Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Revision: Chapter 1-6

Applied Multivariate Statistics – Spring 2012

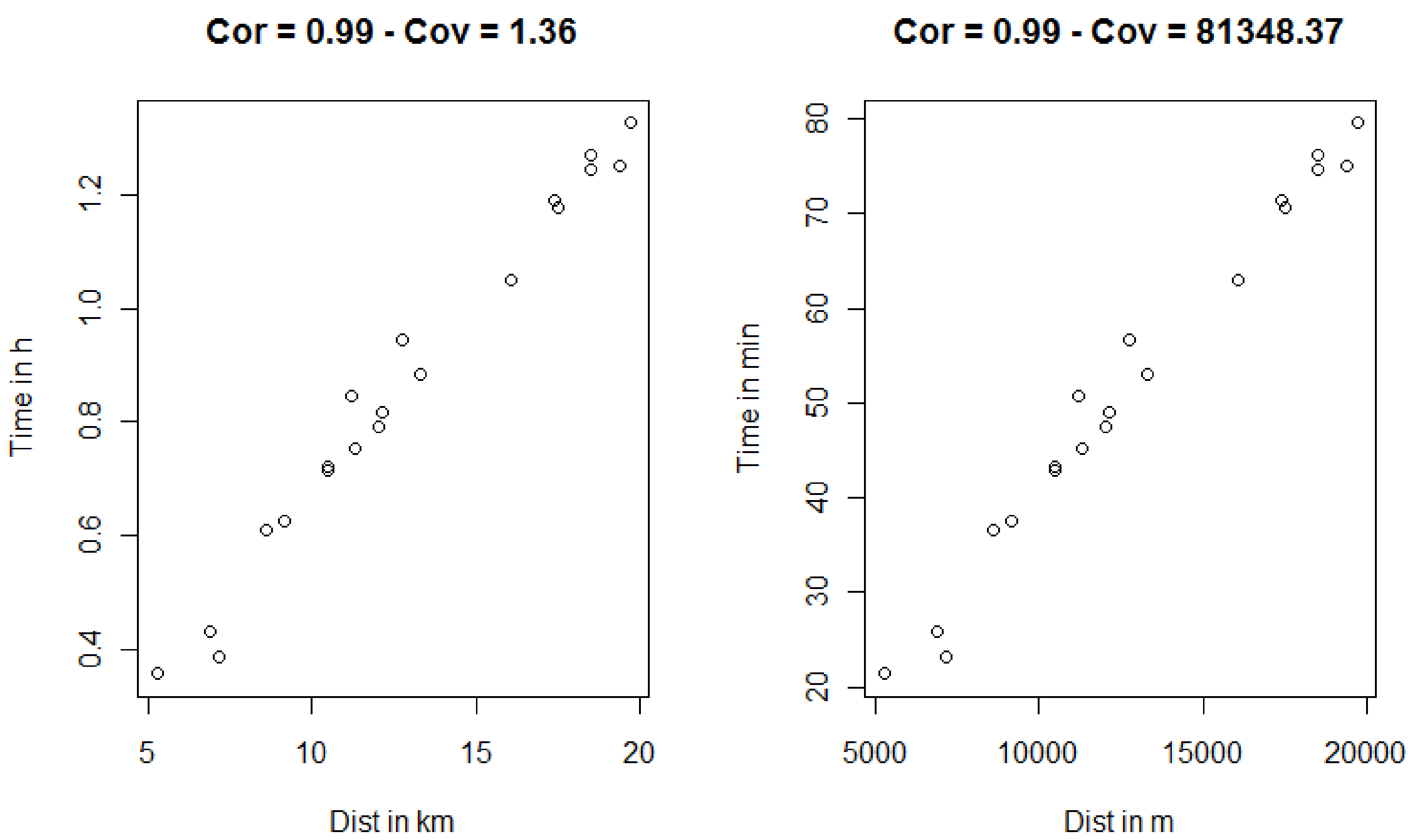
Overview

- Cov, Cor, Mahalanobis, MV normal distribution
- Visualization: Stars plot, mosaic plot with shading
- Outlier: chisq.plot
- Missing values: md.pattern, mice
- MDS: Metric / non-metric
- Dissimilarities: daisy
- PCA
- LDA

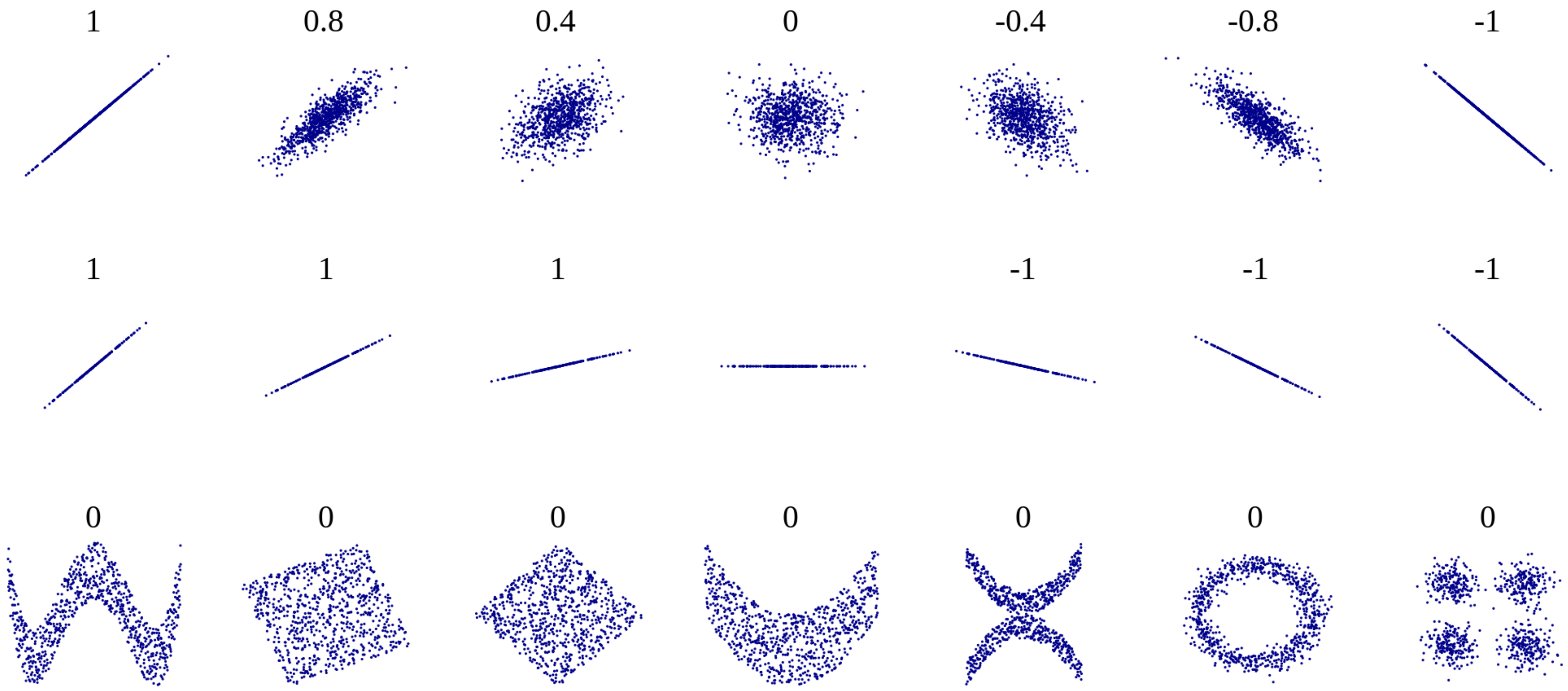
Two variables: Covariance and Correlation

