Eine Einführung in R: Hochdimensionale Daten: $n << p$ Teil II

Bernd Klaus, Verena Zuber

Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE),
Universität Leipzig

20. Januar 2011

Fragestellung: Supervised vs Unsupervised

Ähnlichkeitsmaße und Clustern

- Ähnlichkeitsmaße
- Clustern
- Visualisierung
- Gennetzwerke
  - Einführung
  - Networkmodelle
  - Graphical Gaussian Models

Grundsätzliche Fragestellung

- *Supervised*:
  Mit Daten X soll eine interessierende Variable Y erklärt werden.
  Beispiele:
  - Y kategorial: Differentielle Expression
  - Y metrisch: Lineares Modell
- *Unsupervised*:
  Welche Struktur findet sich in den Daten X?
  Eine interessierende Variable Y soll nicht (oder erst in der weiteren Analyse) untersucht werden.
  Beispiele:
  - Clusterverfahren
  - Netzwerke
  - Principal Component Analysis (PCA)

1. Ähnlichkeitsmaße: Wann sind Variablen ähnlich?

Ähnlichkeitsmaße

Wie kann Ähnlichkeit quantifiziert werden?

Dafür eignen sich Distanzmaße, wie:

- Euklidische Norm:

$$\|x\| = \sqrt{x_1^2 + x_2^2 + ... + x_p^2}$$

- Absolute (Manhatten) Norm:

$$\|x\|_1 = |x_1| + |x_2| + ... + |x_p|$$

- p-Norm:

$$\|x\|_p = \left(x_1^p + x_2^p + ... + x_p^p\right)^{1/p}$$

Clustern

Ein Cluster ist eine Gruppe von Variablen oder Beobachtungen. Es können sowohl Variablen als auch Beobachtungen geclustert werden. Entscheidend ist die Fragestellung!

Ziel des Clustern:

- Varianz in einem Cluster so gering wie möglich
- Varianz zwischen Clustern so stark wie möglich

Beispiel: Genexpressionsdaten

- Bei genetischen Prozessen ist häufig nicht ein Gen alleine beteiligt, sondern ein Netzwerk, das in komplexer Weise interagieren kann -> Gennetzwerke
- Finden sich die AML und ALL Beobachtungen aus den Golub Daten in identischen Clustern wieder oder sind diese bunt gemischt?

Cluster-Algorithmen: Single Linkage

- Start: Jedes der zu clusternden Objekte bildet ein eigenes Cluster
- Repeat: Verbinden der Cluster, die am ähnlichsten sind
- Stop: Ein großes Cluster
- Ergebnis: Baumstruktur, Dendrogramm

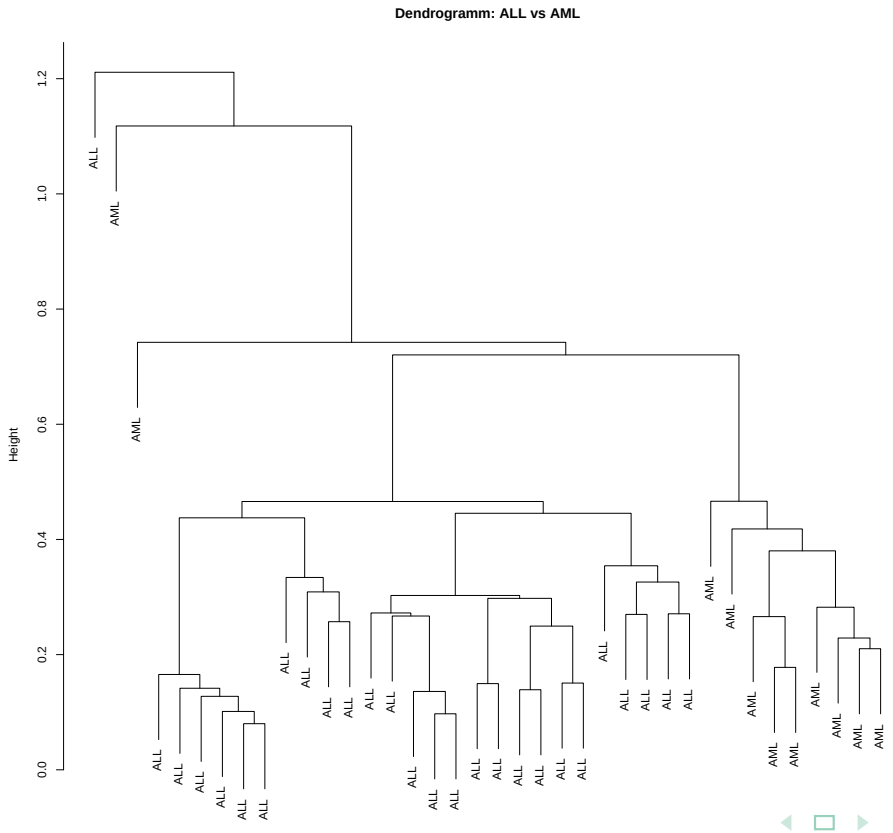
Vorteil: Anzahl der Gruppen muss nicht a priori festgelegt werden.

