Lecture 6: Principal Components

Måns Thulin

Department of Mathematics, Uppsala University
thulin@math.uu.se

Multivariate Methods • 2/5 2011

Outline

- Two examples
  - Arrhythmia data
  - National track records for women
- Finding the principal components
- Interpretation
- Some probability theory

Arrhythmia data

Cardiologists wish to study arrhythmia (abnormal heart activity).

For $n = 452$ patients, $p = 279$ variables were recorded, including age, sex, height, weight, heart beat rate and a number of heart beat related measures.

The goal of the study is to use these data to classify new patients as belonging to one of 16 groups – the patient should be classified either as healthy (no arrhythmia) or as having one of 15 kinds of arrhythmia.

(Classifications methods are studied in block 4!)

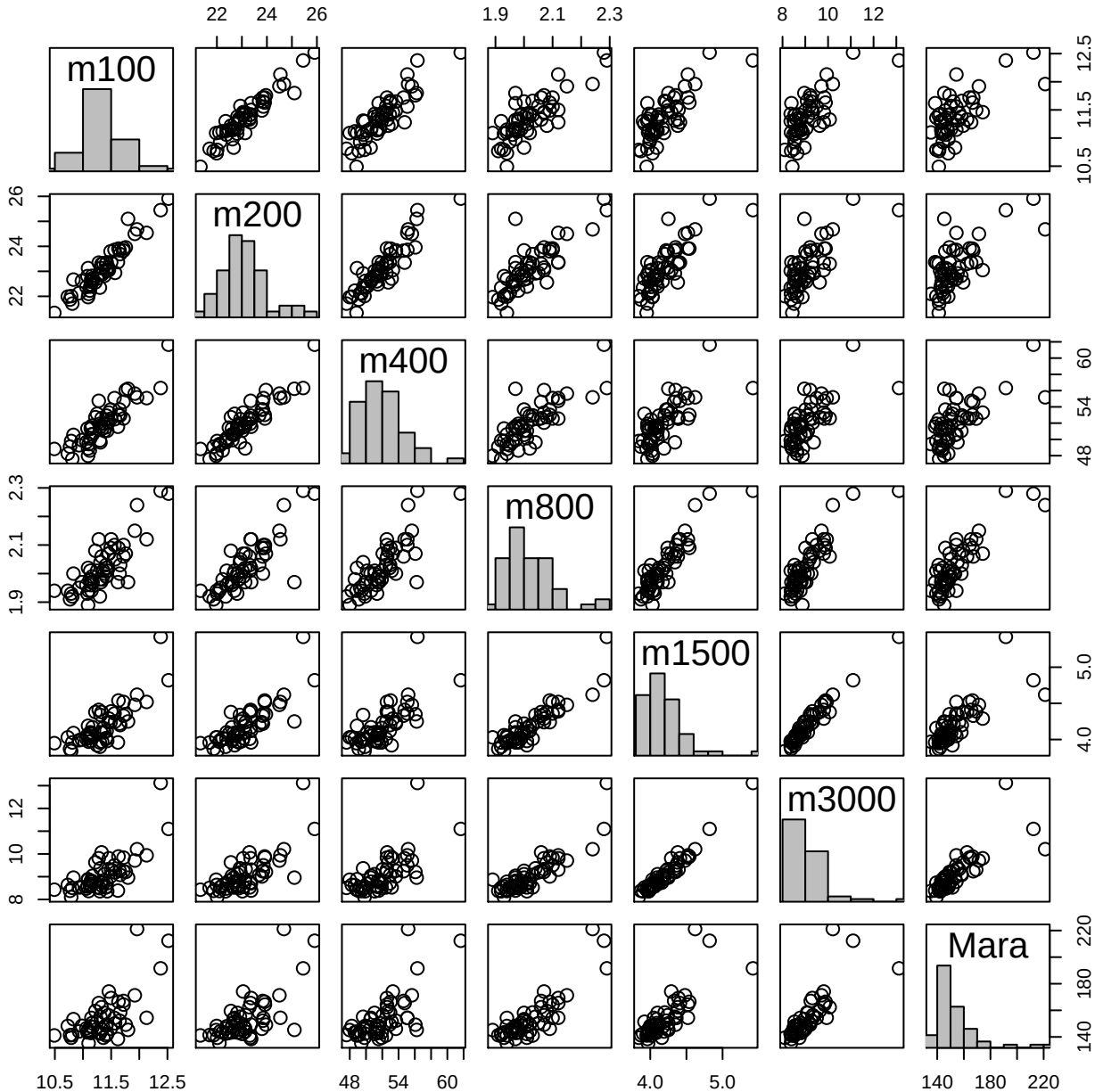
279 dimensions is extremely hard to visualize!