- Covariance: $Cov(X,Y) = E[(X - E[X])(Y - E[Y])] \quad \in [-\infty; \infty]$
- Correlation: $Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \quad \in [-1; 1]$
- Sample covariance: $\widehat{Cov}(x,y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$
- Sample correlation: $r_{xy} = \widehat{Cor}(x,y) = \frac{\widehat{Cov}(x,y)}{\hat{\sigma}_x \hat{\sigma}_y}$
- Correlation is invariant to changes in units, covariance is not (e.g. kilo/gram, meter/kilometer, etc.)

Scatterplot: Correlation is scale invariant



Intuition and pitfalls for correlation
Correlation = LINEAR relation



Covariance matrix / correlation matrix: Table of pairwise values

- True covariance matrix: $\Sigma_{ij} = Cov(X_i, X_j)$
- True correlation matrix: $C_{ij} = Cor(X_i, X_j)$

- Sample covariance matrix: $S_{ij} = \widehat{Cov}(x_i, x_j)$
  Diagonal: Variances
- Sample correlation matrix: $R_{ij} = \widehat{Cor}(x_i, x_j)$
  Diagonal: 1

- R: Functions “cov”, “cor” in package “stats”

Multivariate Normal Distribution: Most common model choice

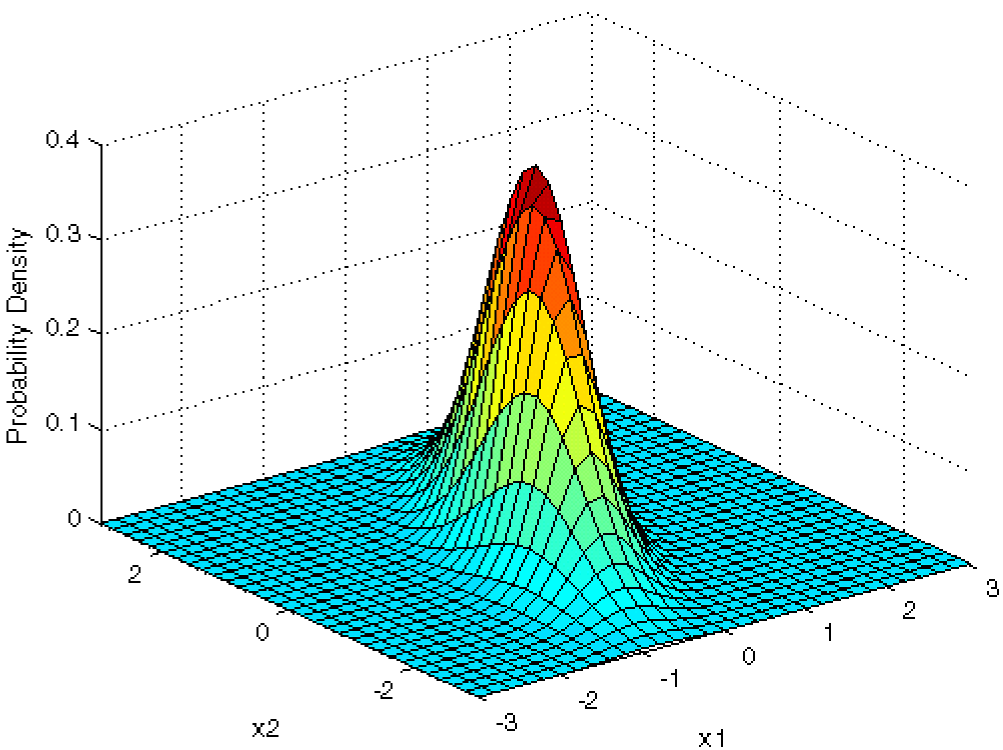
Sq. Mahalanobis Distance $MD^2(x)$

=

Sq. distance from mean in standard deviations

IN DIRECTION OF X

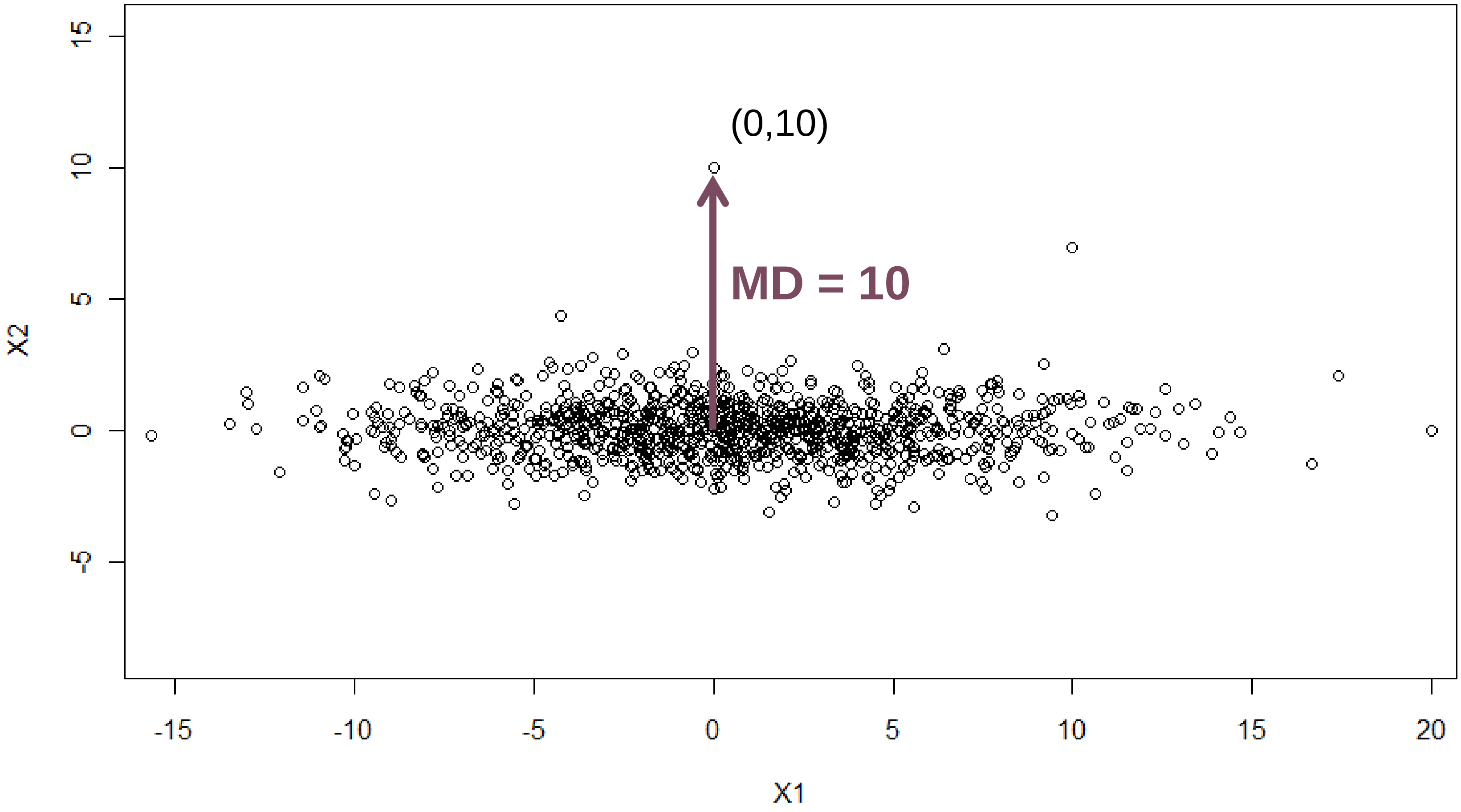
$$f(x;\mu,\Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left(-\frac{1}{2}\cdot \boxed{(x-\mu)^T\Sigma^{-1}(x-\mu)}\right)$$



Mahalanobis distance: Example

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

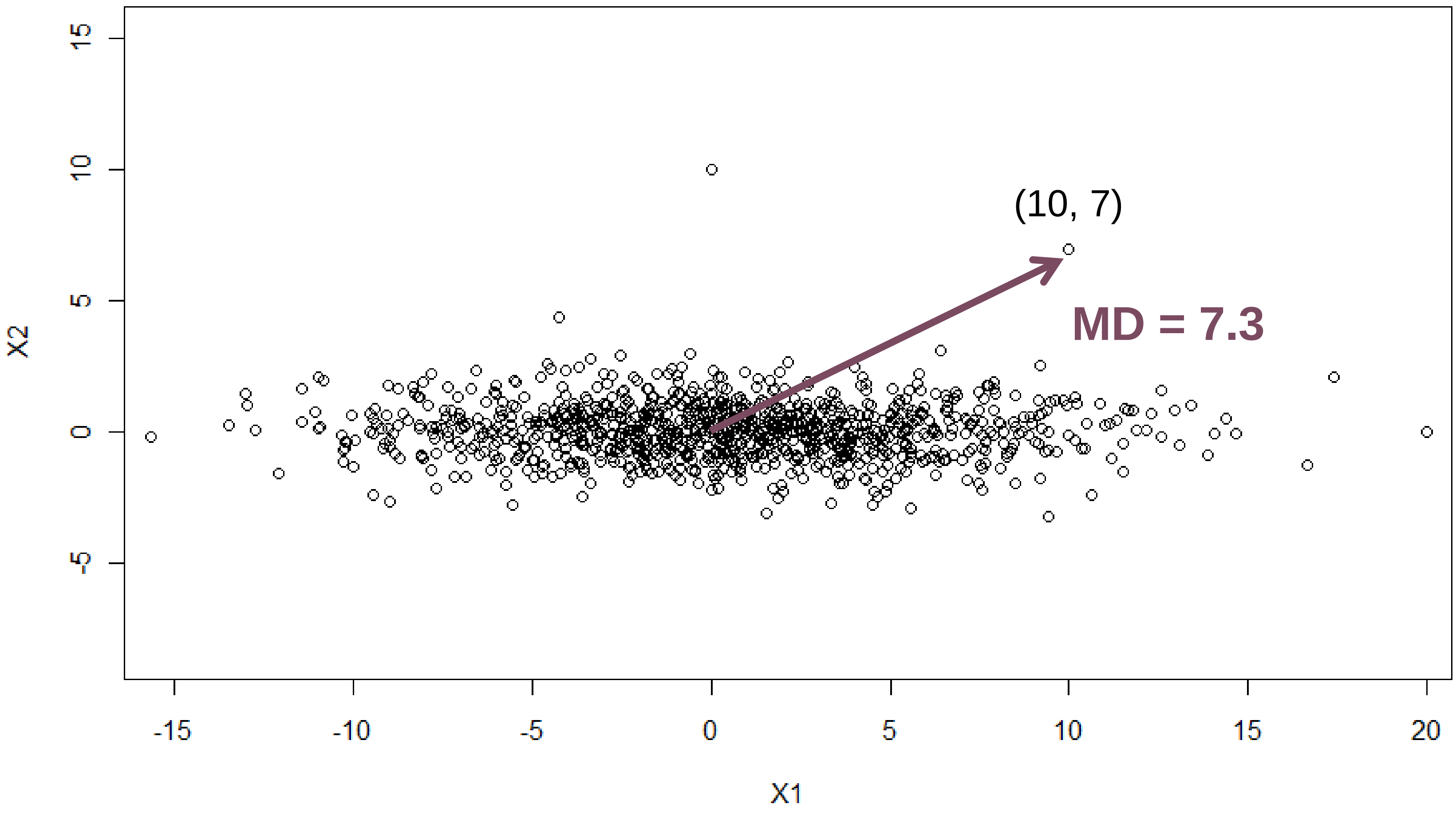
$$\Sigma = \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix}$$



