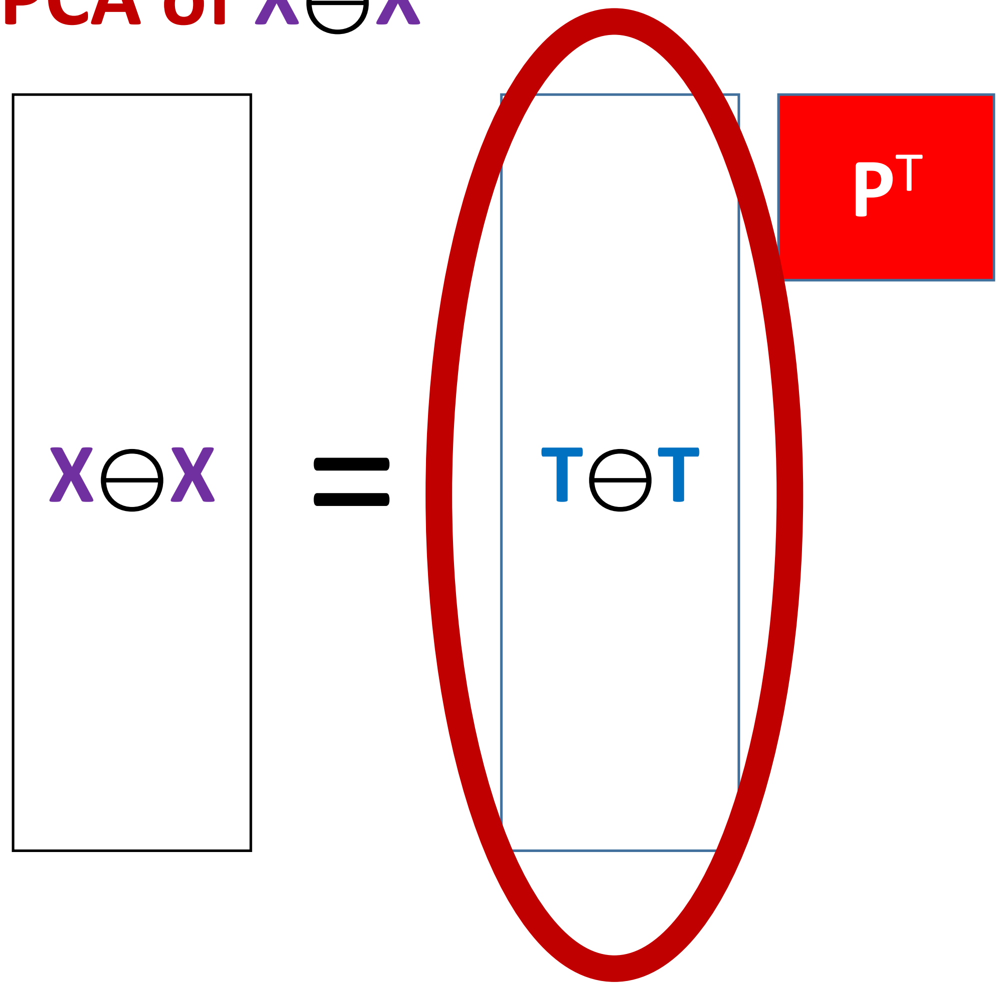
Loading matrices obtained from these two PCA solutions are identical.

Key relationships

PCA of X



PCA of $X \ominus X$



Key result #2:

If we crossdiff-unfold scores from PCA of X, we get scores from PCA of $X \ominus X$.

Paired comparisons

Row objects in **X** *and* **all paired comparisons** have the same PCs

Uncertainty in Paired Comparisons after PCA

Castura, J.C., Varela, P., & Næs, T. (2023) Evaluation of complementary numerical and visual approaches for investigating pairwise comparisons after principal component analysis. *Food Quality and Preference*, 107, 104843.
https://doi.org/10.1016/j.foodqual.2023.104843

The uncertainty cloud of each paired difference **accounts for mutual dependencies** and can be used to obtain…

