<u>Hints at other dimension reduction techniques</u>

In PCA, the aim is to find a low-dimensional (projective) representation of the data that preserves variability.

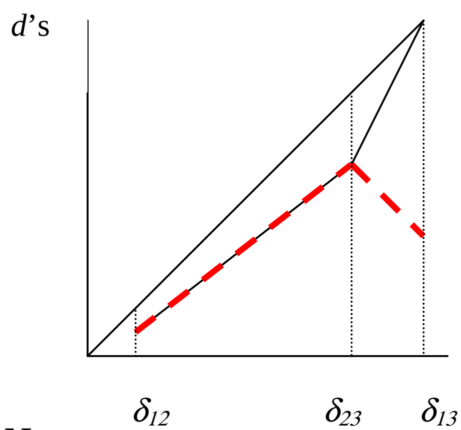
<u>Multi-dimensional Scaling</u> (MDS):
find a low-dimensional representation of the data that preserves "relative positioning" of the points, i.e. distances among them.

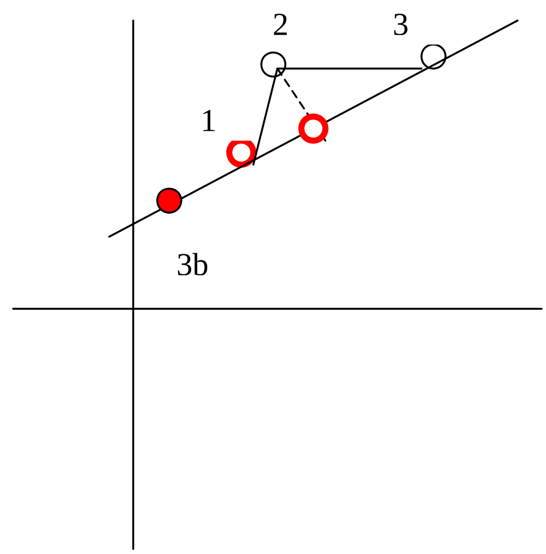
$$X_i \in R^T, \quad i = 1 \ldots N$$

$$\delta_{il} = d(X_i, X_l)$$

$$\delta_{(il)1} \leq \delta_{(il)2} \ldots \leq \delta_{(il)m}, \quad m = \frac{N(N-1)}{2}$$

N=3, *T*=2, *k*=1





$$W_i, \hat{W}_i \in R^k, \quad i = 1 \ldots N$$

$$d_{il} = d(W_i, W_l), \; \hat{d}_{il} = d(\hat{W}_i, \hat{W}_l)$$

$$St(W_1, \ldots W_N) = \min_{\hat{W}\text{'s}:\, \hat{d}\text{'s as close as monotone to } \delta\text{'s}} \left(\frac{\sum_{i<l} (d_{il} - \hat{d}_{il})^2}{\sum_{i<l} d_{il}^{\;2}} \right)$$

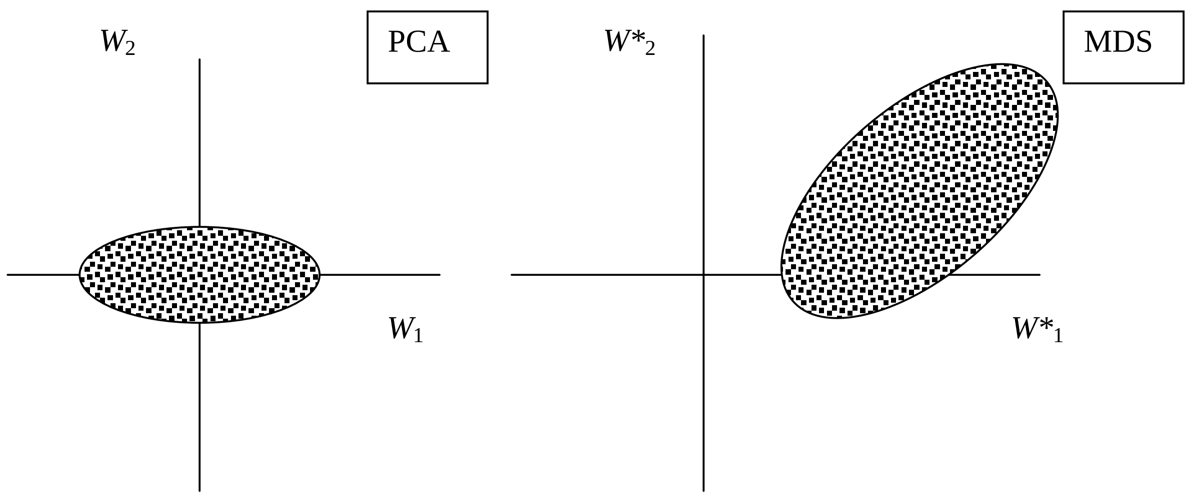
$$St(k) = \min_{W\text{'s}} St(W_1, \ldots W_N) \quad \rightarrow \quad W_1^*, \ldots W_N^*$$