Mahalanobis distance: Example

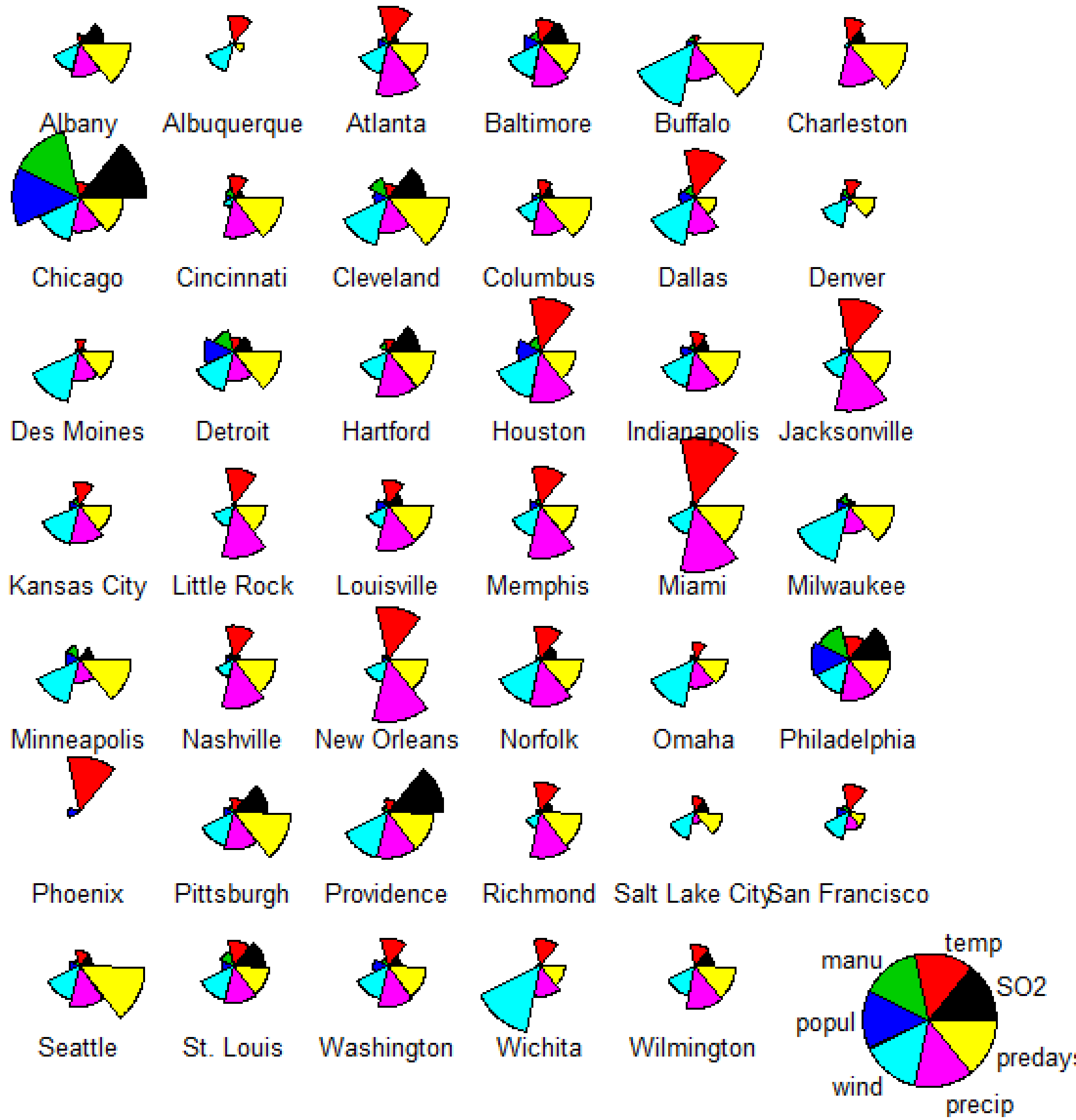
$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix}$$



Glyphplots: Stars

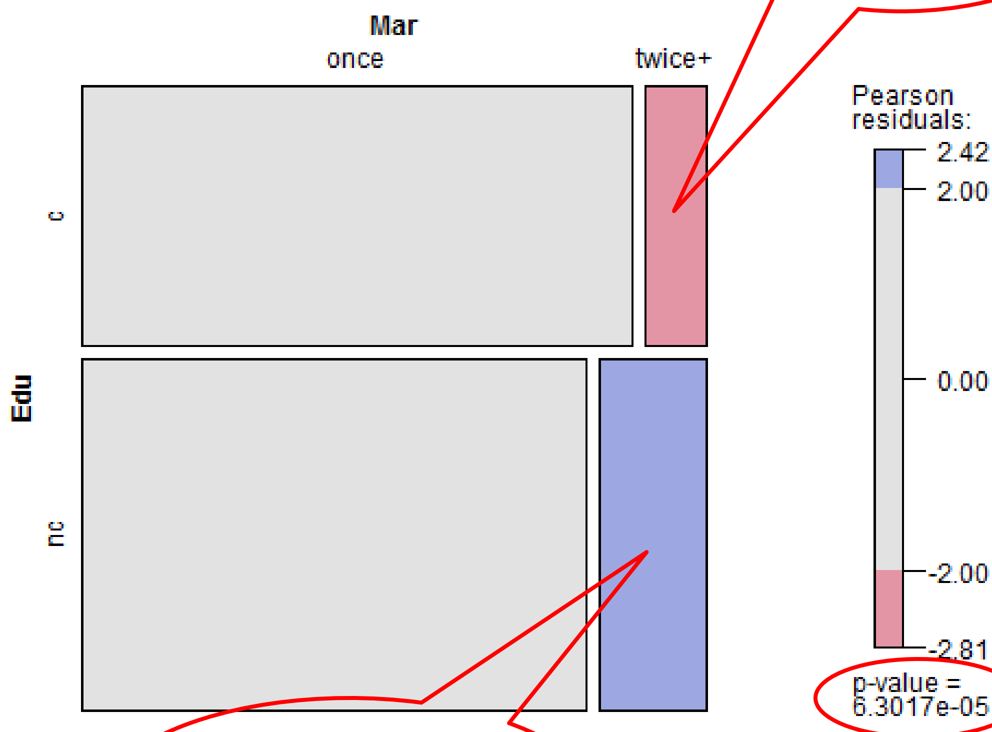
- Which cities are special?
- Which cities are like New Orleans?
- Seattle and Miami are quite far apart; how do they compare?
- R: Function “stars” in package “stats”



Mosaic plot with shading

R: Function “mosaic” in package “vcd”

Suprisingly small observed cell count



Suprisingly large observed cell count

p-value of independence test: Highly significant

Outliers: Theory of Mahalanobis Distance

Assume data is multivariate normally distributed
(d dimensions)

Squared Mahalanobis distance of samples follows a Chi-Square distribution
with d degrees of freedom

Expected value: d

("By definition": Sum of d standard normal random variables has
Chi-Square distribution with d degrees of freedom.)