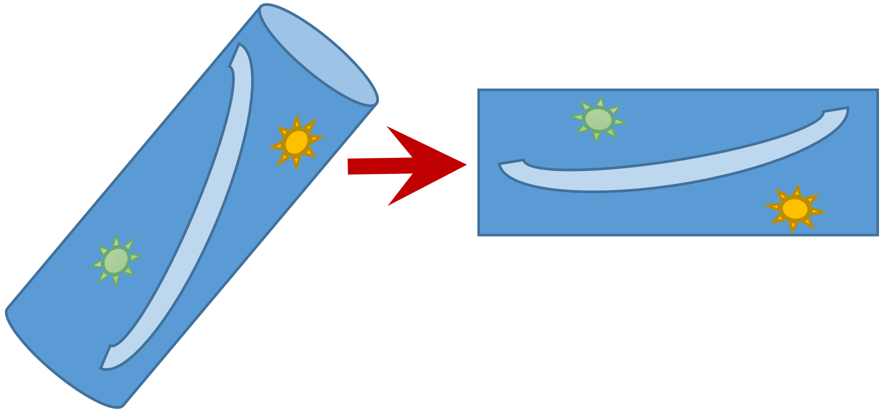
nonparametric uncertainty regions

confidence ellipsoid approximations

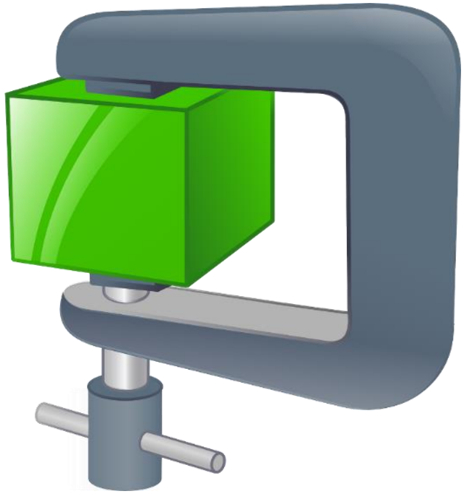
P values

Principal component analysis

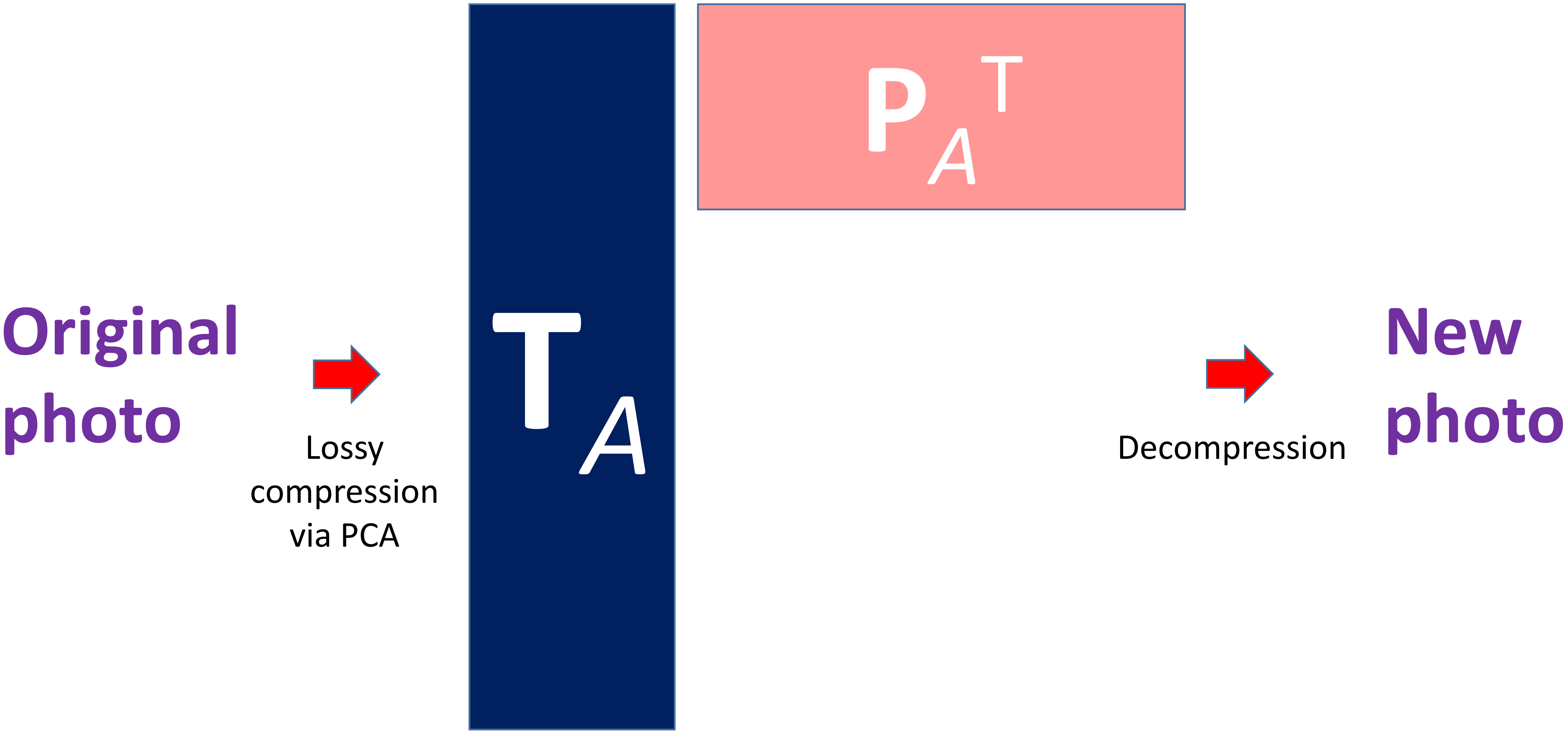
PCA is a statistical method that maximizes the variance in the standardized linear projection of a matrix.



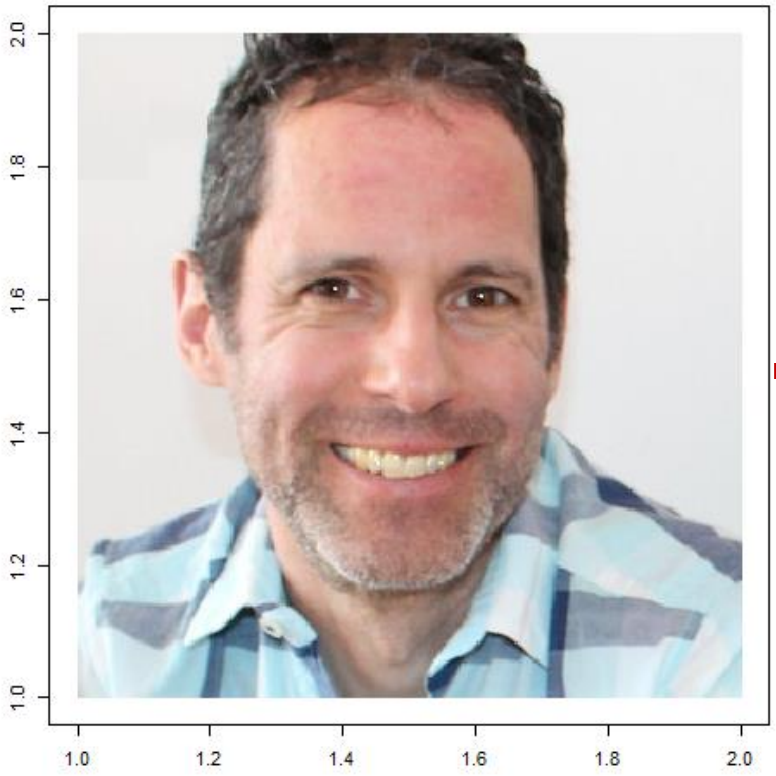
PCA is a method for **data compression** via dimension reduction.



