

4/29/2005

Bayes-Netze II

Achim Tresch

dkfz.

GERMAN
CANCER RESEARCH CENTER
IN THE HELMHOLTZ ASSOCIATION

Übersicht

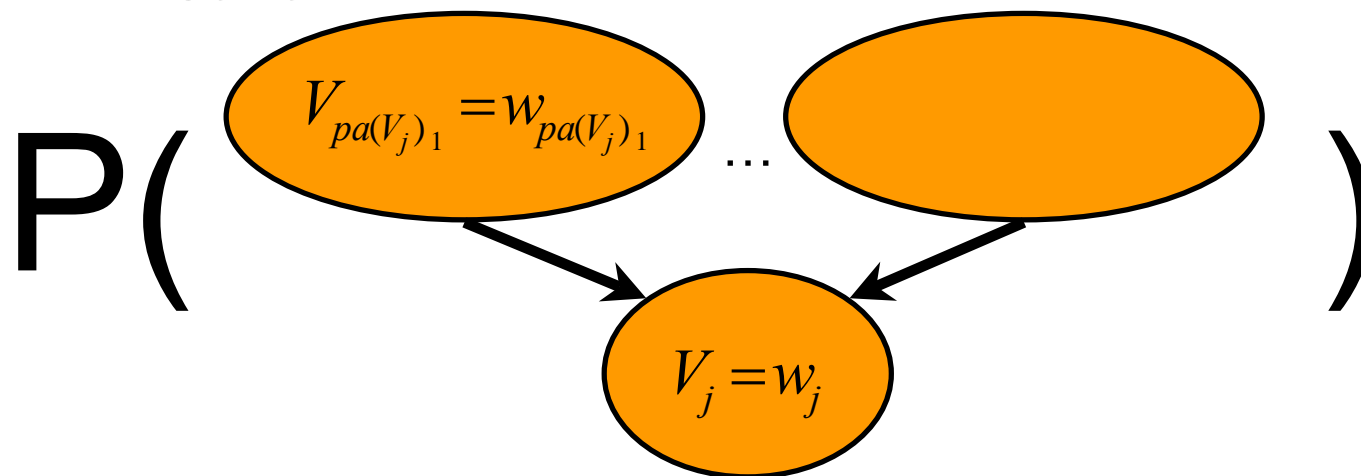
- Lernen von Bayes-Netzen 1 (Forts.)
Parameteridentifikation
- Lernen von Bayes-Netzen 2
„Structure discovery“,
„Netzwerkidentifikation“
- MCMC-Sampling
- Kausalität vs. Korrelation

Parameteridentifikation

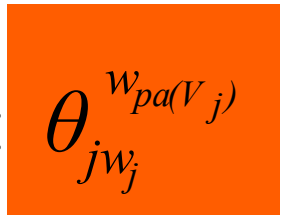
Wie lernt man nun eine lokale bedingte W.verteilung? Wir setzen der Einfachheit halber voraus, jede der Variablen nehme Werte von $1, \dots, r$ an.

Sei $w = (w_1, \dots, w_n)$ eine Beobachtung. Wir sind interessiert an den Wahrscheinlichkeiten $P(w_j = r \mid pa(V_j) = w_{pa(V_j)})$. Erinnerung: Die Notation $w_{pa(V_j)}$ bedeutet $w_{pa(V_j)} = (w_v, v \in pa(V_j))$.

Setze



$$= P(V_j = w_j \mid pa(V_j) = w_{pa(V_j)}) \equiv \theta_{jw_j}^{w_{pa(V_j)}}$$



Parameteridentifikation

Wie lernt man nun eine lokale bedingte W.verteilung? Wir setzen der Einfachheit halber voraus, jede der Variablen nehme Werte von $1, \dots, r$ an. Die Wahrscheinlichkeitstabelle von $P(V_j | pa(V_j))$ sieht dann folgendermaßen aus:

$$P(V_j | pa(V_j)) = \left\{ \begin{array}{c|c} pa(V_j) & P(V_j | pa(V_j)) \\ \hline v \in \{1, \dots, r\}^{q_j} & Multinom(1, \theta_{j1}^v, \dots, \theta_{jr}^v) \end{array} \right.$$

Für jeden der möglichen Elternzustände v muss eine Wahrscheinlichkeitsverteilung $Multinom(1, \theta_{j1}^v, \dots, \theta_{jr}^v)$ gelernt werden.