<u>Not unique</u>: Stress is invariant under translations, orthogonal transformations (rotations, reflections) and overall re-scalings (blow-shrink) of the *W*'s.
(Solution for *k+1* builds on solution for *k*.)

St(k) will decrease as *k* increases, being certainly 0 for *k* >= *min{T,N-1}*

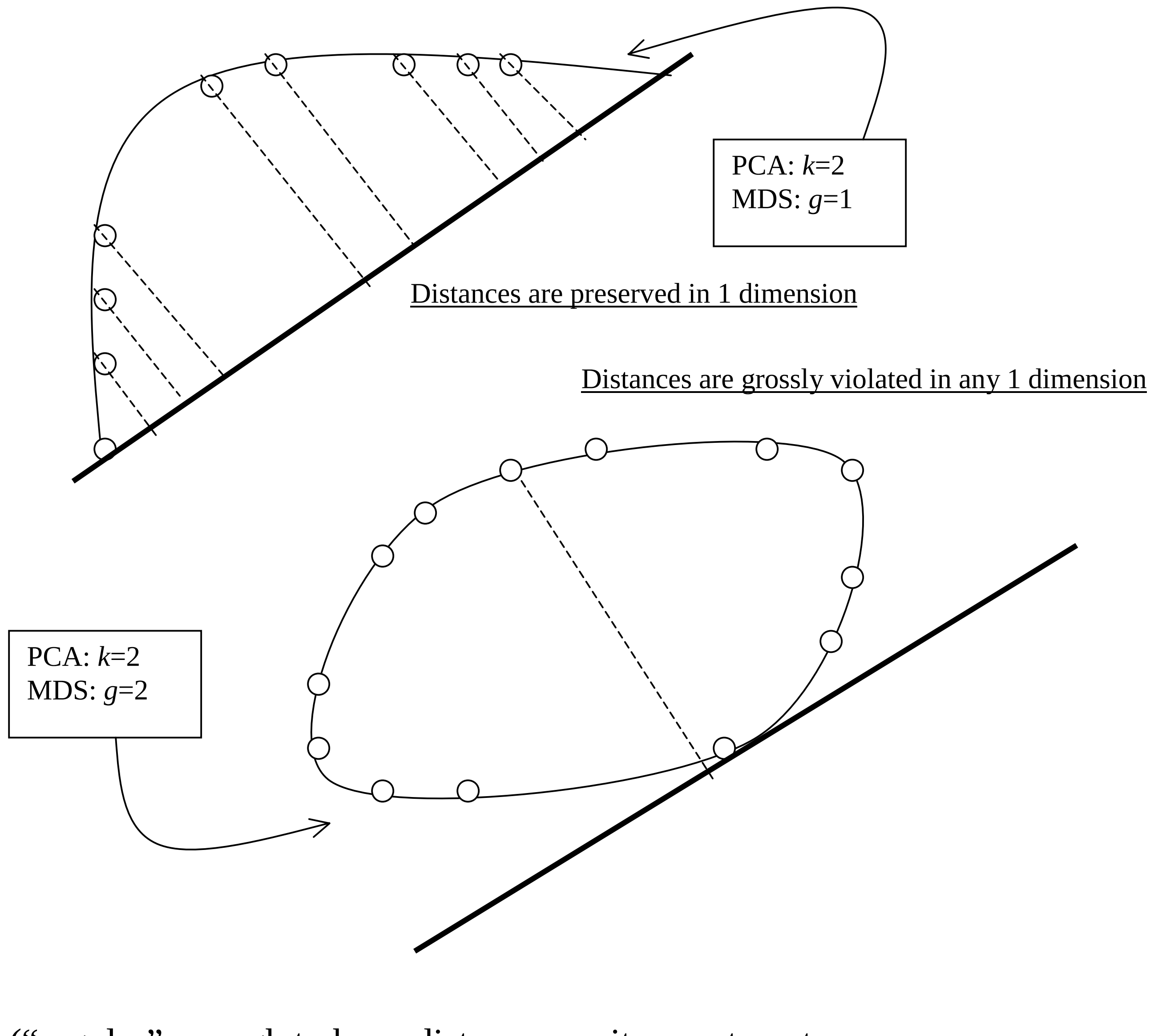
Plot and look for "negligible tails" and/or bends.

