Statistical Learning

Course 1 - Principal Components Analysis

Centrale de Lyon - Bachelor 2nd Year
2024-2025

Unsupervised Learning

Unsupervised vs Supervised Learning:

- The first part of this course focuses on supervised learning methods such as regression and classification.
- In that setting we observe both a set of features $X_1, X_2, ..., X_p$ for each object, as well as a response or outcome variable Y . The goal is then to predict Y using $X_1, X_2, ..., X_p$.
- Here we instead focus on unsupervised learning, we where observe only the features $X_1, X_2, ..., X_p$. We are not interested in prediction, because we do not have an associated response variable Y.

The Goals of Unsupervised Learning

- The goal is to 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? We discuss two methods:
  - principal components analysis, a tool used for data visualization or data pre-processing before supervised techniques are applied (this course)
  - clustering, a broad class of methods for discovering unknown subgroups in data (next course).

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 Analysis: details

- The first principal component of a set of features $X_1, X_2, ..., X_p$ is the normalized linear combination of the features

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

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

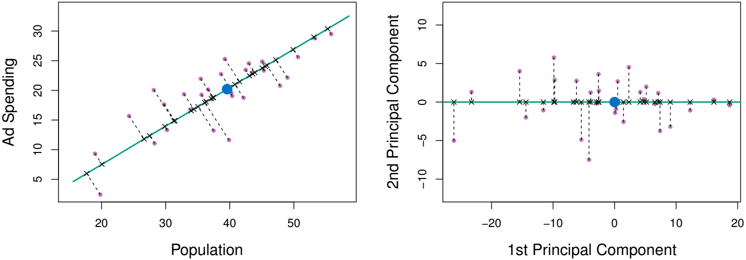

Figure: from An introduction to Statistical Learning

Principal Components Analysis: details

- We refer to the elements $\phi_{11}, ..., \phi_{p1}$ as the loadings of the first principal component; together, the loadings make up the principal component loading vector,

$$\phi_1 = (\phi_{11}, \phi_{21}, ..., \phi_{p1})'$$

.

- We constrain the norm of the principal component loading vector to equal to one, since otherwise setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance.

How to compute the Principal Components

- Suppose we have a $n \times p$ data set X. We assume that each of the variables in X has been centered to have mean zero (that is, the column means of X are zero, $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$ for all $j = 1, ..p$).
- We then look for the linear combination of the sample feature values of the form

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + ... + \phi_{p1}x_{ip} =< \phi_1, x_i > \qquad (1)$$

  for $i = 1, ..., n$ 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}). Hence the sample variance of the $(z_{i1})_{1 \leq i \leq n}$ can be written as $\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2$.

Computation: continued

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

$$maximize_{\phi_{11}\phi_{21}\ldots\phi_{p1}} \frac{1}{n}\sum_{i=1}^{n} z_{i1}^2 \text{ subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1.$$

  Remark that $\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^2$.

- This problem can be solved via solving the eigenvalues and eigenvectors of the sample covariance matrix $V = X'X$. See details

- We refer to Z_1 as the first principal component, with realized values $z_{11}, ..., z_{n1}$ (the scores).

Geometry of PCA

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

Further principal components

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

$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + ... + \phi_{p2}x_{ip},$$

where ϕ_2 is the second principal component loading vector, with elements $\phi_{12}, \phi_{22}, ..., \phi_{p2}$.

Further principal components: continued

- 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.

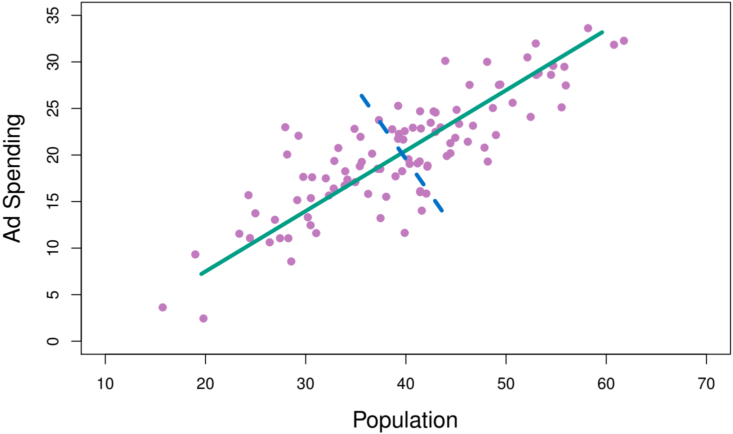

Figure: from An introduction to Statistical Learning

Further principal components: continued

- The principal component directions $\phi_1, \phi_2, \phi_3, ...$ are the sequence of eigenvectors associated with the inverse ordered sequence of eigenvalues $\lambda_1 > \lambda_2 > \lambda_3 ....$ of the matrix V, and the variances of *mth* components are $\frac{1}{n}\sum_{i=1}^{n} z_{im}^2 = \lambda_m$. There are at most $M = min(n-1, p)$ principal components.

PCA find the hyperplane closest to the observations

- The notion of principal components 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.

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 $\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$, and the variance explained by the *mth* principal component is

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

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

Proportion Variance Explained: continued

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

$$\frac{\frac{1}{n}\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{M} Var(Z_j)} = \frac{\lambda_m}{\sum_{j=1}^{M} \lambda_j}.$$

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

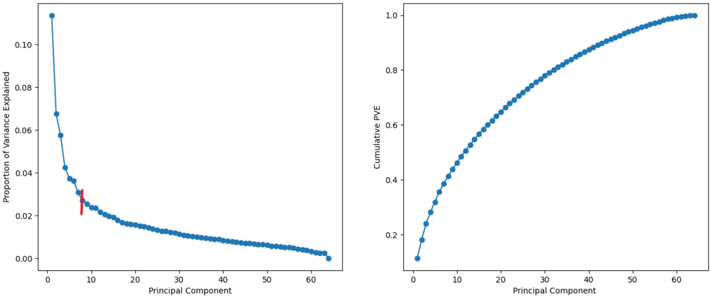

Figure: from An introduction to Statistical Learning

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.
- Most of the time the ’scree plot’ on the previous slide can be used as a guide: we look for an ’elbow’. In the previous graph we keep the first seven components.

Individuals Representation-Variables Correlation

We plot the score vector Z_1 against Z_2 geometrically, this amounts to projecting the original data down onto the space spanned by ϕ_1 and ϕ_2.

Correlation Circle : draw the Correlation Circle by computing the correlation between the orignal variables and the prinicpal components