R-Befehl:

```
plot(hclust(dist(data,method="euclidian"),
method="single"), main="Dendrogramm: ALL vs AML",
                labels=gol.fac)
```

Cluster-Algorithmen: Single Linkage

- Expression von $p = 2$ Genen “CCND3 Cyclin D3” und “Zyxin”
- Frage: Können die $n = 38$ Beobachtungen getrennt nach Läukemie-Typen (ALL vs AML) geclustert werden?



Cluster-Algorithmen: k-means

2 k-means

- Start: Beginne mit k (zufälligen) Clustern und berechne die k-Mittelwerte
- Repeat: Assoziere jede Variable neu mit dem Mittelwert, zu dem es am nächsten liegt. Daraus ergeben sich k neue Mittelwerte
- Stop: Wenn sich die k Mittelwerte nicht mehr stark verändern
- Ergebnis: k Cluster

Nachteil: Anzahl der Gruppen muss a priori festgelegt werden.

R-Befehl:

```
cl <- kmeans(data, centers=2, nstart = 10)
```

Cluster-Algorithmen: k-means

- Expression von $p = 2$ Genen "CCND3 Cyclin D3" und "Zyxin"
- Frage: Können die $n = 38$ Beobachtungen getrennt nach Läukemie-Typen (ALL vs AML) geclustert werden?

```
cluster.2 <- kmeans(data, centers=2, nstart = 10)
```

- `K-means clustering with 2 clusters of sizes 11, 27`
- `Cluster means:`

	CCND3 Cyclin D3	Zyxin
1	0.6355909	1.5866682
2	1.8938826	-0.2947926

- `Clustering vector: 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1`
- `Within cluster sum of squares by cluster: 4.733248 19.842225 (between SS / total SS = 62.0 %)`

Eine Heatmap stellt die Expression aller p Gene aller n Beobachtungen dar. Die Spalten (und/oder Zeilen) werden nach Ihrerer Ähnlichkeit angeordnet. Zusätzlich wird an den Spalten (und/oder Zeilen) ein Dendrogramm geplotet.

