Principal Components – Part 2

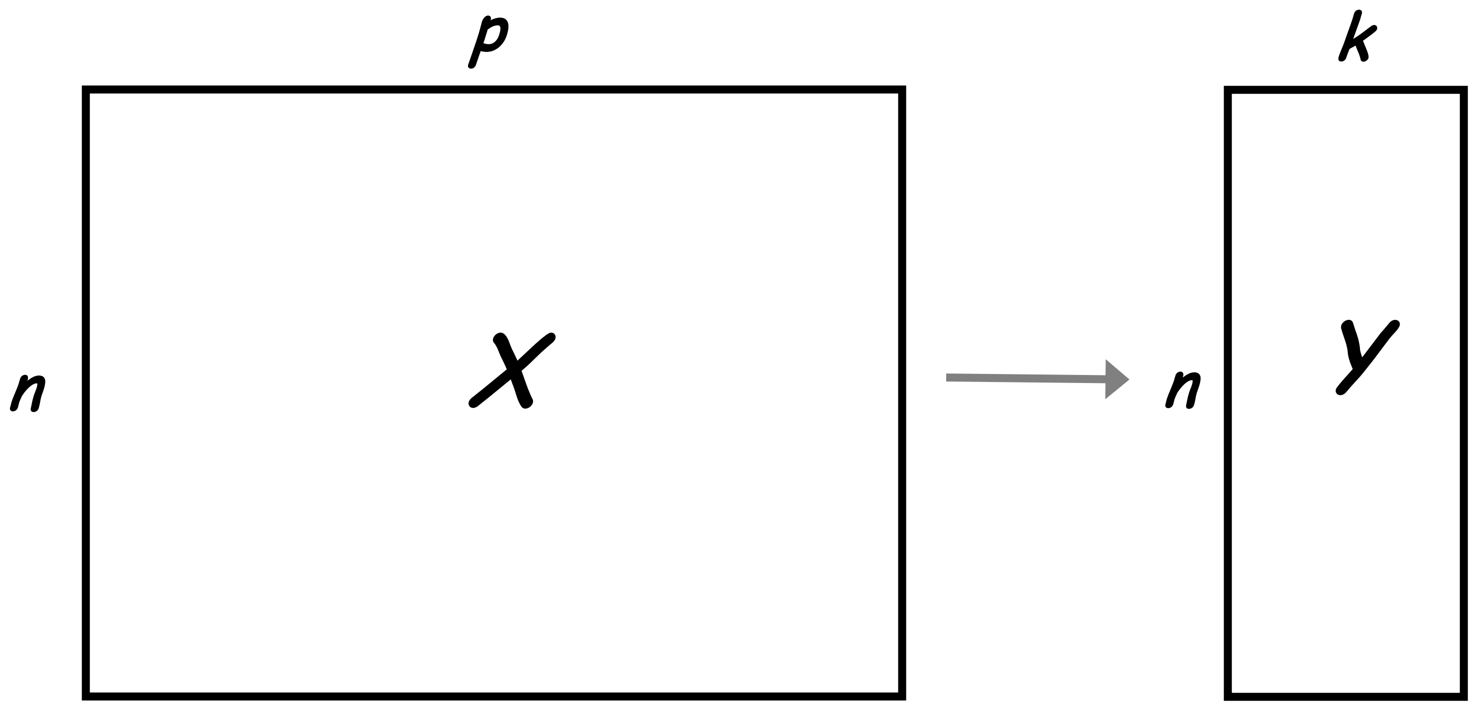
Recall the main idea of Principal Component Analysis (PCA)

Starts with a data matrix of n objects by p variables, which are typically highly correlated, and summarizes it by uncorrelated axes (principal components or principal axes) that are linear combinations of the original p variables

The first k components retain/display as much as possible of the variation among objects.

Data Reduction

Summarization of the data with many (p) variables by a smaller set of (k) derived (synthetic, composite) variables.



Data Reduction

"Residual" variation is information in ***X*** that is not retained in the principal components ***Y***

Balancing act between:

- clarity of representation, ease of understanding
- oversimplification: loss of important or relevant information.

The Number of Principal Components

Q: How many components to retain?

A: There is no definite answer ☹

Things to consider:

- % explained variance (e.g., you may aim for 85%)
- size of eigenvalues (e.g., when PCs are derived from the correlation matrix R, then usually we don't keep eigenvalues < 1)
- Visual inspection via a screeplot (remove eigenvalues smaller than the "elbow"/bend)
- Subject-matter interpretation
- Hypothesis test (see Rencher, Chapter 12)

Large Sample Inference

Confidence intervals for $\lambda_1 \geq \lambda_2 \geq \ \dots \geq \lambda_p \geq 0$ and $\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_p$ can be obtained when we assume the sample $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ is from a multivariate normal distribution.

Main result:

$$\sqrt{n}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) \approx \mathrm{N}_p(0, 2\boldsymbol{\Lambda}^2)$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix with the eigenvalues on the main diagonal, $\boldsymbol{\lambda}$ is the vector of population eigenvalues and $\hat{\boldsymbol{\lambda}}$ is the vector of estimated eigenvalues from the sample.

Confidence Intervals for λ

Result:

A large sample 100(1 - α)% CI for λ_i is:

$$\frac{\hat{\lambda}_i}{1 + z_{\frac{\alpha}{2}}\sqrt{\frac{2}{n}}} \leq \lambda_i \leq \frac{\hat{\lambda}_i}{1 - z_{\frac{\alpha}{2}}\sqrt{\frac{2}{n}}}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal distribution.

Ellipse Charts

It is possible to use the PCs to construct confidence ellipses for the pairs $(\hat{y}_{i1}, \hat{y}_{i2}),\ i = 1, \dots n$.

Formula:

$$\frac{\hat{y}_1^2}{\hat{\lambda}_1} + \frac{\hat{y}_2^2}{\hat{\lambda}_2} \leq \chi_2^2(\alpha)$$

These are used as contours to check for outliers in the dataset.