UNIVERSITÄT LEIPZIG

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

Bernd Klaus, Verena Zuber
Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE),
Universität Leipzig

12. Januar 2012

1. Hochdimensionale Daten ($n << p$)

2. Analyse von Genexpressionsdaten
   - Differentielle Expression
   - Penalisierte Regressionsverfahren

Hochdimensionale Daten: $n << p$
Eine Einführung in die Analyse von Genexpressionsdaten

Hochdimensionale Datentypen

Hochdimensionale Daten zeichnen sich meist dadurch aus, dass die Anzahl der Beobachtungen (n) wesentlich kleiner ist als die Anzahl der Variablen (p) ist.
Kurz schreibt man dafür auch $n << p$. Beispielfelder sind

- Bildgebende Verfahren: FMRI, DTI
- Finanzwesen
- Recommender-Systeme : "Kunden, die Produkt XY gekauft haben, werden auch Produkt YZ kaufen"
- Genomweite Assoziationsstudien (GWAS): "Welche Genvarianten treten bei welchen Krankheiten auf?"
- Im Fokus hier: Genexpressionsdaten

Kurze Einführung: Was sind Genexpressionsdaten?

- Genexpression bezeichnet die Ausprägung des Genotyps – also der genetischen Information (Gen, DNA)
- In DNA-Microarrays kann die Menge an mRNA (Messenger RNA) einer Vielzahl von Genen simultan bestimmt werden
- ... sie gilt als Maß für die Genexpression!
- statistische Methoden helfen dann z.B. **"differentielle Expression"** von Genen zu finden
- d.h. Unterschiede in der Expression von Genen unter verschiedenen Bedingungen (z.B. Krebs- und Normalgewebe)

Genexpressionsdaten

1. Struktur der Daten:
   - n Beobachtungen oder Messungen
   - p Gene (z.B. *Affy-Gene ST 1.0*-Chip: 28 132 Gene)
2. Dateneigenschaften:
   - Wichtig: Präprozessierug: Entfernung von experimentiellen Artefakten, z.B. Unterschiede zwischen verschiedenen Microarray Chips, Qualitätskontrolle uvm.
   - **Nach der Präprozessierung** ist die Genexpression eine **metrische Größe** (unser Ausgangspunkt)
   - Oft liegt eine **starke Korrelation** innerhalb bestimmter **Gengruppen** vor (*Pathways*, *Eigengenes*)
   - Durch die aufwendige und komplexe Datenerhebung sind Genexpressionsdaten anfällig gegenüber Messartefakten, Ausreißer und Ähnlichem

Genexpressionsdaten II

Typische Fragestellungen

- Differentielle Expression:
  Welche Gene unterscheiden sich zwischen bestimmten Gruppen?
- Zusammenhang zu stetigen Merkmal:
  Welche Gene stehen im Zusammenhang mit einer stetigen Zielgröße (z.B.: Alter, Blutdruck, BMI, Plasmakonzentration im Blut)?
- Gennetzwerke: Wie beeinflussen sich verschiedene Gene gegenseitig?

Eine Einführung in die Analyse von Genexpressionsdaten

Golub-Daten I

- Genexpressionsdaten mit $n = 38$ Beobachtungen (Chips) und $p = 3051$ Genen
- Weitere Information: Faktorvariable `golub.cl` beschreibt welche Form von Leukämie $(k = 2)$ bei der entsprechenden Beobachtungen vorliegt
  - ALL: acute lymphoblastic leukemia ($n_0 = 27$)
  - AML: acute myeloid leukemia ($n_1 = 11$)
- Die Daten sind schon präprozessiert
- Ziel der Studie: Einen Klassifikator für die beiden Leukämiesubtypen (ALL, AML) zu finden, der auf Genexpressionsdaten basiert

Golub-Daten II

REPORTS

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

T. R. Golub,[1,2*†] D. K. Slonim,[1†] P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5*]

Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). Here, a generic approach to cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias as a test case. A class discovery procedure automatically discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without previous knowledge of these classes. An automatically derived class predictor was able to determine the class of new leukemia cases. The results demonstrate the feasibility of cancer classification based solely on gene expression monitoring and suggest a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge.

1. Differentielle Expression

Differentielle Expression

- Gegeben: $k \leq 2$ Gruppen (Y kategorial)
- Frage: Welche Gene sind einer Gruppe stärker oder schwächer exprimiert als in einer anderen Gruppe?
- Wie kann man differentielle Expression quantifizieren?
  1. Fold Change: Mittelwertsdifferenz zwischen zwei Gruppen
  2. t-score: Mittelwertsdifferenz zwischen zwei Gruppen standardisiert mit der Standardabweichung
  3. Regularisierte t-score: t-Score mit veränderter Varianzschätzung, z.B. der SAM – score ("Significance analysis of microarrays applied to the ionizing radiation response" Tusher, Tibshirani und Chu, PNAS 2000)
  4. Wilcoxen-score: Nonparametrisches Analogon zum t-score
- Ziel: Eine Rangliste an Genen, die die höchste differentielle Expression aufweisen

Problematik: Multiples Testen

- Ähnlich wie bei ANOVA post-hoc tests, aber viel größere Anzahl an Tests
- Bsp.: Durchführen von $p = 2$ Tests zum Fehlerniveau 1. Art (H_0 ablehnen, obwohl H_0 wahr)

Multiples Testen, $\alpha = 0.05$

Wahrscheinlichkeit in beiden Tests **keinen** α-Fehler zu machen:

$$0.95 * 0.95 = 0.9025 \quad < \quad 0.95$$

$\Rightarrow$ Je mehr Tests durchgeführt werden, desto wahrscheinlicher ist es, eine Nullhypothese fälschlicherweise abzulehnen!

Bei der Analyse von Genexpression besonders relevant, da sehr viele Tests durchgeführt werden (p sehr groß!)

Problematik: Multiples Testen II

Einige Verfahren für multiples Testen

- Bonferoni Korrektur (sehr konservativ):
  Bei p Tests führe jeden Test zum Niveau α/p durch, um das Gesamtniveau von α zu halten — da p sehr groß unbrauchbar für Genexpressionsdaten
- Family Wise Error Rate (FWER):
  Kontrolle der Wahrscheinlichkeit mindestens eine wahre Nullhypothese fälschlicherweise abzulehnen
- False Discovery Rate (FDR):
  Kontrolle der falsch positiven Ergebnisse, d.h. bei z.B. FDR < 0.05 wird nur bei 5%, der als signifikant erkannten Tests, fälschlicherweise die Nullhypothese abgelehnt

Aktueller Überblicksartikel:

Benjamini: "Simultaneous and selective inference: Current successes and future challenges" Biometrical Journal, 2010

2. Penalisierte Regressionsverfahren

Penalisierte Regressionsverfahren

- Gegeben: Y metrisch, Bsp.: Alter, BMI, Blutdruck, Plasmakonzentration im Blut
- Frage: Welche Gene stehen im Zusammenhang mit Y?
- Wie kann man den Zusammenhang quantifizieren?
  1. Marginale Korrelation
  2. β-Koeffizient aus dem linearen Modell

Stichwort: Singularität

Vorsicht: Das einfache lineare Modell kann bei $n << p$ Datensätzen nicht angepasst werden (Stichwort: Invertieren von singulären Matrizen). Es muss regularisiert/penalisiert werden!

Die bekanntesten Verfahren dafür sind:

- Lasso: Tibshirani 1996
- Elastic Net: Zou und Hastie 2005