If points are very close to a k-dimensional subspace, so that projecting on it does preserve distances, PCA and MDS will have equivalent results: lead to *k* (e.g. 2), and



are the same, modulo translation, orthogonal transformation and overall re-scaling.

But MDS can reduce the dimension further if points are close to "regular" regions of a $g<k$ dimensional manifold (embedded into a k-dimensional affine space)



("regular" enough to have distances on it monotone to g-dimensional Euclidean distances).

MDS can also be employed to assess dimensionality and provide a low-dimensional graphical representation when the starting point of the analysis is not a cloud of points in T dimensions, but a collection of N objects for which one can specify a consistent dissimilarity matrix.

Also, recalling that dimension reduction is <u>*NOT*</u> clustering, one may still want to reduce the dimension prior to clustering:

- To eliminate "artifacts" (un-wanted variation patterns)… then PCA may make sense, but need reasoning!

- Otherwise, MDS may present advantages, as its objective is to preserve distances among points (as opposed to variability: there is in principle no reason why interesting clustering should occur in linear sub-regions of large variability).

Reference:

R. Gnanadesikan: *Methods for statistical data analysis of multivariate observations*. Wiley.

MDS is not implemented in Minitab, but it is implemented in S+.

<u>Extension of MDS</u> that allows one to capture “less-regular” low-dimensional manifolds:

J. Tenenbaum, V. De Silva, J.C. Langford (2000)
A global geometric framework for non-linear dimensionality reduction. *Science* **290**, 2319—2323.

“… builds on classical MDS, but seeks to preserve the intrinsic geometry of the data, as captured in the geodesic manifold distances between all pairs of data points”.

See also, based on a different principle:

S.T. Roweis, L.K. Saul (2000)
Non-linear dimensionality reduction by locally linear embedding. *Science* **290**, 2323—2326.

<u>Factor Analysis</u>:

Introduce a <u>decomposition model</u>: additive superposition of a *structural* and a *structure-void* term, uncorrelated to one another

$$X_i - \overline{X} = X_{i,o} + \varepsilon_i \quad , \quad X_{i,o}, \varepsilon_i \in R^T$$

$$\overline{X}_o = \overline{\varepsilon} = 0_T$$

$$S_o = \frac{1}{N}\sum_{i=1}^{N} X_{i,o} X_{i,o}' \quad , \quad S_\varepsilon = \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i \varepsilon_i'$$

$$\frac{1}{N}\sum_{i=1}^{N} \varepsilon_i X_{i,o}' = \underline{0}$$

Cross-covariances

This induces an additive decomposition of the var/cov matrix

$$S = \frac{1}{N}\sum_{i=1}^{N} (X_{i,o} + \varepsilon_i)(X_{i,o} + \varepsilon_i)' = \underbrace{S_o}_{\text{Structural component}} + \underbrace{S_\varepsilon}_{\text{Structure-void component}}$$

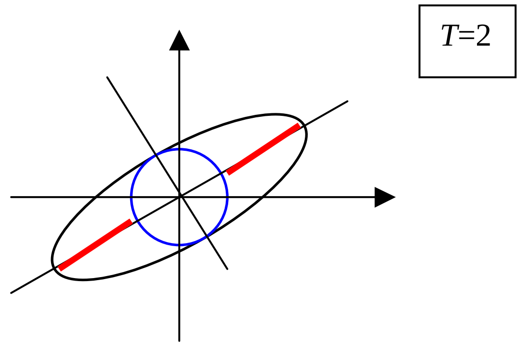
The idea is that the $X_{i,o}$'s actually live in a low dimension:

$Span(S_o)$, $dim(S_o) = K < T$.

Issue: the terms in the decomposition of the profiles and thus the components in the decomposition of the var/cov matrix are ***unobservable***.

