Today

1. Introduction to unsupervised learning
2. Principal Components Analysis (PCA)
3. PCA v.s. Clustering Analysis

1. Introduction to Unsupervised learning
   - "Most of this course focuses on "supervised learning" methods such as regression and classification
   - We observe both a set of features (predictors, or variables) $X_1 \cdots X_p$ for each object, as well as a response or outcome variable Y. The goal is then to predict Y using $X_1, \cdots X_p$.

   $\Rightarrow$ Here, we instead focus on "unsupervised learning": we observe only the features $X_1, \cdots X_p$. We are not interested in prediction, because there is no associated response variable Y.

* The goal of Unsupervised Learning

Discover interesting things about the measurements:
- Is there an informative way to visualize the data?
- Can we discover subgroups among the variables or among the observations?

Two methods $\left\{ \right.$
- principal components analysis
  - a tool for data visualization or data pre-processing before supervised techniques are applied
- Clustering - discover unknown subgroups in data

* Challenge of Unsupervised Learning

- Unsupervised learning is more subjective than supervised learning, as there is no simple goal for the analysis, such as prediction of a responce.
- But techniques for unsupervised learning are of growing importance in a number of fields:
  - e.g. 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.

* Advantages of Unsupervised learning

- It is often easier to obtain "unlabelled data" (lab instrument or a computer), than labelled data, which can require human intervention.

e.g. it is difficult to automatically assess the overall sentiment of a movie review: is it favorable or not?

2 Principal Component Analysis (PCA)

- 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

* Details

- The first principal component of a set of features $X_1 \cdots X_p$ is the normalized linear combination of the features

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \cdots + \phi_{p1} X_p$$

Z_1 has the largest variance. By normalized, $\sum_{j=1}^{p} \phi_{j1}^2 = 1$.

$\Rightarrow$ $\phi_{11} \cdots \phi_{p1}$ are loadings of the first PC ; together, they make up the principal component loading vector
$$\underline{\phi}_1 = (\phi_{11}, \cdots \phi_{p1})^T.$$

(Note: we constrain the loadings so that their sum of squares is equal to one, since otherwise setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance.)

- Suppose $X_{n \times p}$ & each of the variables in X has been centered to have mean zero

look for the linear combination of the sample feature values of the form
$$z_{i1} = \phi_{11} x_{i1} + \cdots + \phi_{p1} x_{ip} \quad i = 1, \cdots n$$
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} $\Rightarrow$ the sample variance of z_{i1} can be written as $\frac{1}{n} \sum_{i=1}^{n} z_{i1}^2$

$\Rightarrow$ the first PC loading vector solves the optimization problem
$$\underset{\phi_{11} \cdots \phi_{p1}}{\text{maximize}} \ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

can be solved by a singular-value decomposition of X.

* Geometry of PCA
- the loading vector ϕ_1 with elements $\phi_{11}, \phi_{21} \cdots \phi_{p1}$ defines a direction in 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 $z_{i2} = \phi_{12} x_{i1} + \phi_{22} x_{i2} + \ldots + \phi_{p2} x_{ip}$

  where $\underline{\phi_2} = \{\phi_{12}, \phi_{22}, \phi_{32} \ldots \phi_{p2}\}$ is the second principal component loading vector, with elements $\phi_{12}\ \phi_{22} \ldots \phi_{p2}$
- It turns out that 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 $\phi_1, \ldots \phi_k \ldots$ are the ordered sequence of right singular vectors of the matrix X, and the variances of components are $\frac{1}{n}$ times the squares of the singular values. There are at most $\min(n-1, p)$ principal components.

* Interpretation of PCA

PCA find the hyperplane closest to the observations

- The first PC loading vector has a very 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 PCs as the dimensions that are closest to the n observations extends beyond just the first principal component.

- For instance, 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.

* Notes: ① scaling of the variables matters

- If the variables are in different units, scaling each to have standard deviation equal to 1 is recommended.
- 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 centred to have mean 0) is defined as

$$\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2}_{\text{"sample variance"}}$$

and the variance explained by the mth principal component is

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

$\Rightarrow$ It can be shown that $\sum_{j=1}^{p} Var(X_j) = \sum_{m=1}^{M} Var(Z_m)$

with $M = \min(n-1, p)$

- Therefore, the PVE of the mth principal component is given by the positive quantity between 0 and 1.

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

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

* How many principal components should we use?
  If we use PCs as a summary of our data, how many components are sufficient?
  - no simple ~~question~~ answer to this question, as cross-validation is not available for this purpose. (Why not?)
  - The "scree plot" can be used as a guide: we look for an "elbow"

3. PCA v.s. 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