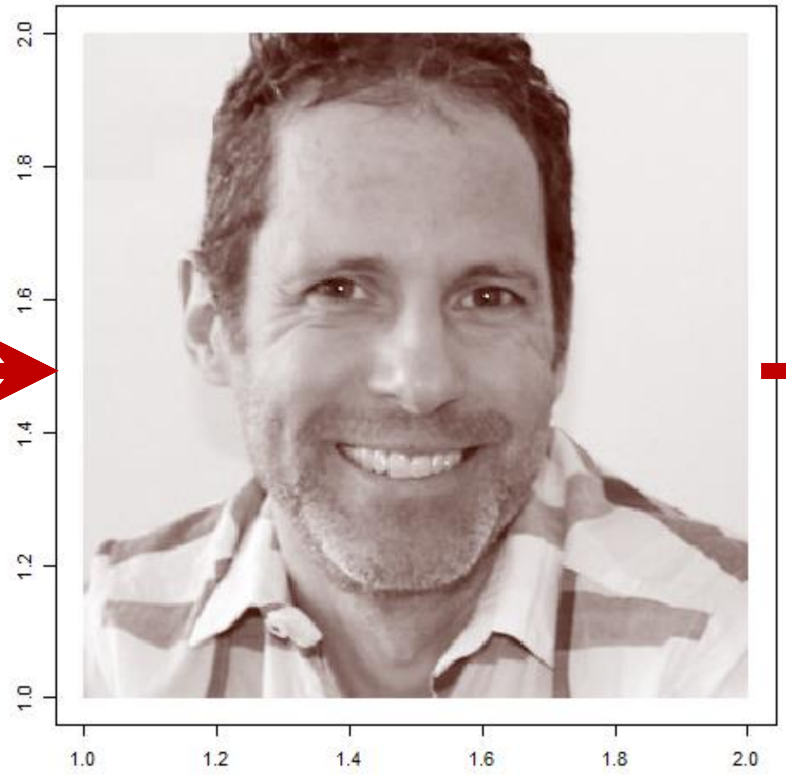
PCA of a Photograph



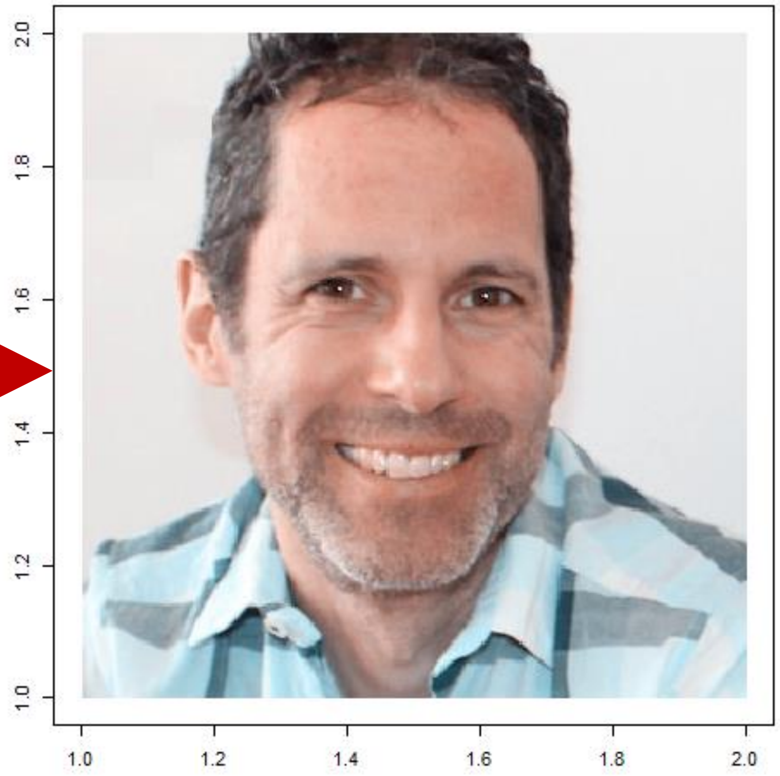
Lossy compression – example 1



Original image
has 3 components (RGB)