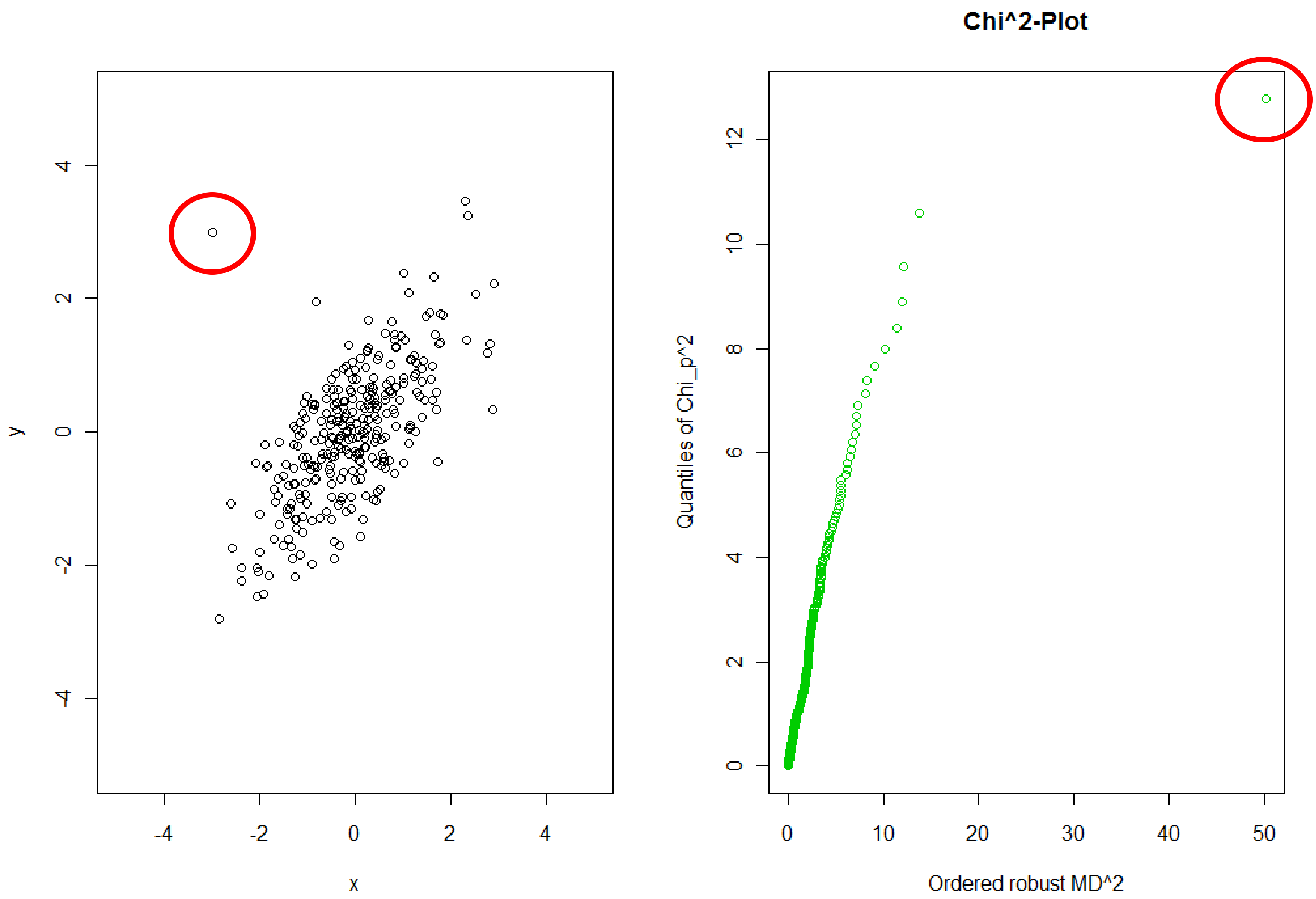
Outliers: Check for multivariate outlier

- Are there samples with estimated Mahalanobis distance that don’t fit at all to a Chi-Square distribution?
- Check with a QQ-Plot
- Technical details:
  - Chi-Square distribution is still reasonably good for estimated Mahalanobis distance
  - use robust estimates for μ, Σ

- R: Function «chisq.plot» in package «mvoutlier»

Outliers: chisq.plot

Outlier easily detected !



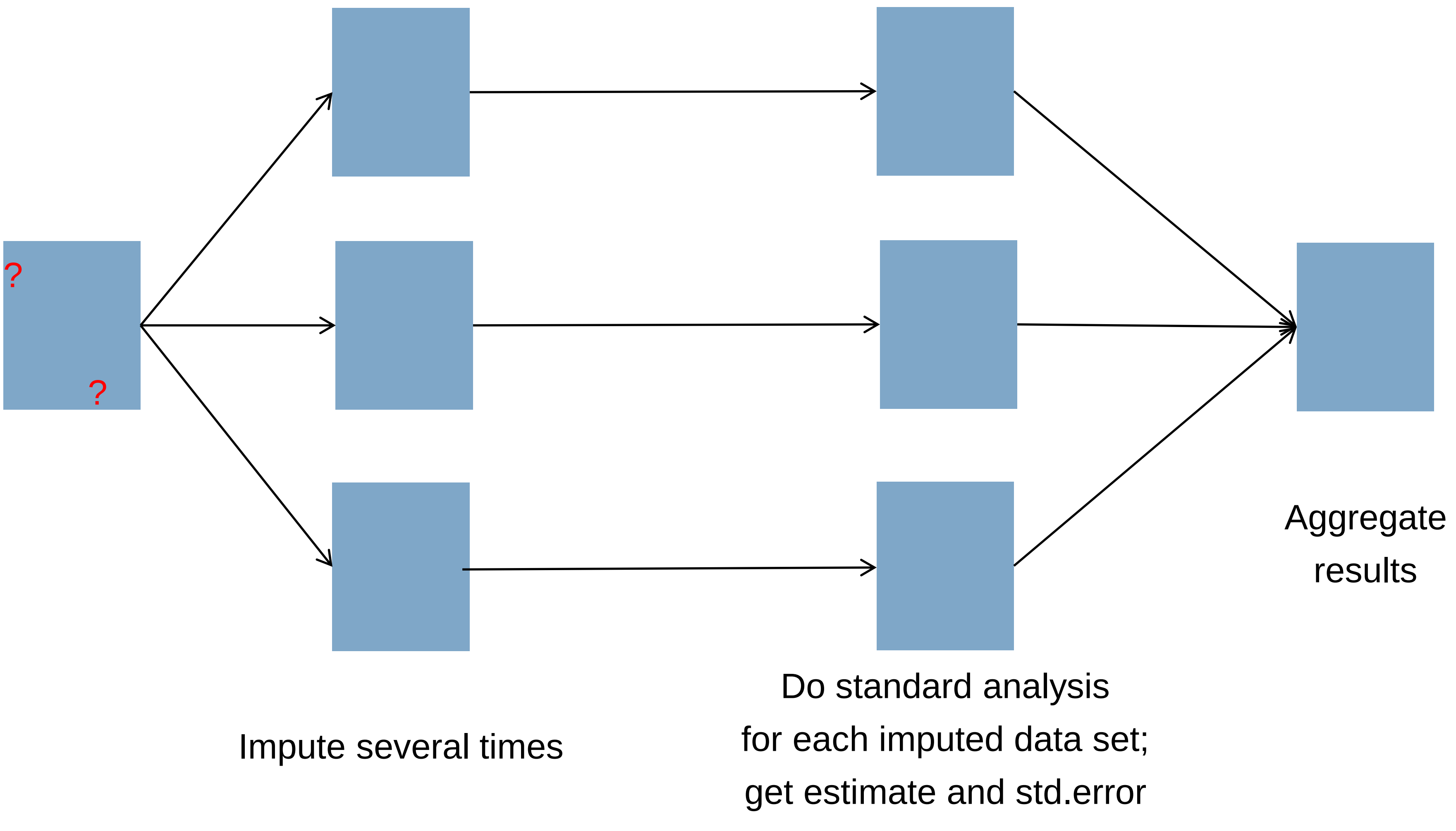
Missing values: Problem of Single Imputation