Is it possible to study only a subset of the variables without losing too much information? Can most of the information in the data set be described by a few linear combinations of the variables?

National track records for women

In the first computer exercise, we studied national track records for women for $p = 7$ distances and $n = 54$ countries.

National track records for women



National track records for women

In the first computer exercise, we studied national track records for women for $p = 7$ distances and $n = 54$ countries.

It is of interest to somehow summarize these data and develop an index of athletic performance for these countries.

With observations $\mathbf{x}_i = (x_{i1}, \ldots, x_{i7})'$, a natural idea is to look at the linear combination

$$\frac{1}{7}x_{i1} + \frac{1}{7}x_{i2} + \ldots + \frac{1}{7}x_{i7}.$$

With $\mathbf{a} = (1/7, 1/7, \ldots, 1/7)'$ we can write this as $\mathbf{a}'\mathbf{x}_i$.

Is there a better choice of $\mathbf{a}$?

Summarizing the data

For data sets like the two mentioned, we wish to:

- Reduce dimensionality
  - Visual investigation
  - Simplify further analysis
  - Detect outliers
  - Test for normality
  - Do inference when $p > n$
- Summarize the data
  - Which variables are important?
  - Create index

The general problem is to investigate linear combinations $\mathbf{a}'\mathbf{x}_i$.

Summarizing the data: choosing $\mathbf{a}$

We would like to find $\mathbf{a}_1, \ldots, \mathbf{a}_k$, where $k < p$, such that $\mathbf{a_1'x}_i, \ldots, \mathbf{a_k'x}_i$ contains as much information as possible.

- The "information" in the data set is the variation of the variables.
- **Idea:** Find $\mathbf{a_1}$ that maximizes the (sample) variance of $\mathbf{a_1'x}_i$.
  - There is a problem with this approach...
  - For $b > 1$, $b\mathbf{a_1'x}_i$ has greater variance than $\mathbf{a_1'x}_i$.
- **Revised idea:** Find $\mathbf{a_1}$, normalized such that $\mathbf{a_1'a_1} = \sum_{j=1}^{p} a_{1j}^2 = 1$, that maximizes the (sample) variance of $\mathbf{a_1'x}_i$.
  - How should we choose $\mathbf{a_2}$?
  - If we change a_{s1} and a_{t1} just a little, we can find an $\mathbf{a}$ with $\mathbf{a'a} = 1$ and that gives a variance close to that of $\mathbf{a_1'x}_i$...
  - It seems reasonable to require that $\mathbf{a_2}$ is uncorrelated to $\mathbf{a_1}$!
- Then find the normalized $\mathbf{a_2}$ that is (sample) uncorrelated to $\mathbf{a_1}$ and maximizes the variance of $\mathbf{a_2'x_i}$ under this constraint.
  - Then find the normalized $\mathbf{a_3}$ that is uncorrelated to both $\mathbf{a_1}$ and $\mathbf{a_2}$ and that maximizes the variance of $\mathbf{a_3'x_i}$ under this constraint. And so on...

Summarizing the data: choosing $\mathbf{a}$

How do we find the $\mathbf{a_1}, \ldots, \mathbf{a_k}$ that satisfy those criterions?

See blackboard!

Call $\mathbf{a_1'x}, \ldots, \mathbf{a_p'x}$ the prinicpal components.

Theorem. The ith principal component is given by

$$y_i = \mathbf{e_i'x} = e_{i1}x_1 + e_{i2}x_2 + \ldots + e_{ip}x_p$$

where $(\lambda_1, \mathbf{e_i}), \ldots, (\lambda_p, \mathbf{e_p})$ are the eigenvalue-eigenvector pairs of $\mathbf{S}$.

The sample variance of $y_i = \mathbf{e_i'x}$ is λ_i and the sample covariance of y_i and y_j, $i \neq j$ is 0.

Furthermore

$$tr(\mathbf{S}) = \sum_{i=1}^{p} s_{ii} = \lambda_1 + \lambda_2 + \ldots + \lambda_p \qquad \text{and}$$

$$r_{y_i,x_j} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{s_{jj}}}$$

Summarizing the data: how many PCs?