Sei q_j die Anzahl der Eltern von V_j . Bei r möglichen Zuständen der Zufallsvariablen gibt es also r^{q_j} verschiedene gemeinsame Elternzustände, und es müssen $(r-1) \cdot r^{q_j}$ Parameter gelernt werden (pro Verteilung sind wegen $\theta_{j1} + \dots + \theta_{jr} = 1$ nur $r-1$ Parameter zu lernen).

Parameteridentifikation

Für jeden Elternzustand geschieht dies beispielsweise mittels des in der vorigen Vorlesung präsentierten Verfahrens

$$P(\theta_{j1}^v, \dots, \theta_{jr}^v) = \text{Dir}(\alpha_{j1}^v, \dots, \alpha_{jr}^v)$$
$$P(\theta_{j1}^v, \dots, \theta_{jr}^v | D) = \text{Dir}(\alpha_{j1}^v + n_{j1}^v, \dots, \alpha_{jr}^v + n_{jr}^v)$$

Wobei n_{jt}^v die Zahl der Beobachtungen mit $V_j = t$, $V_{pa(V_j)} = v$ ist, und die α_{jt}^v das a-priori-Wissen repräsentieren.

Problem: Um eine multinomiale posterior-Verteilung vernünftig zu lernen, sollte die Zahl der Beobachtungen die Zahl der Parameter deutlich überschreiten. Daher sollte die Zahl der Beobachtungen $(r-1) \cdot r^{q_j}$ um ein Vielfaches übertreffen. Für $r = 3$, $q_j = 3$ wäre dies $2 \cdot 3^3 = 54$, d.h., man braucht weit über 100 Beobachtungen.

Das exponentielle Wachstum der benötigten Beobachtungszahl mit q_j führt dazu, dass man oft als zusätzliche Modellannahme alle q_j durch 2 oder 3 begrenzt (was z.B. bei der Schätzung genregulatorischer Netzwerke sehr fragwürdig ist).

Parameteridentifikation

Sei $\Theta_j^v = (\Theta_{j1}^v, \dots, \Theta_{jr}^v)$, für alle $j = 1, \dots, n$, $v \in \{1, \dots, r\}^{q_j}$

Und $\Theta_j = (\Theta_j^v, v \in \{1, \dots, r\}^{q_j})$

(dies sind alle Parameter, die die lokale bedingte W.verteilung des Knotens V_j festlegen)

Wie setzt man nun die posteriors der Parameter der lokalen Verteilungen $P(\Theta_j | D)$ zu einem posterior der gesamten Parameterverteilung zusammen?

Annahme: Globale Parameterunabhängigkeit

$$P(\Theta) = \prod_{j=1}^n P(\Theta_j)$$

Der globale posterior ist das Produkt der (lokalen) posteriors aller Knoten.

Parameteridentifikation

Annahme: Lokale Parameterunabhängigkeit

$$\begin{aligned} P(\Theta_j) &= P(\Theta_j^v, v \in \{1, \dots, r\}^{q_j}) \\ &= \prod \left\{ P(\Theta_j^v), v \in \{1, \dots, r\}^{q_j} \right\} \end{aligned}$$

Der lokale posterior ist das Produkt der posteriors aller gelernten Parametersets $\Theta_j^v = (\Theta_{j1}^v, \dots, \Theta_{jr}^v)$.

Diese Annahme erscheint gewagt, sie ist oft nicht erfüllt.