- Too optimistic: Imputation model (e.g. in Y = a + bX) is just estimated, but not the true model
- Thus, imputed values have some uncertainty
- Single Imputation ignores this uncertainty
- Coverage probability of confidence intervals is wrong

- Solution: Multiple Imputation
  Incorporates both
  - residual error
  - model uncertainty (excluding model mis-specification)

- R: Package «mice» for Multiple Imputation using chained equations

Multiple Imputation: MICE



Idea of MDS

- Represent high-dimensional point cloud in few (usually 2) dimensions keeping distances between points similar
- Classical/Metric MDS: Use a clever projection
  - guaranteed to find optimal solution only for euclidean distance
  - fast
  R: Function "cmdscale" in base distribution
- Non-metric MDS:
  - Squeeze data on table = minimize STRESS
  - only conserve ranks = allow monotonic transformations before reducing dimensions
  - slow(er)
  R: Function "isoMDS" in package "MASS"

Distance: To scale or not to scale…

- If variables are not scaled
  - variable with largest range has most weight
  - distance depends on scale
- Scaling gives every variable equal weight
- Similar alternative is re-weighing:

$$d(i,j) = \sqrt{w_1(x_{i1} - x_{j1})^2 + w_2(x_{i2} - x_{j2})^2 + ... + w_p(x_{ip} - x_{jp})^2}$$

- Scale if,
  - variables measure different units (kg, meter, sec,…)
  - you explicitly want to have equal weight for each variable
- Don't scale if units are the same for all variables
- Most often: Better to scale.

Dissimilarity for mixed data: Gower’s Dissim.

- Idea: Use distance measure between 0 and 1 for each variable: $d_{ij}^{(f)}$
- Aggregate: $d(i,j) = \frac{1}{p}\sum_{i=1}^{p} d_{ij}^{(f)}$

- Binary (a/s), nominal: Use methods discussed before
  - asymmetric: one group is much larger than the other
- Interval-scaled: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{R_f}$
  x_{if}: Value for object i in variable f
  R_f: Range of variable f for all objects
- Ordinal: Use normalized ranks; then like interval-scaled based on range

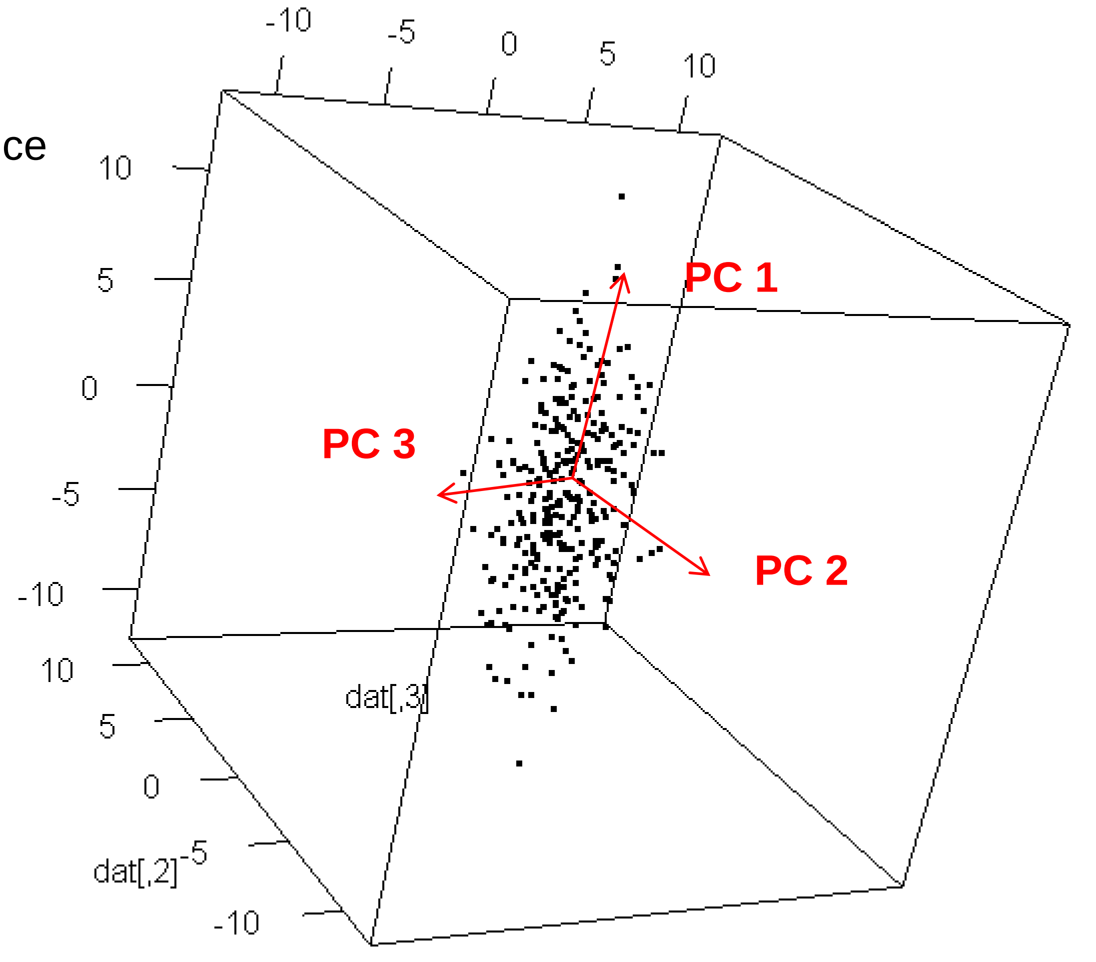
- R: Function “daisy” in package “cluster”