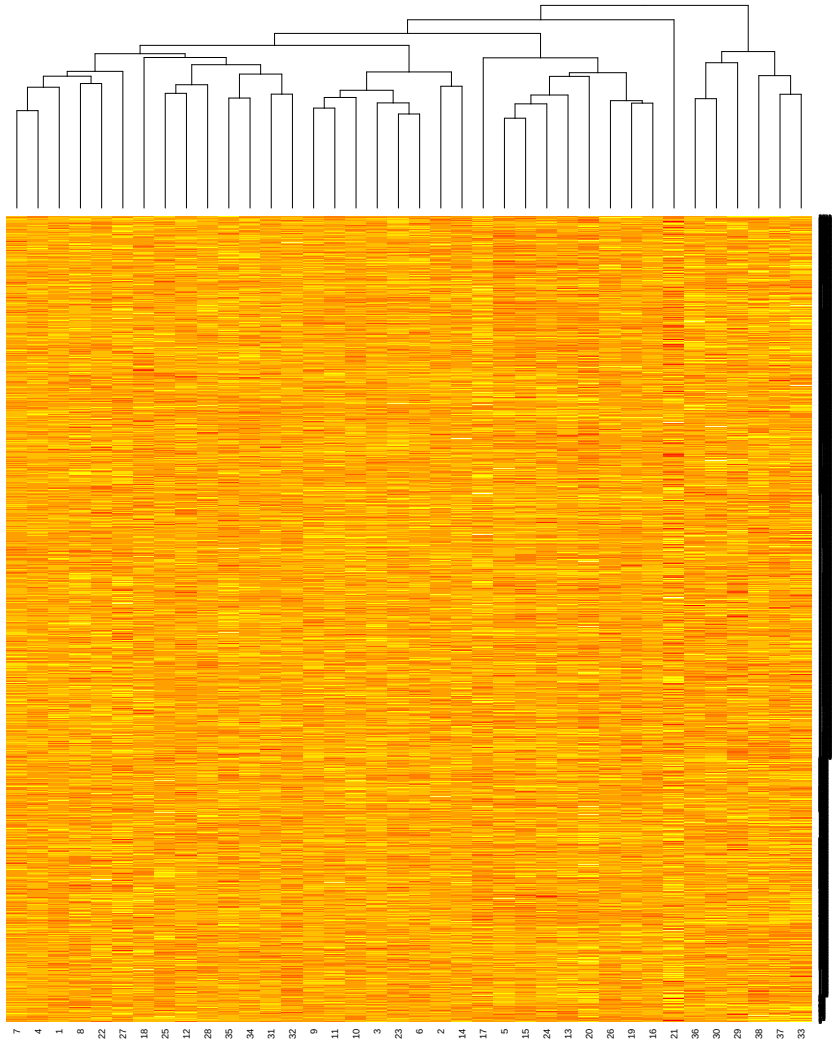
R-Befehl: `heatmap( ))` nur bedingt empfehlenswert

- `heatmap(x, Rowv=NULL, Colv=NULL,))`
- `x`: Datenmatrix
- `Rowv=NA`: Option falls Ordnen der Zeilen nicht erwünscht
- `Colv=NA`: Option falls Ordnen der Spalten nicht erwünscht

Weitere Pakete:

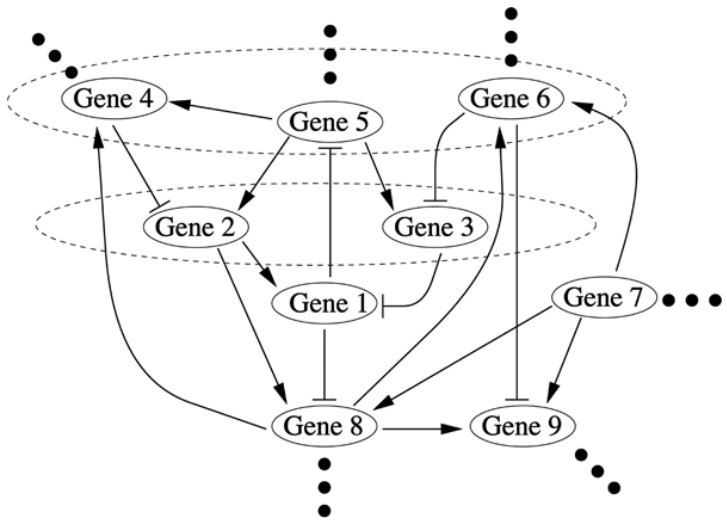
- heatmap.plus mit Befehl `heatmap.plus( )`
- gplot mit Befehl `heatmap.2( )`
- compHclust mit Befehl `compHclust.heatmap( )`

Heatmap der Golub-Daten



"Reverse Engineering" von Gennetzwerken

Oft will man die regulatorischen Zusammenhänge zwischen verschieden verstehen ...



$\Rightarrow$ Wie schätzt man diese kausalen Netzwerke ?

