Dimensionality Reduction

Jay Urbain, PhD

Credits:

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) An Introduction to Statistical Learning, with applications in R, www.StatLearning.com, Springer-Verlag,

The Goals of Unsupervised Learning

- Goal: discover interesting things about objects and their attributes.
  - Is there an informative way to visualize the data?
  - Can we discover subgroups among the variables or among the observations?
- Two important methods:
  - Clustering, a broad class of methods for discovering unknown subgroups in data.
  - Principal components analysis, a tool used for data visualization or data pre-processing before supervised techniques are applied.

Challenge of Unsupervised Learning

- Unsupervised learning is more subjective than supervised learning: no simple goal for the analysis, such as prediction of a response.
- But techniques for unsupervised learning are of growing importance in a number of fields:
  - subgroups of breast cancer patients grouped by their gene expression measurements,
  - groups of shoppers characterized by their browsing and purchase histories,
  - movies grouped by the ratings assigned by movie viewers.

An advantage of unsupervised learning

- It is often easier to obtain unlabeled data — from a lab instrument or a computer — than labeled data, which can require human intervention.
- For example:
  - It is difficult to automatically assess the overall sentiment of a movie review: is it favorable or not?
  - There are relationships in the data beyond a simple rating.

Principal Components Analysis

- PCA produces a low-dimensional representation of a dataset. It finds a sequence of linear combinations of the variables that have maximal variance, and are mutually uncorrelated.
- Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization.

Principal Components

- Suppose we wish to visualize *n* observations with measurements on a set of *p* features, $X_1, X_2, \ldots, X_p$ as part of exploratory data analysis.
- Could examine 2-D scatter plots of the data, each of which contains the *n* observations' measurement on two of p.
- However there are *(n choose 2) = n*(n-1)/2* such scatter plots.
  - E.g., if *p=10,* there are *45* plots!
- Clearly, a better method is needed to visualize the *n* observations when *p* is large.

Principal Components

- We would like to find a low-dimensional representation of the data that captures as much of the information as possible.
- For instance, if we can obtain a two-dimensional representation of the data that captures most of the information, then we can plot the observations in this low-dimensional space.
- PCA – principal components analysis – provides such a tool.

Principal Components Analysis: details

- The **first principal component** of a set of features $X_1, X_2, \ldots, X_p$ is the normalized linear combination of the p features:

$$Z_1 = \varphi_{11}X_1 + \varphi_{21}X_2 + \ldots + \varphi_{p1}X_p$$

that has the largest variance. By normalized, we mean that

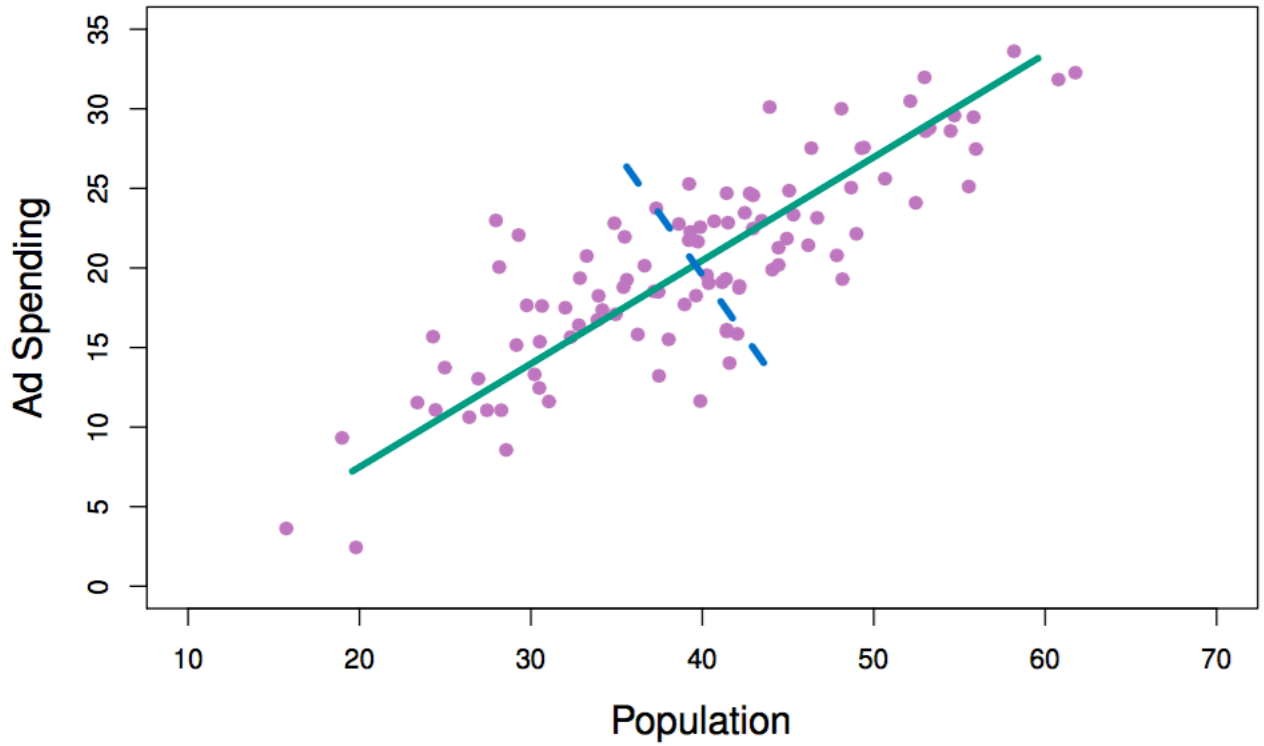
$$\sum_{j=1}^{p} \phi_{j1}^2 = 1$$

- We refer to the elements $\varphi_{11}, \ldots, \varphi_{p1}$ as the loadings of the first principal component.
- Together, the loadings make up the principal component loading vector:

$$\phi_1 = (\phi_{11}\ \phi_{21}\ \ldots\ \phi_{p1})^T$$

- The loadings are constrained so that their sum of squares is equal to one. Otherwise setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance.

PCA: Advertising Dataset Example



The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component direction, and the blue dashed line indicates the second principal component direction.

Computation of Principal Components

- Given an $n \times p$ data set X (n instances and p features)
- Since we are only interested in variance, assume that each of the variables in X has been centered to have mean zero (i.e., the column means of X are zero).
- We then look for the linear combination of the sample feature values of the form:

$$z_{i1} = \varphi_{11} X_{i1} + \varphi_{21} X_{i2} + \ldots + \varphi_{p1} X_{ip} \quad (1)$$

for $i = 1, \ldots, n$ that has largest sample variance, subject to the constraint that:

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1$$

Since each of the x_{ij} has mean zero, then so does z_{i1} (for any values of φ_{j1}).
Therefore the sample variance of the z_{i1} can be written as:

$$\frac{1}{n} \sum_{i=1}^{n} z_{i1}^2$$

Computation: continued

- Plugging in equation (1) the first principal component loading vector solves the optimization problem:

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^2 \text{ subject to } \sum_{j=1}^{p}\phi_{j1}^2 = 1$$

- This problem can be solved via a ***singular-value decomposition*** of the matrix *X*, a standard technique in linear algebra.
- We refer to Z_1 as the first principal component, with realized values $z_{11}, \ldots, z_{n1}$

Geometry of PCA

- The loading vector φ_1 with elements $\varphi_{11}, \varphi_{21}, \ldots, \varphi_{p1}$ defines a direction in p-dimensional feature space along which the data vary the most.
- If we project the n data points $x_1,\ldots,x_n$ onto this direction, the projected values are the principal component scores $z_{11},\ldots,z_{n1}$ themselves.

Further principal components

- The second principal component is the linear combination of $X_1, \ldots, X_p$ that has maximal variance among all linear combinations that are ***uncorrelated*** with Z_1.
- The second principal component scores $z_{12}, z_{22}, \ldots, z_{n2}$ take the form

$$z_{i2} = \varphi_{12} x_{i1} + \varphi_{22} x_{i2} + \ldots + \varphi_{p2} x_{ip},$$

- where φ_2 is the second principal component loading vector, with elements $\varphi_{12}, \varphi_{22}, \ldots, \varphi_{p2}$

Further principal components: continued

- Constraining Z_2 to be uncorrelated with Z_1 is equivalent to constraining the direction φ_2 to be orthogonal (perpendicular) to the direction φ_1. And so on.
- The principal component directions φ_1, φ_2, φ_3, . . . are the ordered sequence of *right singular vectors* of the matrix X, and the variances of the components are 1 times the n squared of the singular values.
- There are at most $min(n - 1, p)$ principal components.