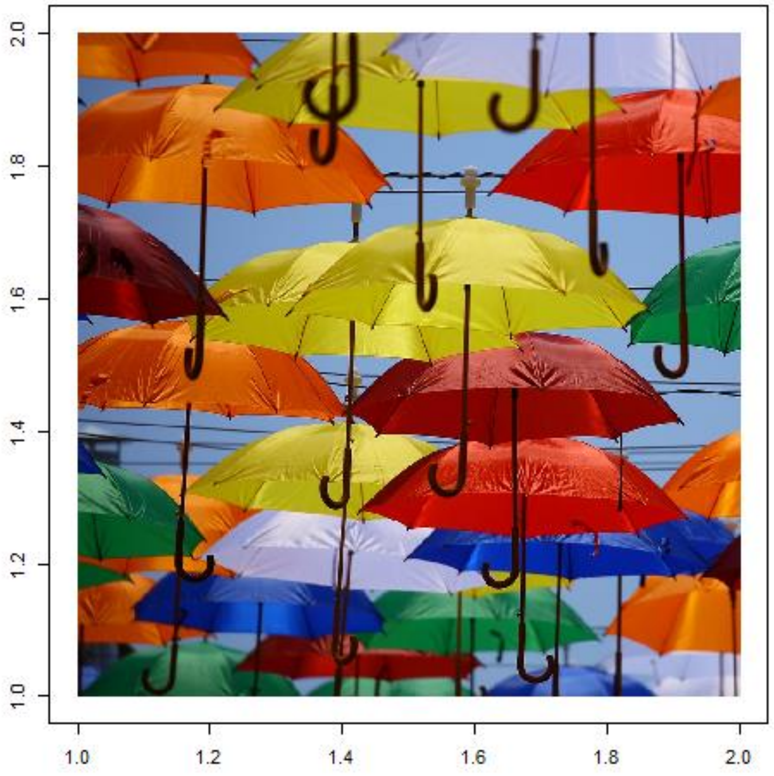


Compression to 1 PC
93% of RGB variance
extracted

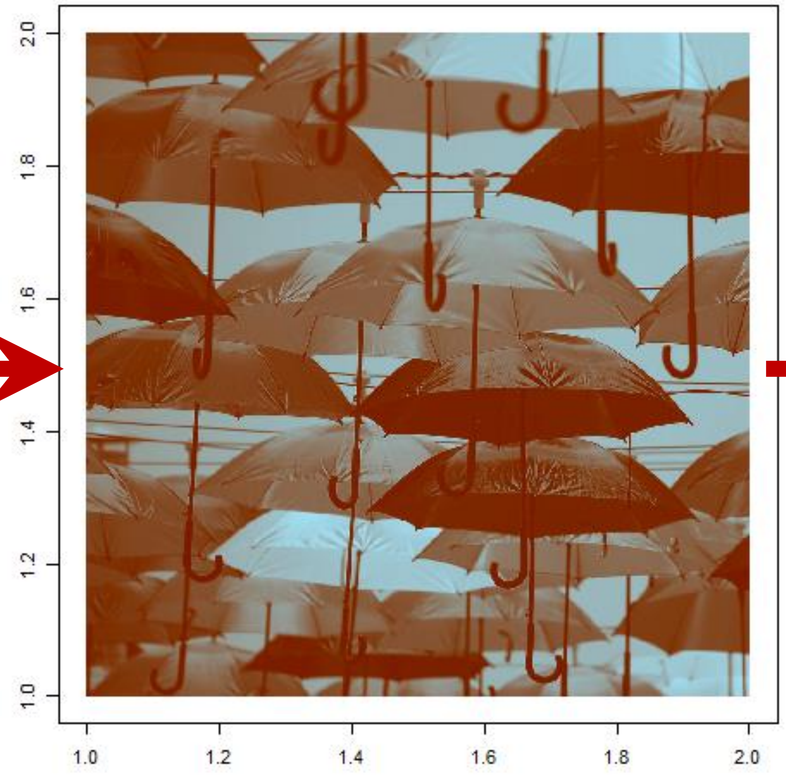


Compression to 2 PCs
97% of RGB variance
extracted

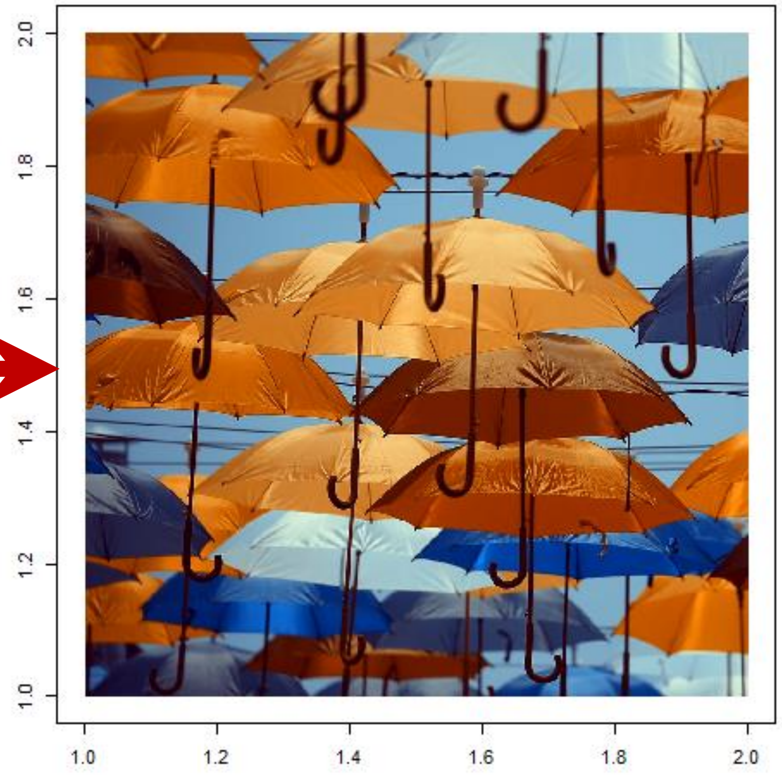
Lossy compression – example 2



Original image
has 3 components (RGB)



Compression to 1 PC
57% of RGB variance
extracted



Compression to 2 PCs
92% of RGB variance
extracted

Investigating a Subset of Paired Comparisons after PCA

Castura, J.C., Varela, P., & Næs, T. (2023). Investigating only a subset of paired comparisons after principal component analysis. *Food Quality and Preference*, 110, 104941. https://doi.org/10.1016/j.foodqual.2023.104941

When are only a subset of paired comparisons relevant?

Examples:

1. **Many Test Products *vs* One Control**

   Focus on Test-Control pairs,
   not Test-Test pairs

2. **Temporal sensory data**

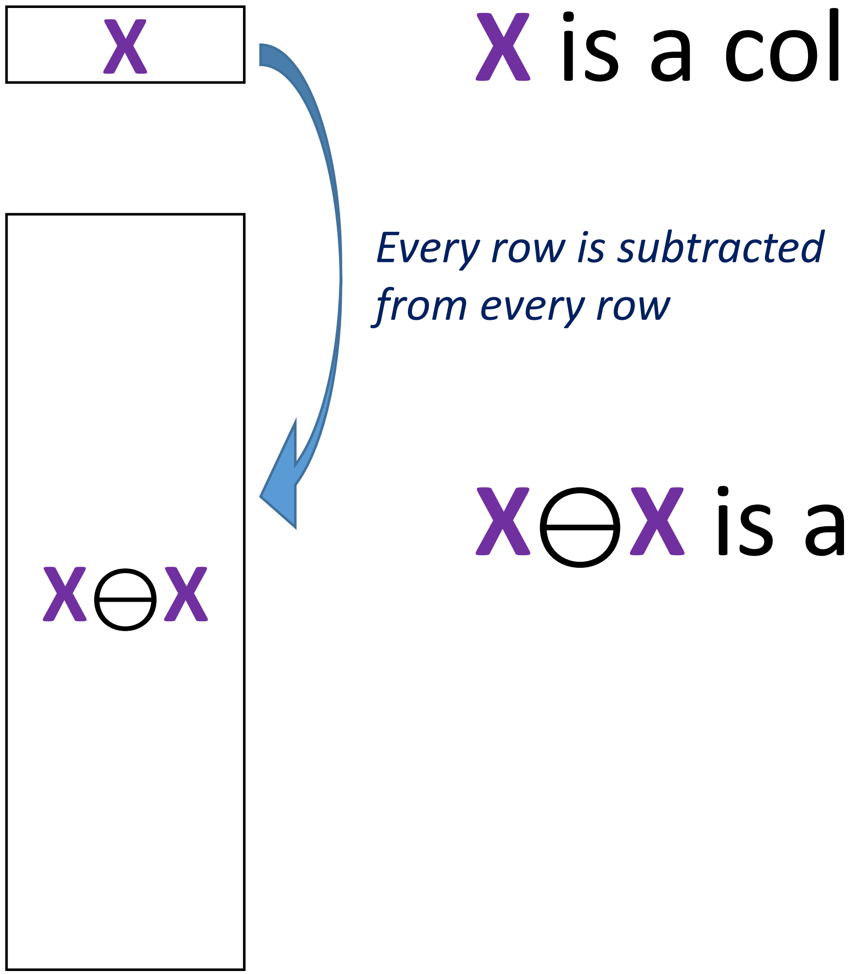
   Focus on Within-timepoint pairs,
   not Across-timepoint pairs

Investigating only a subset of paired comparisons

"…the interrelationships between the variables might be different for the subset of paired comparisons than it is for all paired comparisons. So the covariance matrix for a matrix of all paired comparisons and the covariance matrix of selected paired comparisons will differ depending on the data. "

Castura, J.C., Varela, P., & Næs, T. (2023). Investigating only a subset of paired comparisons after principal component analysis. *Food Quality and Preference*, 110, 104941.