Durch viele verschiedene statistische Methoden ...

Es gibt viele verschiedene Verfahren z.B.

- Graphische Modelle (Bayesian networks, GGMs)
- Zeitreihenmodelle (VAR model, state space models etc.)
- Biochemische Modelle ((stochastiche) Differentialgleichungen etc.)

Hier: Schätzung von Graphical Gaussian Models (GGMs) mittels `GeneNet`

Correlation und Relevance Networks

- Einfache Grundidee: berechne die Korrelationsmatrix der Gene
- Bestimme Schwelle (z.B. 0.8) und verbinde alle Gene miteinander, die eine Korrelation über dieser Schwelle haben
- Nachteil: Korrelation ermöglicht keine Aussage über kausale Zusammenhänge!
- Bsp.: Im Sommer gibt es viele Herzinfarkte und es wird viel Eis gegessen, daraus kann man aber nicht folgern, dass Eisessen zum Herzinfarkt führt!

Graphical Gaussian Model

Eines der einfachsten graphischen Modelle:

- Ausgangspunkt
  - Korrelationsmatrix einer multivariaten Normalverteilung mit den Parametern μ und $\Sigma = \sigma_{ij}$, $i, j = 1, \ldots, p$
- GGMs basieren auf folgender Tatsache
  - Die bedingte Korrelation der Gene i and j, gegeben alle anderen Gene, ist σ_{ij}^{-1}
  - Die bedingte Korrelation heißt auch partielle Korrelation $\Rightarrow$ misst die direkte Interaktion!
- Ist diese "groß genug", ist im Graph eine Kante zwischen i und j, ansonsten nicht!
- Analog können Richtungen bestimmt werden (hier keine Details)

Schätzung von GGMs mittels "GeneNet"

- Wir untersuchen im Beispieldatensatz 800 Gene einer Pflanzengattung, die einen Einfluss auf den Tag / Nachtzyklus haben, die Daten `arth800` befinden sich im Paket `GeneNet`
- Zuerst berechnen wir die partielle Korrelation

```
data(arth800)
pcor <- ggm.estimate.pcor(arth800.expr)
```

Schätzung von GGMs mittels "GeneNet" II

- Dann bestimmen wir für jede Korrelation zwischen zwei Genen, die Wahrscheinlichkeit, dass diese auch wirklich von 0 verschieden ist $\Rightarrow$ multiples Testproblem!
- Diese Wahrscheinlichkeit ist 1- fdr = Prob(edge != 0)
- Analog wird für die Richtungen verfahren!

```
arth.edges <- ggm.test.edges(pcor, direct=TRUE,
                             fdr = TRUE )
```

- fdr: Entscheide, ob für jede Korrelation und für jede Richtung im Graph der fdr-Wert ausgerechnet werden soll
- direct : Soll ein gerichteter Graph ausgerechnet werden?

Schätzung von GGMs mittels "GeneNet" III

- Nun können wir das fertige Netzwerk durch Extraktion der signifkanten und gegebenenfalls gerichteten Kanten erstellen

```
arth.net <- extract.network(arth.edges,
method.ggm=c( "prob" , "qval" , "number"),
cutoff.ggm=0.8, method.dir=c("prob" ,"qval"
,"number" , "all") , cutoff.dir=0.8)
```

- `method.ggm / mehtod.dir` wie sollen signifikante /gerichtete Kanten ausgewählt werden? (lfdr, Fdr oder Anzahl)
- `cutoff.ggm/ cutoff.dir` Abschneidepunkt für signifikante Kanten / Richtungen

