How Many Components?

- How many components are important?
  - No definitive answer in the PCA framework.
  - The *factor analysis* model allows maximum likelihood estimation, hence hypothesis testing.

- For insight, use all that have a substantive interpretation.

- For other uses such as regression, we need an objective rule.
J&W: use “scree” plot (plot of eigenvalues), and look for an “elbow”; very subjective.

Simple rule of thumb: use all eigenvalues larger than the average.

   – Note: for the correlation matrix, the average is always 1, so the rule is: use all eigenvalues $> 1$.

Overland and Preisendorfer (1982) proposed a rule (“Rule N”) based on comparison of observed eigenvalues with the distribution of eigenvalues in the case of iid $N(0, \sigma^2)$. 
Large Samples

• Suppose that the rows of the data matrix $X$ are a random sample of size $n$ from $N_p(\mu, \Sigma)$.

• Assume that $\Sigma$ has distinct eigenvalues

$$\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0.$$  

• Then, approximately for large $n$,

$$\sqrt{n} \left( \hat{\lambda} - \lambda \right) \sim N_p \left[ 0, 2 \times \text{diag} \left( \lambda^2 \right) \right].$$
• In other words, $\hat{\lambda}_i$ and $\hat{\lambda}_k$ are approximately independent for $k \neq i$, and

$$\hat{\lambda}_i \sim N \left( \lambda_i, \frac{2\lambda_i^2}{n} \right).$$

• Note: if

$$\frac{n\bar{\lambda}}{\lambda} \sim \chi_n^2$$

then similarly, approximately,

$$\bar{\lambda} \sim N \left( \lambda, \frac{2\lambda^2}{n} \right).$$
• So we could also state that, approximately,

\[ \frac{n\hat{\lambda}_i}{\lambda_i} \sim \chi_n^2. \]

• Simulations suggest that this is a better approximation for small \( n \), if the eigenvalues are well separated.

• Asymptotics suggest that the degrees of freedom for \( \hat{\lambda}_i \) could be \( n - i + 1 \) instead of \( n \).

• Also \( \sqrt{n} (\hat{e}_i - e_i) \) is approximately \( N_p (0, E_i) \), where

\[
E_i = \lambda_i \sum_{\substack{k=1 \atop k \neq i}}^p \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} \times e_k e_k'.
\]
• J&W discuss a test for equal correlations:

\[ H_0 : R = \begin{bmatrix}
1 & \rho & \ldots & \rho \\
\rho & 1 & \ldots & \rho \\
\vdots & \vdots & \ddots & \vdots \\
\rho & \rho & \ldots & 1
\end{bmatrix} \]

against the alternative of a general (unstructured) \( \Sigma \) or \( R \).

• \( H_0 \) is equivalent to equality of all eigenvalues of \( R \) but one.

• Likelihood ratio test (and Lawley’s test) give large-sample \( \chi^2 \) statistic with \((p + 1)(p - 2)/2\) d.f.
Other related tests of interest:

- *compound symmetry* is a variance-components structure; it is stronger, requiring equal correlations *and* equal variances;

- equivalent to equality of all eigenvalues of $\Sigma$ but one.

- *sphericity* is the necessary and sufficient condition for univariate repeated measures to be valid; it is a weaker condition: $\sigma_{i,j} = \left(\sigma_{i,i} + \sigma_{j,j}\right) / 2 - \tau^2$.

- no simple characterization in terms of eigen structure.