Crossdiff-unfolding



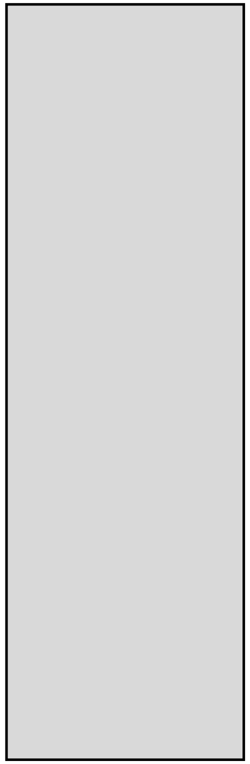
X is a column-centered $(J \times M)$ matrix

X$\ominus$**X** is a column-centered $(J^2 \times M)$ matrix

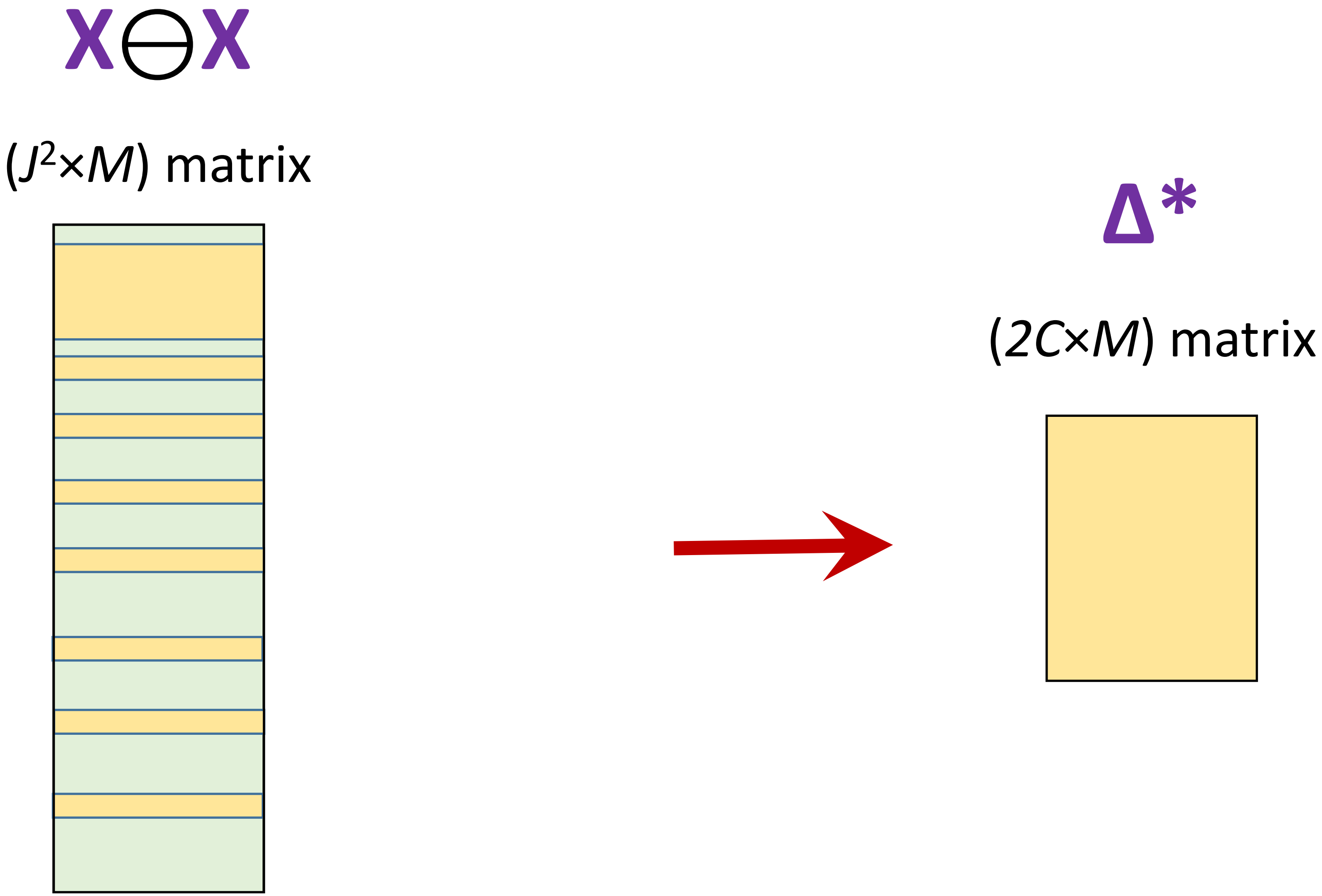
Rows of $\mathbf{X} \ominus \mathbf{X}$ contain all paired comparisons

