Ungerichtete Graphische Modelle mit Normalverteilung

Juliane Schäfer
Institut für Statistik, Ludwig-Maximilians-Universität München

23. Mai 2003

Inhalt

1. Datensituation und Fragestellung
2. Graphische Modelle mit Normalverteilung
3. Modellselektion
4. Modellvalidierung
5. Anwendung auf Genexpressionsdaten von Eisen et al. (1998) (Toh und Horimoto, 2002)
6. Diskussion

Datensituation und Fragestellung

- Microarray Technologie: Genexpressionsprofile
- Zugrunde liegender Mechanismus: Regulatorische Netzwerke, Genfunktionen

Daten:

$$y^{ij} = \text{el}\Big(\text{Gen } i\,(j)\Big) \qquad 1 \leq i \leq L,\ 1 \leq j \leq N$$

Empirischer Mittelwert von Gen i

$$\overline{y}^i = \frac{1}{N} \sum_{k=1}^{N} y^{ik}$$

Stichprobenkovarianz-Matrix S:

$$s^{ij} = \frac{1}{N} \sum_{k=1}^{N} \left(y^{ik} - \overline{y}^i\right) \left(y^{jk} - \overline{y}^j\right)$$

Korrelationskoeffizienten-Matrix C:

$$c^{ij} = s^{ij} / \left(s^{ii} s^{jj}\right)^{\frac{1}{2}} \qquad 1 \leq i, j \leq L$$

Graphische Modelle

Gene A und B stark korreliert

1. Direkte Wechselbeziehung
2. Indirekte Wechselbeziehung
3. Regulierung durch ein gemeinsames Gen

- Statistisches Modell zur Analyse der Zusammenhangsstruktur mehrerer Variablen

- Konzept der bedingten Unabhängigkeit

- Visualisierung durch Graphen $G = (V, E)$ (Markov Eigenschaften)

Graphische Modelle mit Normalverteilung

Annahme: Jeder der L-dimensionalen Vektoren ist Realisierung einer **multivariaten Normalverteilung** mit Parametern μ und Σ.

Bedingte Verteilung eines Gen-Paares i und j, gegeben die übrigen $L-2$ Gene, ist eine bivariate Normalverteilung.

Elemente der zugehörigen Kovarianzmatrix werden aus denen der inversen, ursprünglichen $(L \times L)$-Kovarianzmatrix berechnet.

Inverse Kovarianzmatrix $\Omega = \Sigma^{-1}$

$\omega^{ij} = 0$: Gene i und j bedingt unabhängig gegeben die übrigen $L-2$ Gene

GGM: Bedingte Unabhängigkeit von i und j $\leftrightarrow$
partieller Korrelationskoeffizient $\rho^{ij,\,\text{Rest}} = -\omega^{ij}/(\omega^{ii}\omega^{jj})^{\frac{1}{2}} = 0.$

Modellselektion

Wermuth & Scheidt Algorithmus (1977):

Schritt 0: Vollständiger Graph/ volles Modell $G(0)$,
Korrelationskoeffizientenmatrix $C(0)$

Schritt 1: PCCM $P(\tau)$ aus $C(\tau)$

Schritt 2: Ersetze das Element mit dem kleinsten absoluten Wert in $P(\tau)$ durch 0.

Schritt 3: Rekonstruktion von $C(\tau+1)$ aus $P(\tau)$

Schritt 4: Stopp-Kriterien:

$$\text{dev}\,1 \quad = \quad N \log\left(|C(\tau+1)| \,/\, |C(0)|\right) \overset{a}{\sim} \chi^2(\tau+1)$$

$$\text{dev}\,2 \quad = \quad N \log\left(|C(\tau+1)| \,/\, |C(\tau)|\right) \overset{a}{\sim} \chi^2(1)$$

Entscheidungsregel:

p-Wert $\leq \alpha$ (z. B. $\alpha = 0.05$): Verwirf $C(\tau+1)$. Ende der Iteration.

Sonst: Entferne die Kante, deren partieller Korrelationskoeffizient in $P(\tau)$ auf Null gesetzt wurde, aus $G(\tau)$ und erzeuge damit $G(\tau+1)$. Erhöhe τ um 1, und wiederhole die Schritte 1-4.

- Vorhandensein einer Kante z. N. α abgesichert
- Fehlen basiert nicht zwangsläufig auf starker Evidenz für bedingte Unabhängigkeit

Modellvalidierung

Biologische Betrachtungsweise: Übereinstimmungen mit experimentellen Studien, Kausalitätsbeziehungen

Statistische Betrachtungsweise: Bootstrap

1. N-maliges Ziehen mit Zurücklegen aus der ursprünglichen Stichprobe
2. Berechnung der partiellen Korrelationskoeffizientenmatrix P^{*b} für jede Bootstrap-Stichprobe
3. Wiederhole 1. und 2. B-mal unabhängig voneinander (z. B. $B = 100$)
4. Bootstrap-Wahrscheinlichkeit für $\{i, j\}$