Illustration: USA arrests data

- For each of the fifty states in the United States, the data set contains the number of arrests per 100,000 residents for each of three crimes: *Assault, Murder, and Rape.*
- Also record *Urban Pop* (the percent of the population in each state living in urban areas).
- The principal component score vectors have length $n = 50$, and the principal component loading vectors have length $p = 4$.
- PCA was performed after standardizing each variable to have mean zero and standard deviation one.

USAarrests data: PCA plot

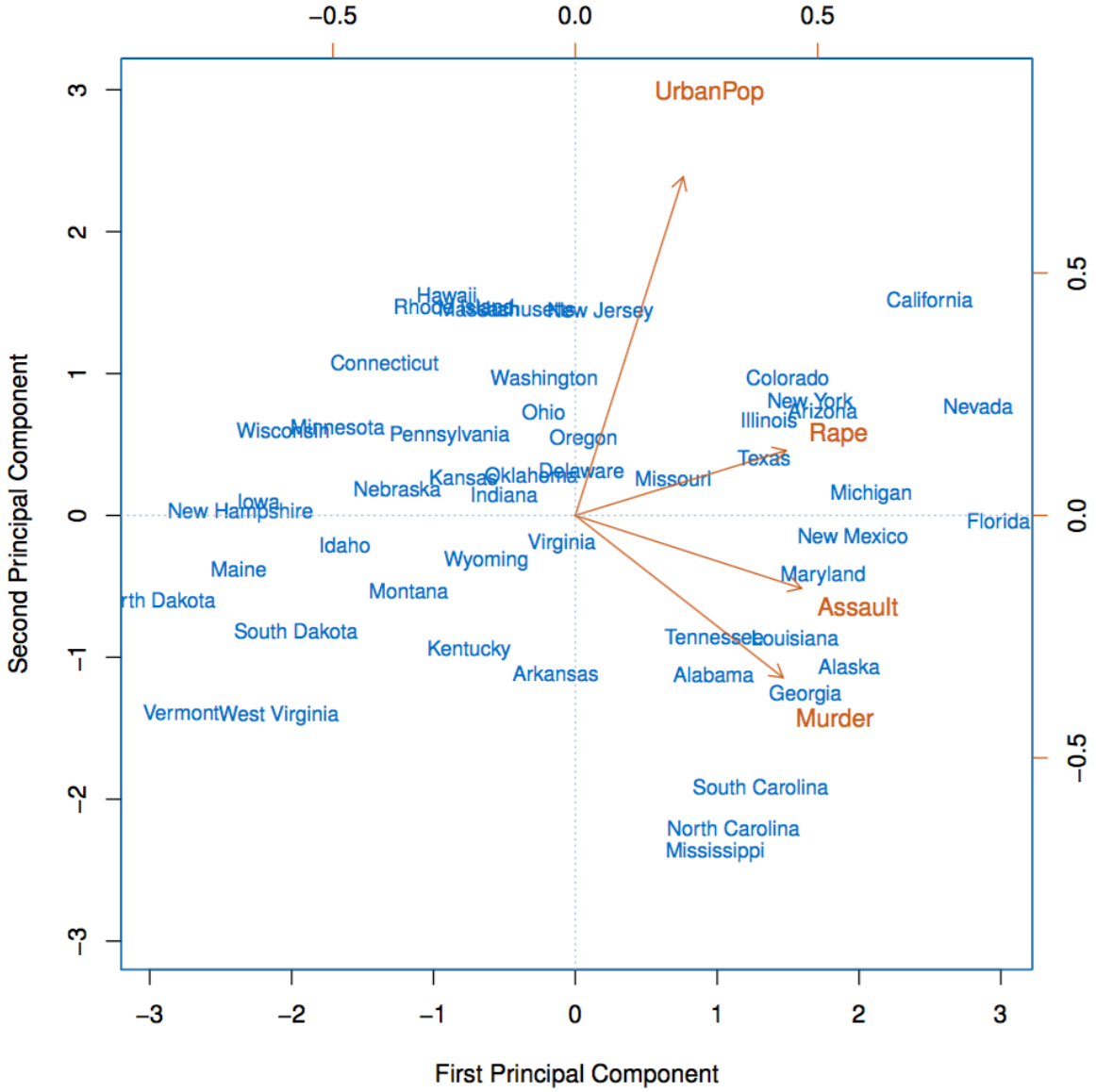


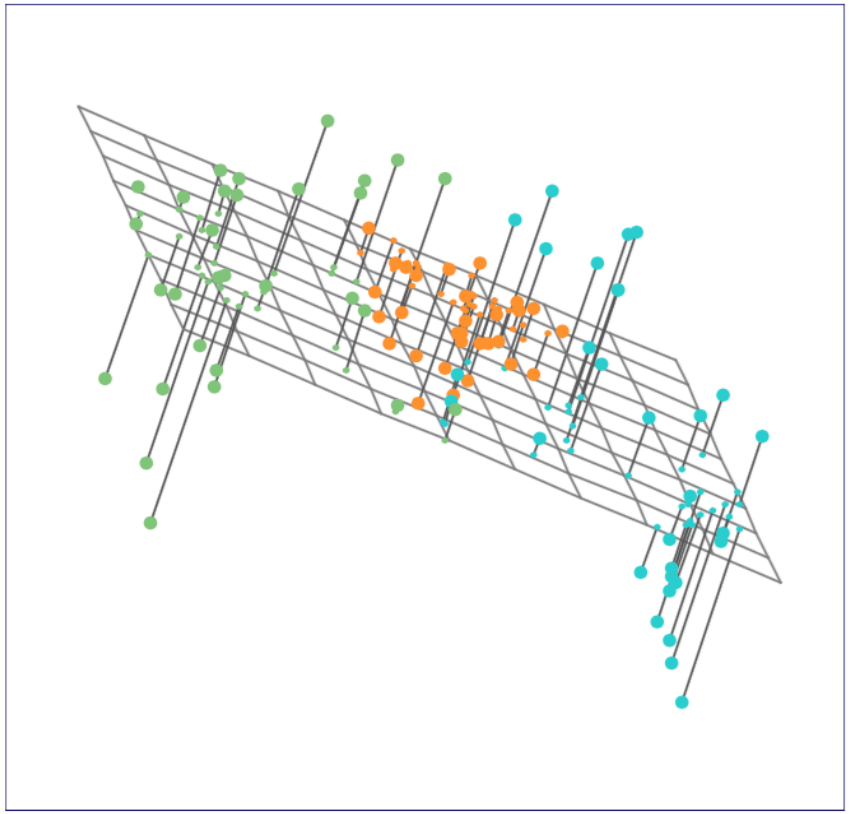