$\mathbf{X} \ominus \mathbf{X}$

($J^2 \times M$) matrix

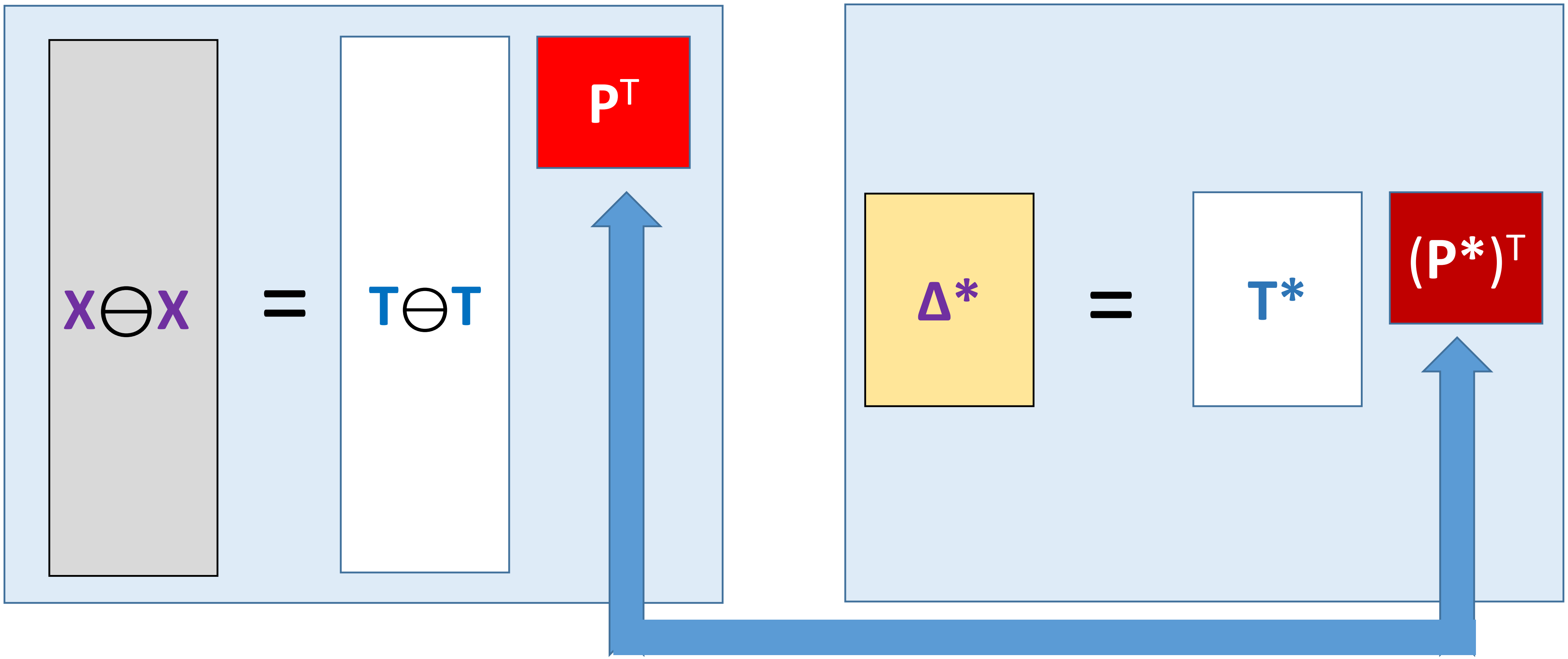


Matrix Δ* contains only *C* relevant paired comparisons



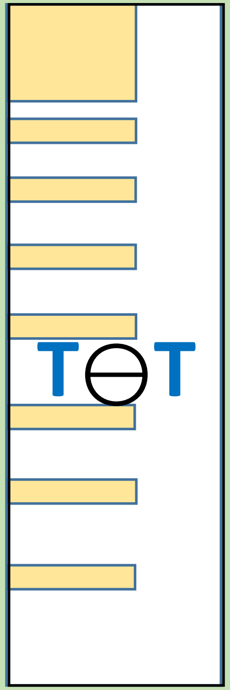
Δ* contains a subset of the rows in X⊖X

PCs of X⊖X and PCs of Δ* are usually different

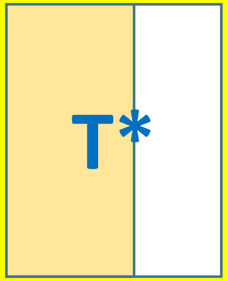


Calculate the relevant sum-of-squares extracted

Sum of squares of **relevant rows** in *A* PCs

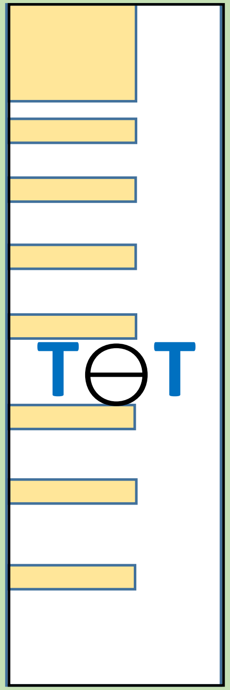


Sum of squares of **all rows** in *A* PCs

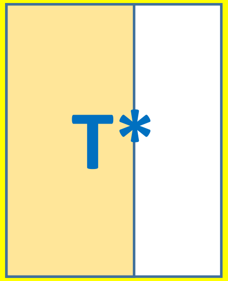


Gain of focusing on *A* PCs of Δ* instead of *A* PCs of X⊖X

Sum of squares of **relevant rows** in *A* PCs



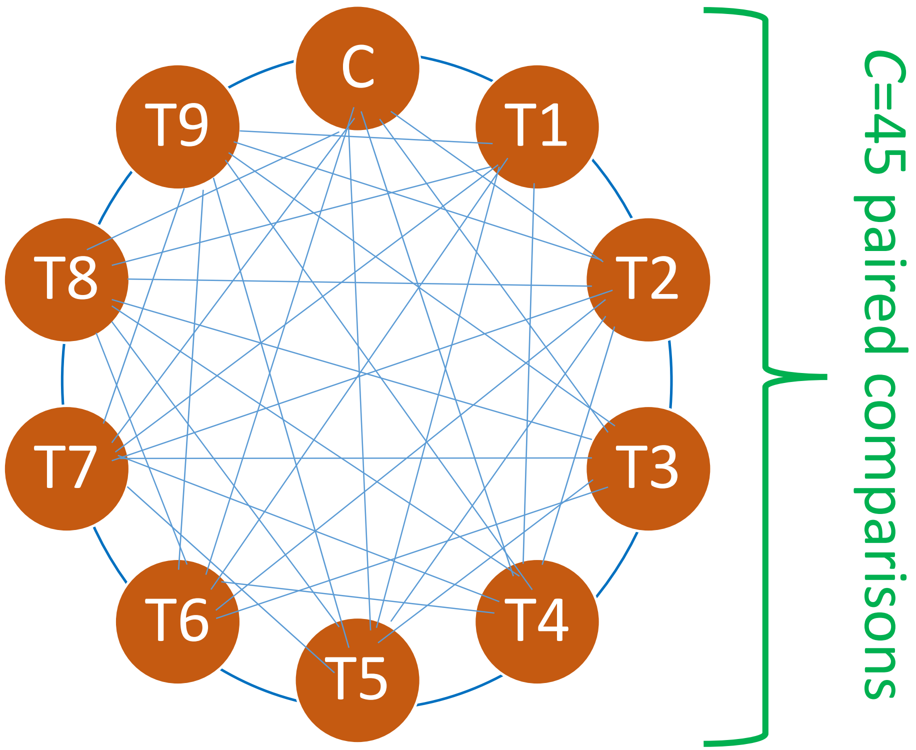
Sum of squares of **all rows** in *A* PCs