$$\frac{1}{B} \sum_{b=1}^{B} \mathcal{I}_{\{\rho_{ij}^{*b} \neq 0\}}$$

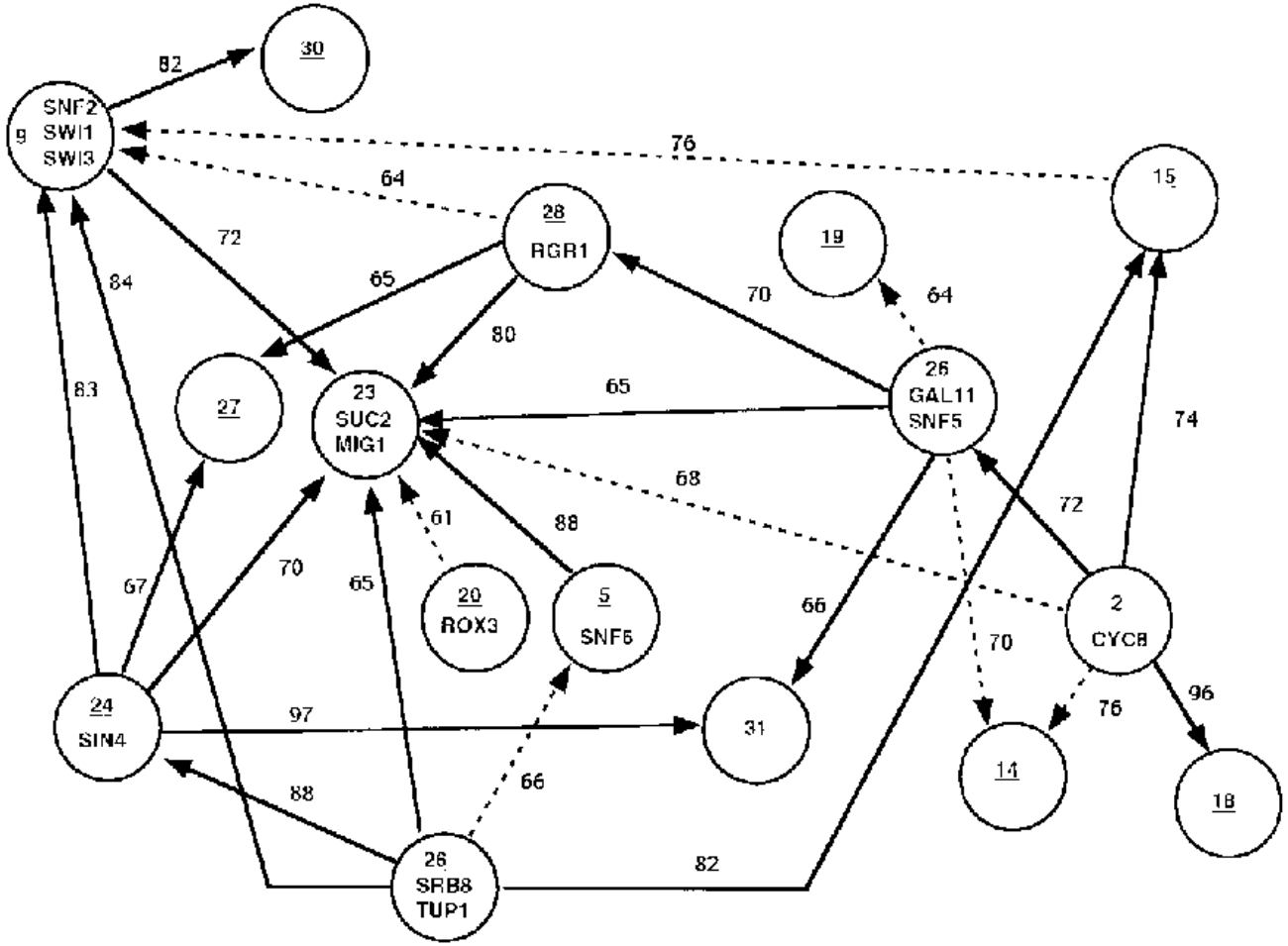
Analyse der Genexpressionsdaten von Eisen et al. (1998) durch Toh und Horimoto (2002)

Clusteranalyse

Starke Ähnlichkeit bezüglich des Expressionsmusters
$\rightarrow$ lineare Abhängigkeiten
$\rightarrow$ Probleme bei der Berechnung der Inversen von Σ

Mittleres Expressionslevel in Cluster k unter Bedingung j

$$\text{el}\Big(\text{Cluster } k\,(j)\Big) = \frac{1}{n}\left(\sum_{\text{Gen } i \in \text{ Cluster } k} \text{el}\Big(\text{Gen } i\,(j)\Big)\right) \quad 1 \leq k \leq M,\, 1 \leq j \leq N$$


30
82
SNF2
9 SWI1
SWI3
76
15
64
28
RGR1
19
72
84
65
80
70
64
26
GAL11
SNF5
83
27
23
SUC2
MIG1
65
74
68
72
61
88
67
70
65
20
ROX3
5
SNF6
66
2
CYC8
70
24
SIN4
97
31
76
96
14
18
88
66
26
SRB8
TUP1
82

Diskussion

- Transkriptionsebene
- Clusteranalyse, Repräsentation des Expressionsverhaltens eines Clusters
- Behandlung der Zeitreihendaten als unabhängige Realisierungen einer multivariaten Zufallsgröße
- Normalverteilungsannahme
- ungerichtete Zusammenhänge, Kausalitätsbeziehungen aus sachlogischen Überlegungen

Literatur

Toh, H. and Horimoto, K. (2002) Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics* **18**, 287-297.

Toh, H. and Horimoto, K. (2002) System for automatically inferring a genetic network from expression profiles. *Journal of Biological Physics* **28**, 449-464.

Wermuth, N. and Scheidt, E. (1977) Fitting a covariance selection to a matrix. Algorithm AS 105. *Appl. Stat.* **26,** 88-92.