Figure details

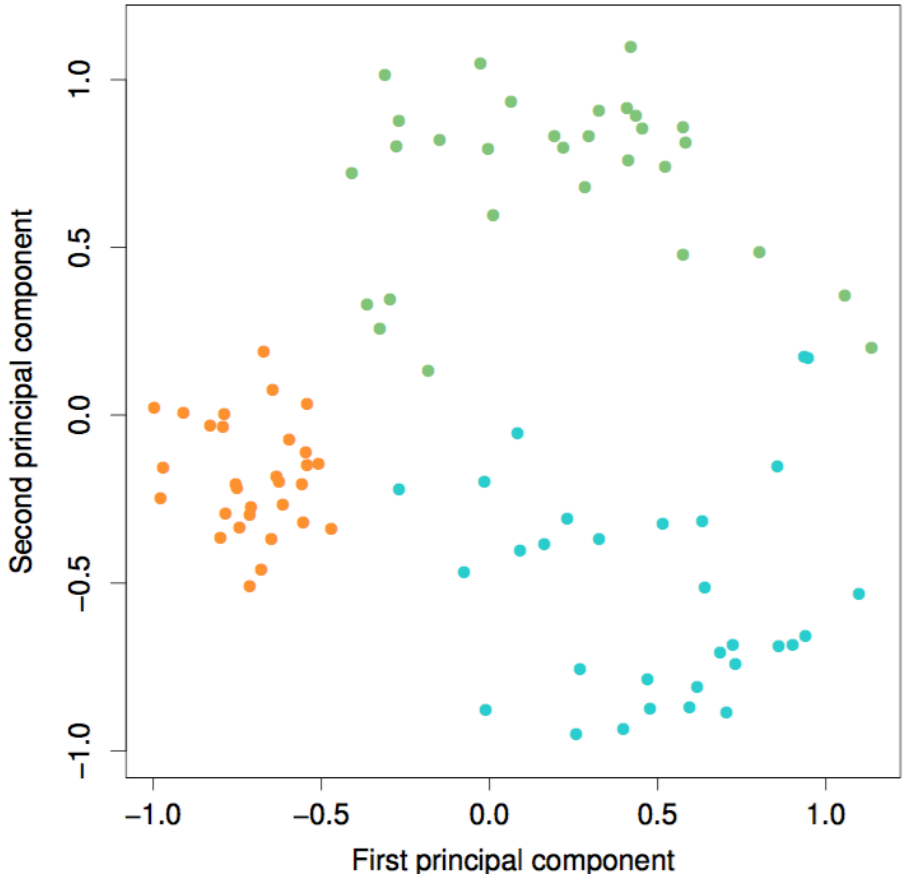
- First two principal components for the USArrests data.
- The blue state names represent the scores for the first two principal components.
- The orange arrows indicate the first two principal component loading vectors (with axes on the top and right).
  - For example, the loading for Rape on the first component is 0.54, and its loading on the second principal component 0.17 [the word Rape is centered at the point (0.54, 0.17)].
- This figure is known as a *biplot*, because it displays both the principal component scores and the principal component loadings.