PCA: Goals

- Goal 1: Dimension reduction to a few dimensions while explaining most of the variance
(use first few PC's)
- Goal 2: Find one-dimensional index that separates objects best
(use first PC)

PCA (Version 1): Orthogonal directions

- PC 1 is direction of largest variance
- PC 2 is
  - perpendicular to PC 1
  - again largest variance
- PC 3 is
  - perpendicular to PC 1, PC 2
  - again largest variance
- etc.

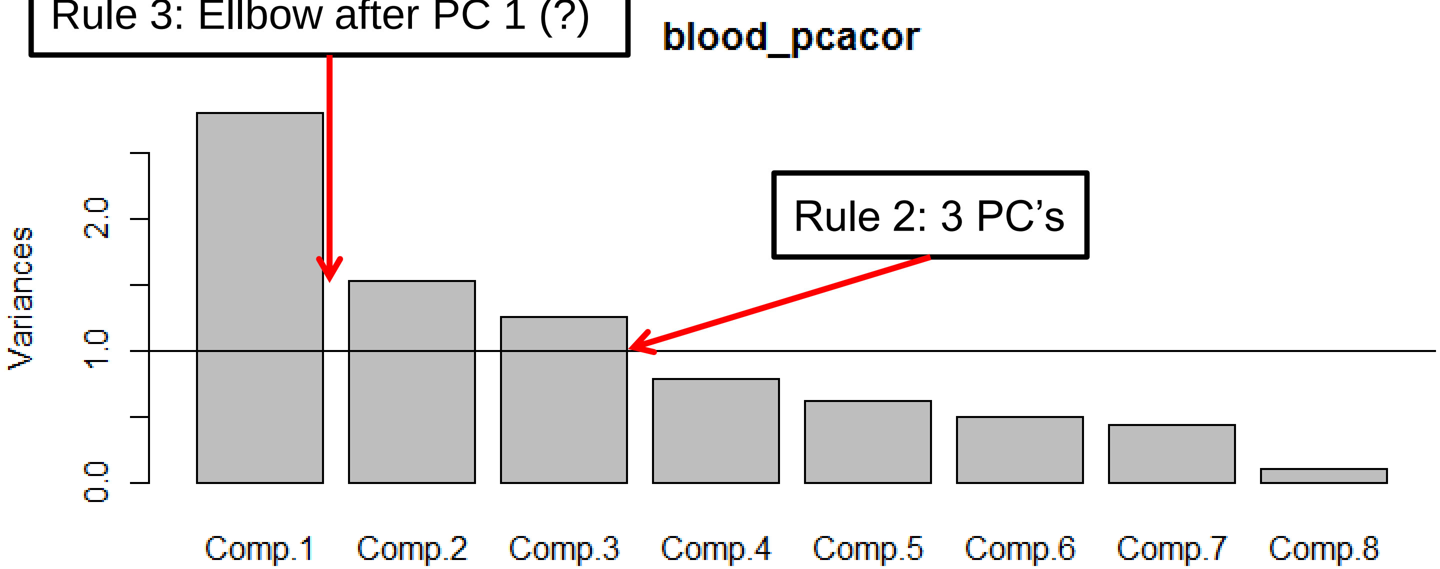


How many PC's: Blood Example

Rule 1: 5 PC's

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6
Standard deviation     1.6710100 1.2375848 1.1177138 0.88227419 0.78839505 0.69917350
Proportion of Variance 0.3490343 0.1914520 0.1561605 0.09730097 0.07769584 0.06110545
Cumulative Proportion  0.3490343 0.5404863 0.6966468 0.79394778 0.87164363 0.93274908
                           Comp.7     Comp.8
Standard deviation     0.66002394 0.31996216
Proportion of Variance 0.05445395 0.01279697
Cumulative Proportion  0.98720303 1.00000000
```

Rule 3: Ellbow after PC 1 (?)

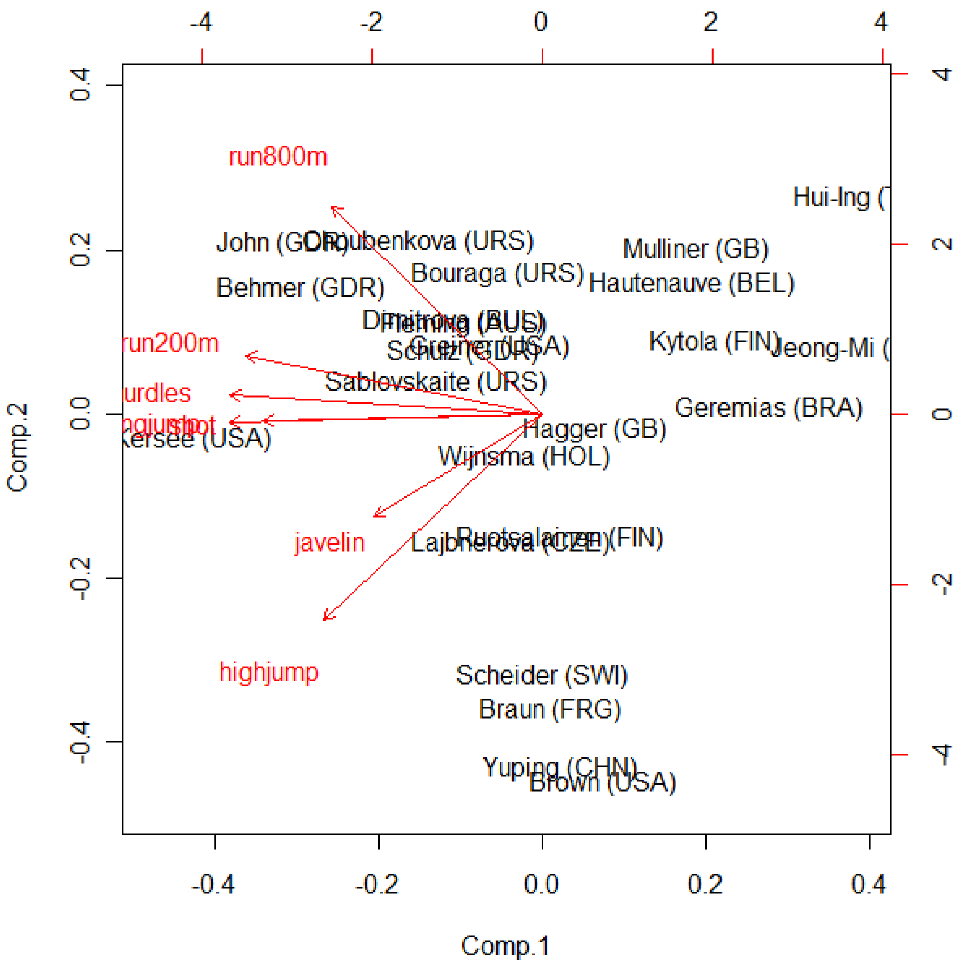


Biplot: Show info on samples AND variables

Approximately true:

- Data points: Projection on first two PCs
  Distance in Biplot ~ True Distance
- Projection of sample onto arrow gives original (scaled) value of that variable