Gain = 100(/ - 1)%

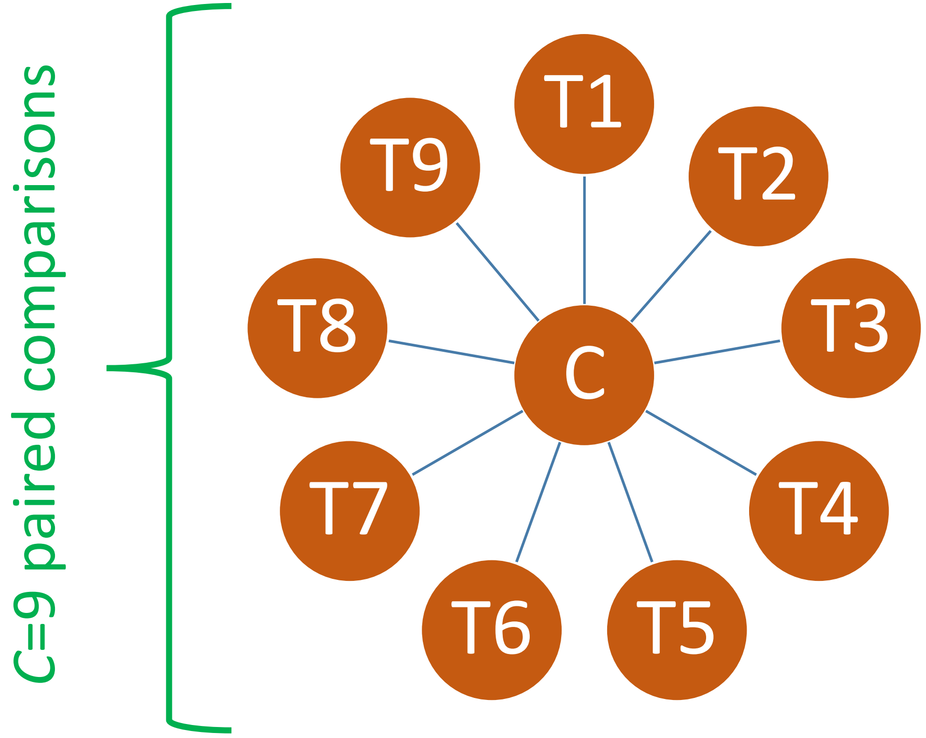
Example 1. QDA of multiple products vs a control

All Paired Comparisons



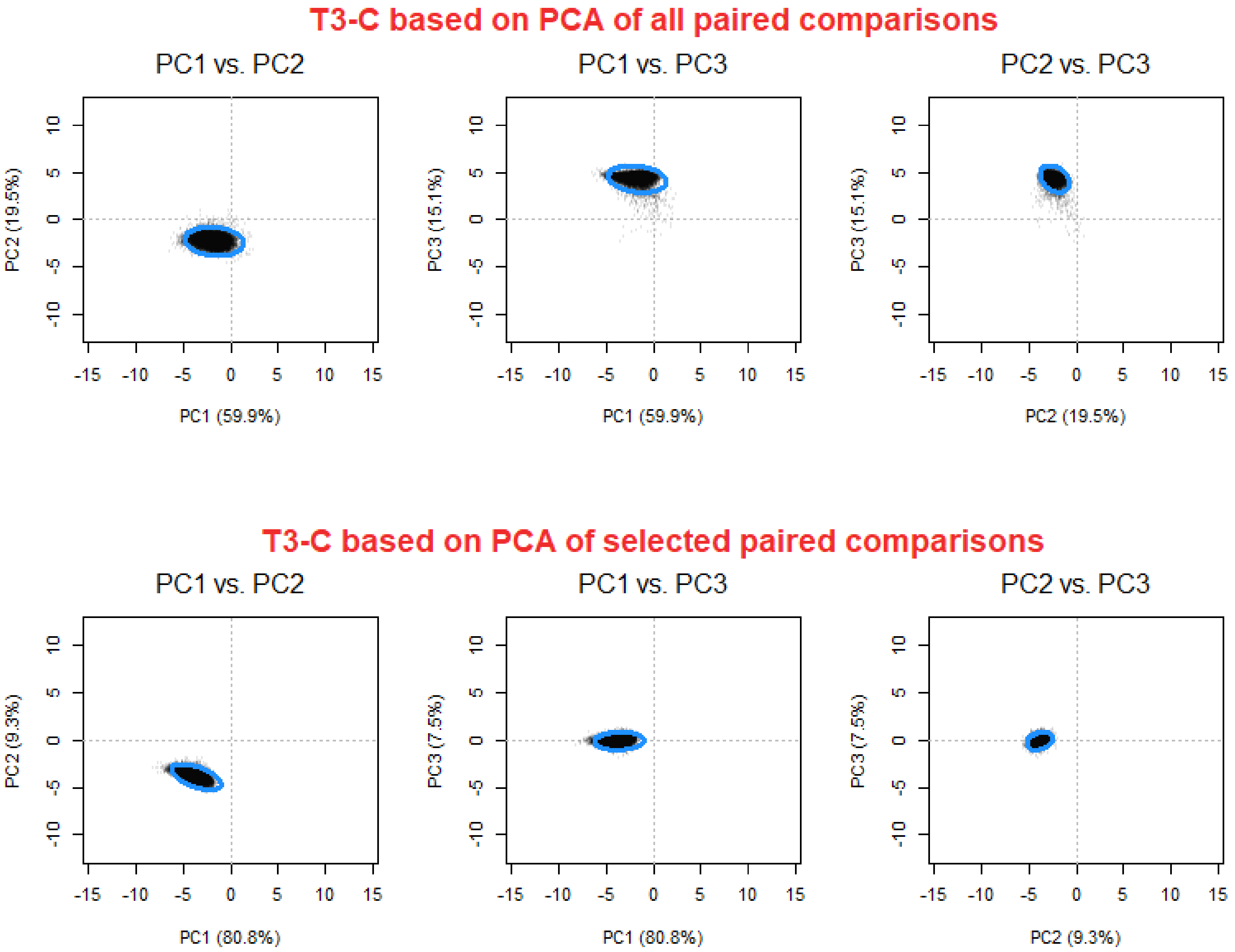
$\mathbf{X} \ominus \mathbf{X}$ has $J^2=100$ rows