PCA Loadings

	PC1	PC2
Murder	0.5358995	-0.4181809
Assault	0.5831836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

Another Interpretation of Principal Components





PCA find the hyperplane closest to the observations

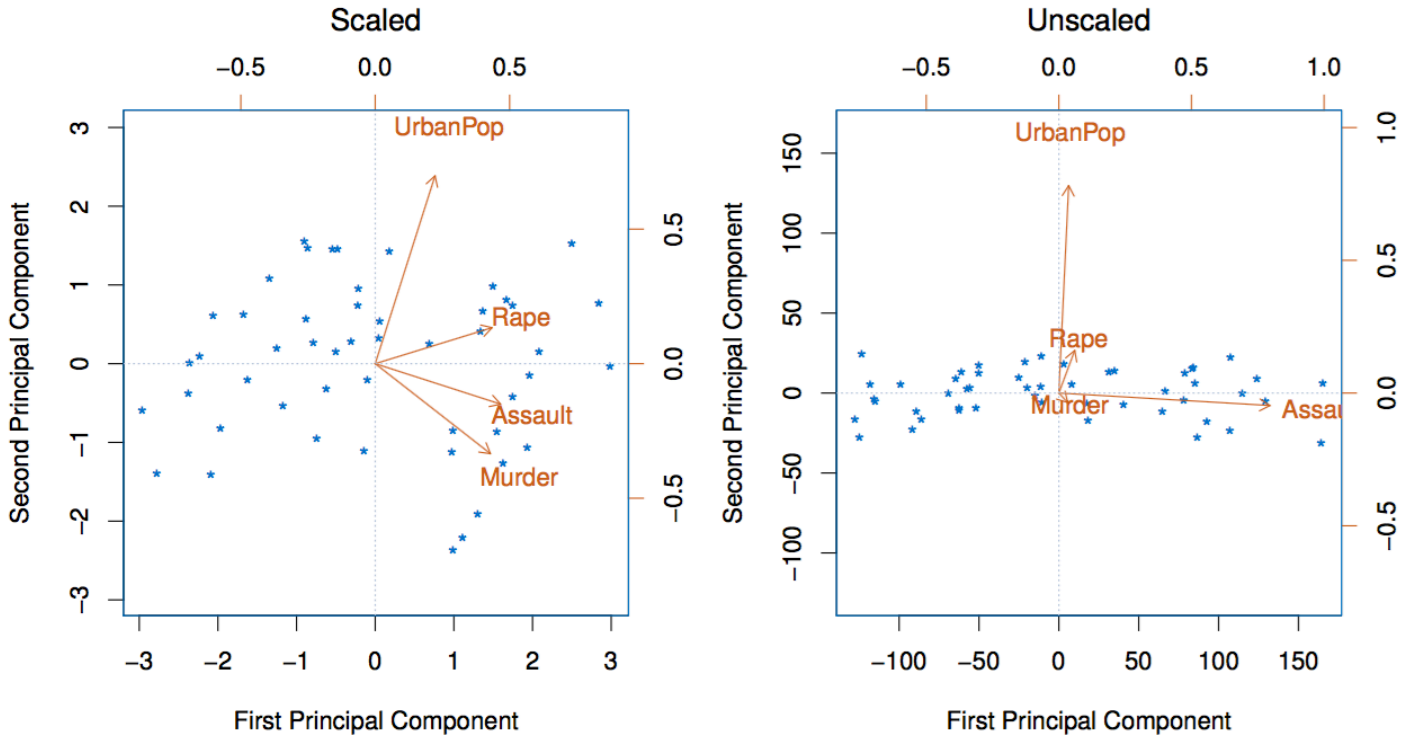
- The first principal component loading vector has a special property:
  - It defines the line in *p-dimensional* space that is closest to the *n* observations (using average squared Euclidean distance as a measure of closeness)
- The notion of principal components as the dimensions that are closest to the *n* observations extends beyond just the first principal component.
- I.e., the first two principal components of a data set span the plane that is closest to the *n* observations, in terms of average squared Euclidean distance.