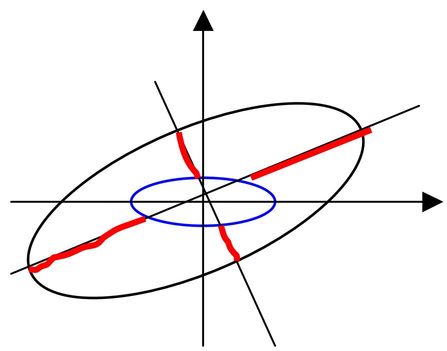
Spherical structure-void var/cov component (structure: departure from sphericity, which involves both correlations and relative spreads along the original coordinate axes)

$$S_\varepsilon = \sigma^2 I_T \quad , \quad \sigma^2 \geq 0$$



Diagonal structure-void var/cov component -- with respect to the original coordinate axes (structure: departure from diagonality in the original coordinate basis, which involves correlations)

$$S_\varepsilon = D(\sigma_j^{\,2}) \quad , \quad \sigma_j^{\,2} \geq 0, j = 1 \ldots T$$



This is the foundation of **Factor Analysis**

Going one step further: *Span(*S_o*)* in bi-jection with R^k, through a choice of orthonormal basis, and write

$$X_{i,o} = \Delta F_i$$

$$\Delta \quad T \times K, \quad F_i \in R^K, \quad \bar{F}^* = 0_K, \quad \frac{1}{N}\sum_{i=1}^{N} F_i F_i' = I_K$$

$$S_o = \Delta\Delta'$$

Coordinates in which the F_i’s are expressed: **<u>latent factors</u>**
K values in each specific F_i : **<u>factor scores</u>** for the *i*th observation
Entries in Δ : **<u>loadings</u>**:

$$X_{i,1,o} = \delta_{1,1} F_{i,1} + \ldots + \delta_{1,K} F_{i,K}$$

How the first factor “loads into” the structural part of the first original coordinate

$$\vdots$$

$$X_{i,T,o} = \delta_{T,1} F_{i,1} + \ldots + \delta_{T,K} F_{i,K}$$

Latent factors (choice of orthonormal basis), factor scores for each observation and loadings are **<u>not unique</u>**, : our decomposition is invariant under rotations in K dimensions (changing orthonormal basis):

$$X_{i,o} = \Delta F_i = \Delta \Theta \Theta' F_i = \Delta^* F_i^*$$

$$\Delta^* \quad T \times K, \quad F_i^* \in R^K, \quad \bar{F}^* = 0_K, \quad \frac{1}{N}\sum_{i=1}^{N} F_i^* F_i^{*}{}' = I_K$$

$$S_o = \Delta\Delta' = \Delta\Theta\Theta'\Delta' = \Delta^*\Delta^{*}{}'$$

Decomposition of original coordinates' variances:

$$s_j^2 = \left(\sum_{l=1}^{K} \delta_{j,l}^2 \right) + \sigma_j^2 \quad , \quad j = 1...T$$

"Communality"

Specific variance (the same for all j's in the spherical case)

Also, in terms of spectral decompositions

Tail *T-K* eigenval's of S equal to σ^2. Eigendirections of S_o compatible with those of S.

$$S_\varepsilon = \sigma^2 I_T$$

$$S = \sum_{j=1}^{T} \lambda_j V_j V_j' = \left(\sum_{j=1}^{T} (\lambda_j - \sigma^2) V_j V_j' \right) + \sigma^2 I_T = \left(\sum_{j=1}^{K} (\lambda_j - \sigma^2) V_j V_j' \right) + \sigma^2 I_T$$

$$S_\varepsilon = D(\sigma_j{}^2) = \sum_{j=1}^{T} \sigma_j{}^2 e_j e_j'$$

$$S = \sum_{j=1}^{T} \lambda_j V_j V_j' = \left(\sum_{j=1}^{T} \lambda_{j,o} V_{j,o} V_{j,o}' \right) + \sum_{j=1}^{T} \sigma_j{}^2 e_j e_j' = \left(\sum_{j=1}^{K} \lambda_{j,o} V_{j,o} V_{j,o}' \right) + \sum_{j=1}^{T} \sigma_j{}^2 e_j e_j'$$

Eigendirections of S_o not necessarily compatible with those of S.

Because of non-observability, the issue is now how to estimate the components of *S*, and chose the dimension (*K*) and an appropriate basis to express the (non-unique) loadings.

Reference: R. Gnanadesikan
Implemented in Minitab.

Projection pursuit and tours:

Exploratory approach.

One, or a whole sequence, of 2D projections chosen according to a criterion (e.g. departures from normality; collection of local maxima).

Looking at the high-dimensional data cloud from a sequence of "viewpoints" that ought to be structurally informative.

A couple of references:

C. Posse (1995). Tools for two-dimensional exploratory projection pursuit. *JCGS*, v.4 n.2.

A. Buja, Cook D., Swayne D.F. (1996) Interactive high dimensional data visualization. *JCGS*, v.5 n. 1. 1996

(see "historic" references therein)