Relevant Paired Comparisons



$\boldsymbol{\Delta}^*$ has $2C=18$ rows

Example 1. QDA of multiple products vs a control



Gain:

1 PC:
15%

2 PCs:
14%

3 PCs:
1%

Castura, Varela, & Næs (2023) [eComponent]
doi:10.1016/j.foodqual.2023.104941

Example 2. Temporal check-all-that-apply

All Paired Comparisons

- 8 yogurts × 56 timepoints
- 448 combinations
- All pairs = 100,028
- 10 attributes

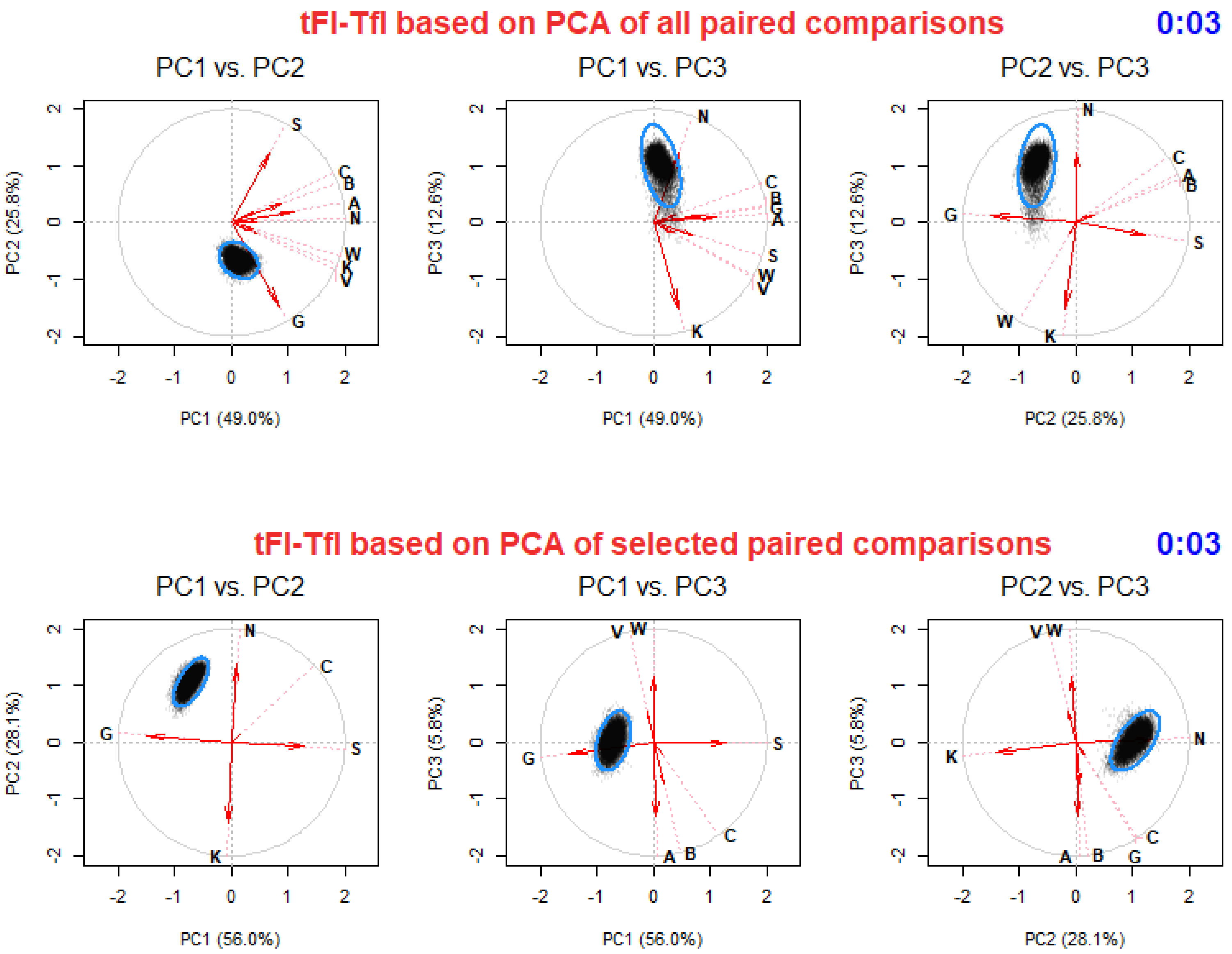
X⊖X has dimension
100028 × 10

Relevant Paired Comparisons

- 28 within-timepoint pairs
- 56 timepoints
- C = 28×56 = 1568
- 10 attributes

Δ* matrix has dimension
3136 × 10

Example 2. Temporal check-all-that apply



Gain:

1 PC: **>3500%**

2 PCs: **52%**

3 PCs: **1%**

Castura, Varela, & Næs (2023) [eComponent]
doi:10.1016/j.foodqual.2023.104941

When only a subset of paired comparison are relevant

Advantages of PCA of Δ* over PCA of X⊖X

- Δ* contains only relevant variance

  …so ***all*** variance extracted by PCA of Δ* is relevant
- Important PCs will tend to have large %VAF
- More natural to focus interpretation on PCs with large %VAF
- Recommended only if a subset of paired comparison are relevant

Advantages of PCA of X⊖X over PCA of Δ*

- Interpretations identical to interpretations of PCA of X
- Conventional so easier to communicate
- Row objects in X are well represented in PCs of X⊖X

John Castura

Paula Varela

Tormod Næs

For further information, please contact jcastura@compusense.com