Scaling of the variables is important

- If the variables are in different units, scaling each to have standard deviation equal to one is recommended.

$$z = \frac{x - \mu}{\sigma}$$

- If they are in the same units, you might (or might not) scale the variables.



Proportion Variance Explained

- To understand the strength of each component, we are interested in knowing the proportion of variance explained (PVE) by each one.
- The **total variance present in a data set** (assuming that the variables have been centered to have mean zero) is defined as:

$$\sum_{j=1}^{p} \mathrm{Var}(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$$

- and the variance explained by the m^{th} principal component is

$$\mathrm{Var}(Z_m) = \frac{1}{n} \sum_{i=1}^{n} z_{im}^2.$$

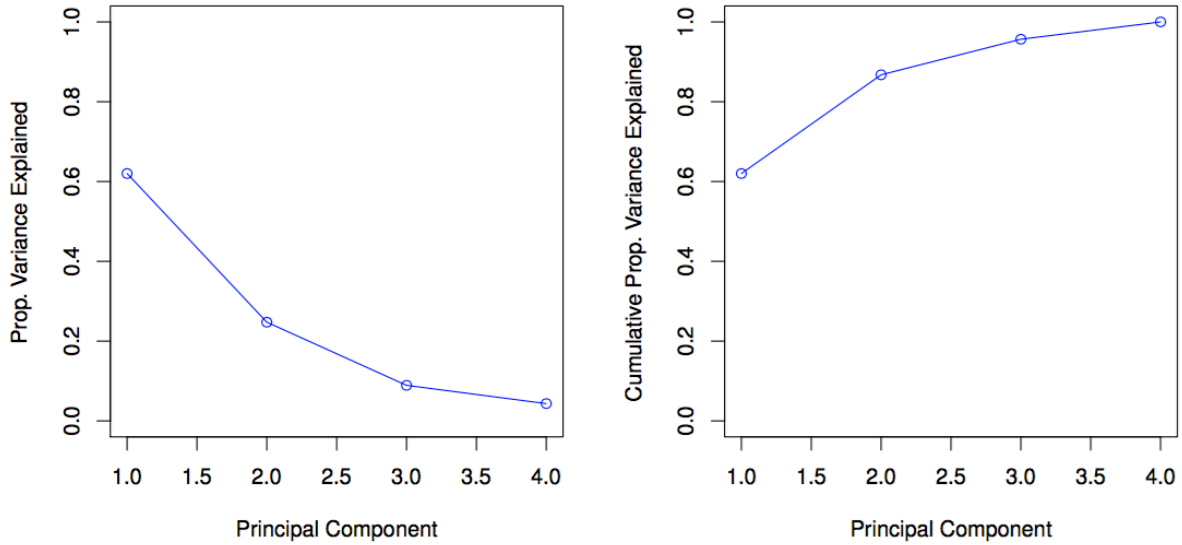
It can be shown that $\sum_{j=1}^{p} \mathrm{Var}(X_j) = \sum_{m=1}^{M} \mathrm{Var}(Z_m)$, with $M = \min(n-1, p)$.

Proportion Variance Explained: continued

- Therefore, the PVE of the m^{th} principal component is given by the positive quantity between 0 and 1.

$$\frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}$$

- The PVEs sum to one. We sometimes display the cumulative PVEs.



How many principal components should we use?

- If we use principal components as a summary of our data, how many components are sufficient?
- No simple answer to this question, as cross-validation is not available for this purpose.
- When could we use cross-validation to select the number of components?
- Proportion of variance explained: look for an “elbow”.

PCA vs Clustering

- PCA looks for a low-dimensional representation of the observations that explains a good fraction of the variance.
- Clustering looks for homogeneous subgroups among the observations.