**Dimension Reduction** (C)

Hints at other dimension reduction techniques. Recall in PCA, the aim is to find a low-dimensional (projective) representation of the data that preserves variability.

**Multi-dimensional Scaling** (MDS):

Find a low-dimensional representation of the data that preserves “relative positioning” of the points, i.e. distances among them.

\[ X_i \in R^T, \quad i = 1...N \]
\[ \delta_{il} = d(X_i, X_l) \]
\[ \delta_{(il)1} \leq \delta_{(il)2} \ldots \leq \delta_{(il)m}, \quad m = \frac{N(N-1)}{2} \]

\[ W_i, \hat{W}_i \in R^k, \quad i = 1...N \]
\[ d_{il} = d(W_i, W_l), \quad \hat{d}_{il} = d(\hat{W}_i, \hat{W}_l) \]

\[ \text{St}(W_1,...W_N) = \min_{\hat{W}'s; \hat{d}'s \text{ as close as monotoneto } \delta's} \left( \sum_{i<l} (d_{il} - \hat{d}_{il})^2 \right) \]

\[ \text{St}(k) = \min_{W's} \text{St}(W_1,...W_N) \rightarrow W_1^*,...,W_N^* \]

*Not unique:* 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 $k \geq min\{T,N-1\}$

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)

\[
\text{PCA: } k=2 \\
\text{MDS: } g=1
\]

Distances are preserved in 1 dimension

\[
\text{PCA: } k=2 \\
\text{MDS: } g=2
\]

Distances are grossly violated in any 1 dimension

(“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 \textit{NOT} 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).

Textbook:
R. Gnanadesikan: \textit{Methods for statistical data analysis of multivariate observations}. Wiley.

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

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

\[
X_i - \bar{X} = X_{i,o} + \varepsilon_i , \quad X_{i,o}, \varepsilon_i \in R^T
\]

\[
\bar{X}_o = \bar{\varepsilon} = 0_T
\]

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

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

This induces an additive decomposition of the var/cov matrix

\[
S = \frac{1}{N} \sum_{i=1}^{N} (X_{i,o} + \varepsilon_i)(X_{i,o} + \varepsilon_i)' = S_o + S_\varepsilon
\]

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

\[
\text{Span}(s_o) , \quad \text{dim}(s_o) = K < T .
\]

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)

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

**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)

\[ S_\varepsilon = D(\sigma_j^2), \quad \sigma_j^2 \geq 0, \quad j = 1...T \]

This is the foundation of **Factor Analysis**
Going one step further: \( \text{Span}(s_o) \) 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: latent factors

\( K \) values in each specific \( F_i \): factor scores for the \( i \)th observation

Entries in \( \Delta \): loadings:

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

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

Latent factors (choice of orthonormal basis), factor scores for each observation and loadings are not unique: 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:

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

“Communality”

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

Also, in terms of spectral decompositions

\[ S_{\epsilon} = \sigma^2 I_T \]

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

\[ S_{\epsilon} = D(\sigma_j^2) = \sum_{j=1}^{T} \sigma_j^2 e_j e_j' \]

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

Tail T-K eigenval’s of S equal to \( \sigma^2 \).

Eigendirections of \( S_o \) compatible with those of \( S \).

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.

Textbook: R. Gnanadesikan. Implemented in Minitab.
**Extensions:**

**Projection pursuit and tours**

- Exploratory approach (some references in reading list)

- One, or a whole sequence of, 2D projection(s) 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.

**Non-linear dimension reduction**

- Principal curves (Hastie and Tibshirani, 1989).

- MDS using an approximation to the geodesic distance, to capture “less-regular” low-dimensional manifolds. (Tenenbaum *et al.* 2000)

- Using local linear embeddings (Roweis and Saul, 2000)