- Arrowlength: Variance of variable
- Angle between Arrows: Correlation

Approximation is often crude;
good for quick overview



Supervised Learning: LDA

$$P(C|X) = \frac{P(C)P(X|C)}{P(X)} \sim P(C)P(X|C)$$

Find some estimate

Prior / prevalence:
Fraction of samples
in that class

Assume:
$X|C \sim N(\mu_c, \Sigma)$

Bayes rule:

Choose class where P(C|X) is maximal

(rule is “optimal” if all types of error are equally costly)

Special case: Two classes (0/1)

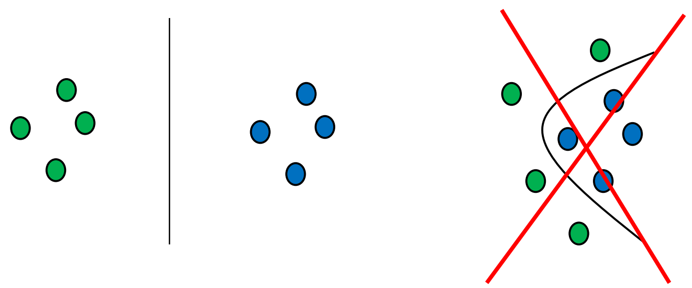
- choose c=1 if P(C=1|X) > 0.5 or
- choose c=1 if posterior odds P(C=1|X)/P(C=0|X) > 1

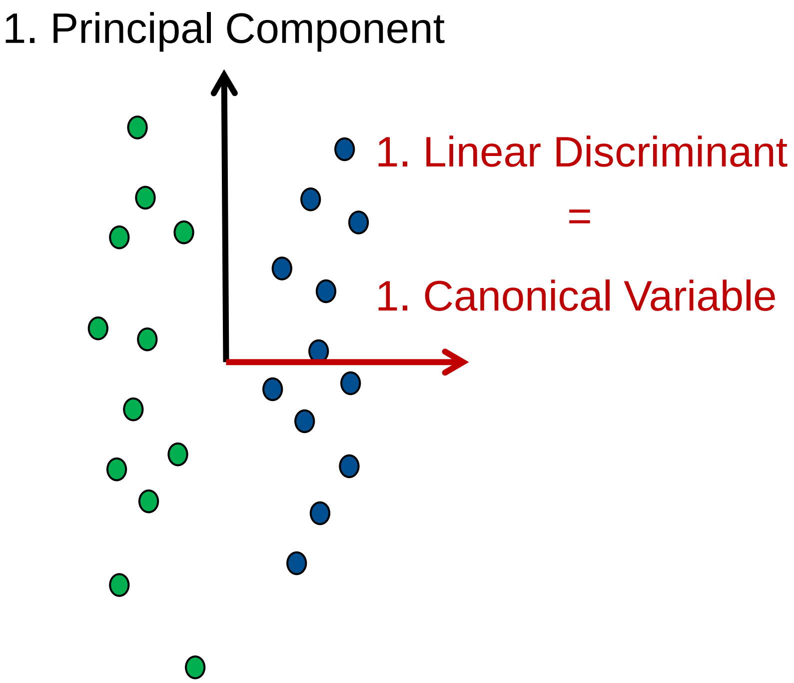
In Practice: Estimate $P(C), \mu_C, \Sigma$

LDA

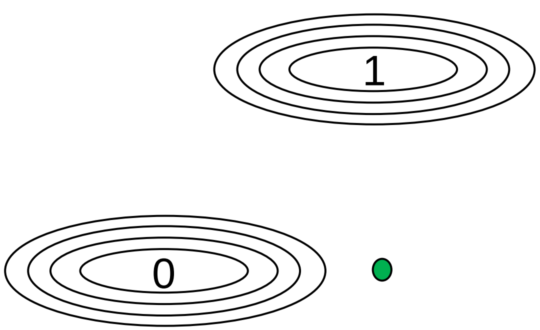
Orthogonal directions of best separation

Linear decision boundary





Balance prior and mahalanobis distance

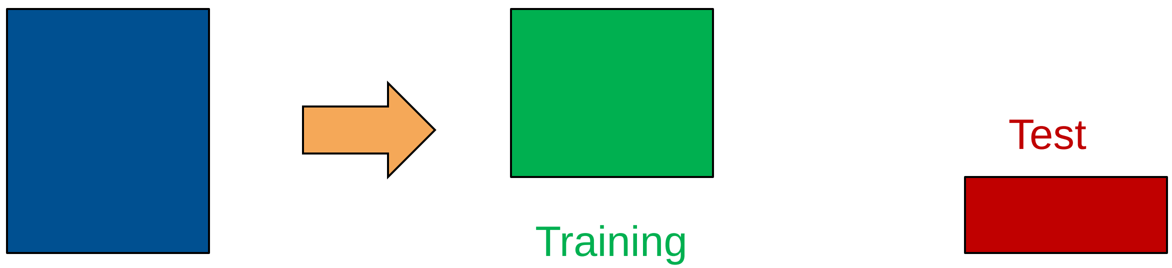


Classify to which class? – Consider:

- Prior
- Mahalanobis distance to class center

LDA: Quality of classification

- Use training data also as test data: Overfitting
  Too optimistic for error on new data
- Separate test data



- Cross validation (CV; e.g. “leave-one-out cross validation):
  Every row is the test case once, the rest in the training data