Schätzung von GGMs mittels "GeneNet" IV

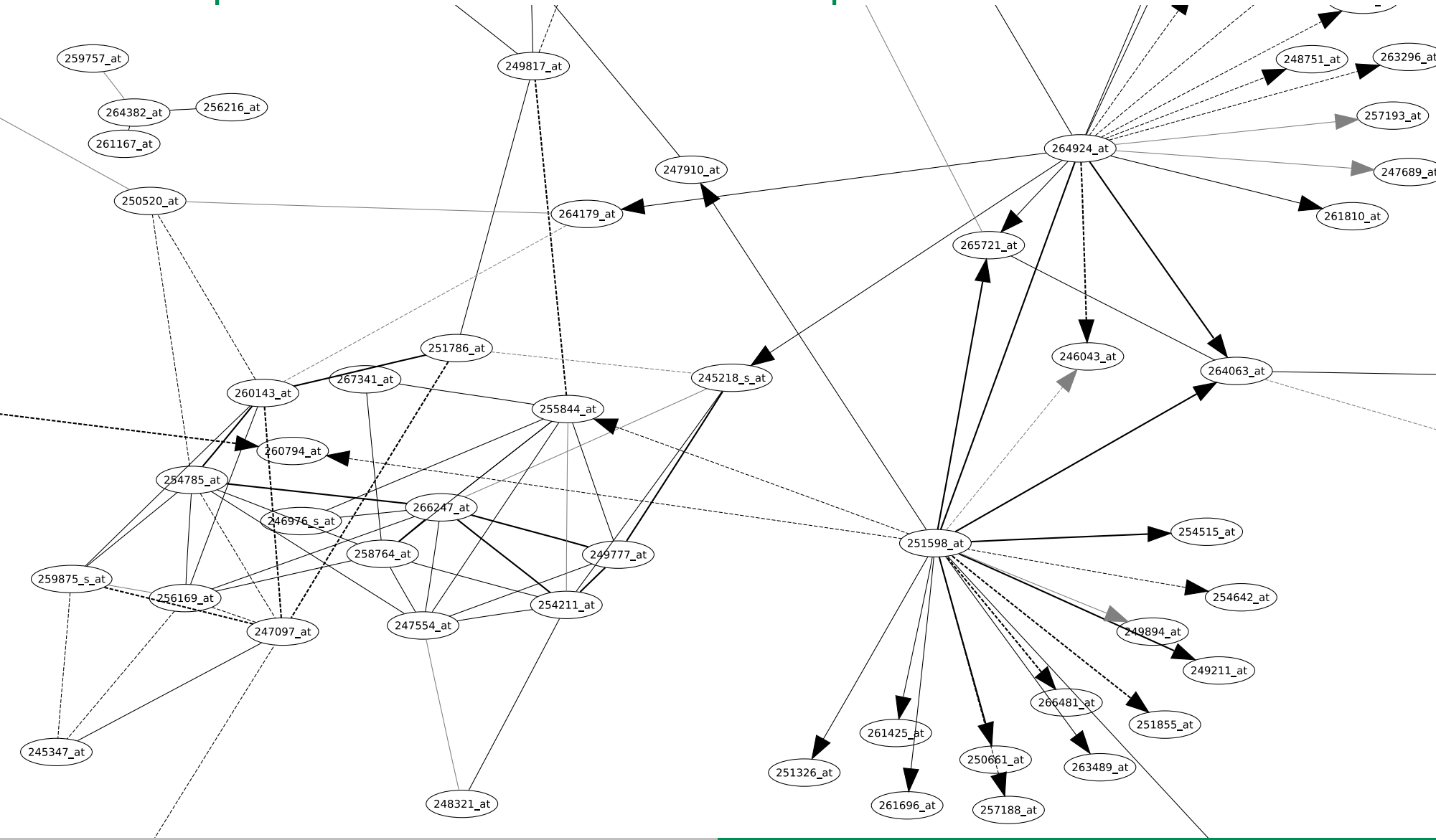
- Nun können wir das fertige Netzwerk plotten

```
ggm.make.dot(filename="arthdyn.dot", edge.list =
    arth.net, node.labels = node.labels,
      main="Arabdiopsis Network"))
```

- benutzt Graphviz, es wird ein .dot – File erstellt
- `edge.list` Ausgabe von ggm.test.edges
- `node.labels` labels für die Knoten
- `main` Titel des Graphen

Arabidopsis Daten: Gerichteter Graph

Arabidopsis Daten: Gerichteter Graph



Arabidopsis Daten: Gerichteter Graph