How many principal components should be used to describe the data?

- Minimum percentage of total variation: 50 %? 80 %? 90 %?
- Look at a screeplot
  - Plot of eigenvalues against the index of the corresponding prinicpal component.
  - Look for an "elbow".
- Several rules-of-thumb exist, but all lack theoretical support.

Summarizing the data: track records

Let's study the track records data with principal components.

See R code!

- Find principal components
- Rescale data to speeds and find PCs
- Use covariance matrix instead of correlation matrix

Summarizing the data: arrhythmia

Let's investige the arrhythmia data with principal components.

See R code!

- Using principal components, the data is reduced from 279 to 26 dimensions, with 90% of the variation accounted for.

Summarizing the data: interpretation

- In many cases the principal components allow nice interpretations. However, the principal components need not always be physically meaningful!
- If the data comes from a normal population, then the confidence region

$$n(\bar{\mathbf{X}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \leq \frac{p(n-1)}{n-p} F_{p,n-p}(\alpha)$$

  is an ellipse with axes given by the eigenvalues $\mathbf{e_i}$ of S. Looking at PCs is essentially the same thing as changing coordinate system to $\mathbf{e_1}, \ldots, \mathbf{e_p}$.
- Unusually small values for the *last* eigenvalue of $\mathbf{S}$ can indicate a linear dependency in the data set.

Population principal components

If the covariance matrix $\mathbf{\Sigma}$ is known, we can (should!) use this instead of $\mathbf{S}$.

In complete analogue to the sample covariance case, we find the following:

Theorem. The ith population principal component is given by

$$Y_i = \mathbf{e_i'X} = e_{i1}X_1 + e_{i2}X_2 + \ldots + e_{ip}X_p$$

where $(\lambda_1, \mathbf{e_i}), \ldots, (\lambda_p, \mathbf{e_p})$ are the eigenvalue-eigenvector pairs of $\mathbf{\Sigma}$.

$Var(Y_i) = \mathbf{e_i'X}$ is λ_i and $Cov(Y_i, Y_j) = 0$ for $i \neq j$,

$$tr(\mathbf{\Sigma}) = \sum_{i=1}^{p} \sigma_{ii} = \lambda_1 + \lambda_2 + \ldots + \lambda_p \qquad \text{and}$$

$$\rho_{y_i,x_j} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{jj}}}$$

Population and sample principal components

Assume that the observations are i.i.d. $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Furthermore, let $\mathbf{e_i}$ be the eigenvectors of $\boldsymbol{\Sigma}$ and $\hat{\mathbf{e}}_i$ be the eigenvectors of $\mathbf{S}$.

Then the sample principal component $y_i = \hat{\mathbf{e}}_i(\mathbf{x} - \bar{\mathbf{x}})$ is a realization of the population principal component $Y_i = \mathbf{e}_i(\mathbf{X} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}, \boldsymbol{\Lambda})$, where $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\lambda})$.

Large sample inference

Assume that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent and $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with eigenvalues $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$ of $\boldsymbol{\Sigma}$ distinct and positive.

Let $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_1 \ \hat{\lambda}_2 \ \ldots \ \hat{\lambda}_p)'$ be the vector of eigenvalues for $\mathbf{S}$. Then

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

- Estimated eigenvalues for $\boldsymbol{\Sigma}$ are asymptotically independent.
- $\hat{\lambda}_i \approx \mathrm{N}(\lambda_i, \ \frac{2}{n}\lambda_i^2)$
- An approximate $(1-\alpha)$ confidence interval for λ_i is

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

Large sample inference

Let $\mathbf{e_i}$ be the eigenvectors of $\mathbf{\Sigma}$ and $\hat{\mathbf{e}}_i$ be the eigenvectors of $\mathbf{S}$. Then

$$\sqrt{n}(\hat{\mathbf{e}}_i - \mathbf{e}_i) \rightarrow_d \mathrm{N}_p(\mathbf{0},\ \mathbf{E}_i)$$

where

$$\mathbf{E}_i = \lambda_i \sum_{k \neq i} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} \mathbf{e}_k \mathbf{e}_k'$$

For large n, $\hat{\lambda}_i$ is approximately independent of the distribution for $\hat{\mathbf{e}}_i$.

Summary

- Two examples
  - Arrhythmia data
  - National track records for women
- Finding the principal components
  - Eigenvectors and eigenvalues of **S**
- Interpretation
- Some probability theory