Beispiel: Gen X reguliert die Expression von Gen Y nach oben, und dieser Effekt ist X-dosisabhängig. Diskretisiert man die Expression von X bzw. Y in die Werte {niedrig, hoch}, so muss man $(2-1) \cdot 2^1 = 2$ Parameter, nämlich

$$P(Y=\text{hoch} \mid X=\text{niedrig}) \text{ und } P(Y=\text{hoch} \mid X=\text{hoch})$$

schätzen. Diese sind bei einem dosisabhängigen Effekt aber hochgradig abhängig, es gilt nämlich stets

$$P(Y=\text{hoch} \mid X=\text{niedrig}) \leq P(Y=\text{hoch} \mid X=\text{hoch})$$

Parameteridentifikation

Bei globaler und lokaler Parameterunabhängigkeit lässt sich der Posterior zerlegen in

$$P(\Theta | D) = \prod_{j=1}^n P(\Theta_j | D) = \prod_{j=1}^n \prod_{v \in \{1..r\}^{q_j}} P(\Theta_j^v | D)$$

(ohne Beweis). Insbesondere erfüllt auch der Posterior die globale und lokale Bedingung für Parameterunabhängigkeit. Im Fall des multinomialen Samplings erhält man

$$P(\Theta | D) = \prod_{j=1}^n \prod_{v \in \{1..r\}^{q_j}} Dir(\Theta_j^v | \alpha_{jt}^v + n_{jt}^v, t = 1, \dots, r)$$

Die einfache Lernregel lautet also:

$$\text{Prior} = P(\Theta) = \prod_{j=1}^n \prod_{v \in \{1..r\}^{q_j}} Dir(\Theta_j^v | \alpha_{jt}^v, t = 1, \dots, r)$$

$$\Rightarrow \text{Posterior} = P(\Theta | D) = \prod_{j=1}^n \prod_{v \in \{1..r\}^{q_j}} Dir(\Theta_j^v | \alpha_{jt}^v + n_{jt}^v, t = 1, \dots, r)$$

Parameteridentifikation

Aus dem globalen posterior können wir die gelernte Wahrscheinlichkeitsverteilung ausrechnen. Sei $w=(w_1, \dots, w_n)$ eine Beobachtung, sei

$$m_{jt}^v = \begin{cases} 1 & \text{falls } w_j = t, w_{pa(V_j)} = v \\ 0 & \text{sonst} \end{cases}$$

Dies ist eine Indikatorvariable, die überprüft, ob in der

Beobachtung w in $(\text{Knoten } j, \text{Eltern von } j) = (t, v)$ gilt. Die Likelihood einer Beobachtung w , gegeben die Parameter Θ , ist (beachte: w ist von den restlichen Beobachtungen D unabhängig):

$$\begin{aligned} P(w | D) &= \int_{\Theta} P(w | D, \theta) P(\theta | D) d\theta \\ &= \int_{\Theta} \prod_{j=1}^n \theta_{jw_j}^{w_{par(V_j)}} \prod_{j=1}^n \prod_{v \in \{1, \dots, r\}} Dir(\Theta_j^v | \alpha_{jt}^v + n_{jt}^v, t = 1, \dots, r) d\theta \end{aligned}$$

Parameteridentifikation

$$\begin{aligned}
 P(w | D) &= \int \prod_{\Theta} \prod_{j=1}^n \theta_{jw_j}^{w_{par}(V_j)} \prod_{j=1}^n \prod_{v \in \{1..r\}^{q_j}} \frac{\Gamma(\alpha_j^{(v_k)} + n_j^{(v_k)})}{\prod_{t=1}^r \Gamma(\alpha_{jt}^{(v_k)} + n_{jt}^{(v_k)})} \prod_{t=1}^r (\theta_{jt}^{(v_k)})^{\alpha_{jt}^{(v_k)} + n_{jt}^{(v_k)} - 1} d\theta \\
 &= \prod_{j=1}^n \prod_{v \in \{1..r\}^{q_j}} \frac{\Gamma(\alpha_j^v + n_j^v)}{\prod_{t=1}^r \Gamma(\alpha_{jt}^v + n_{jt}^v)} \underbrace{\int \prod_{\Theta} \prod_{t=1}^r (\theta_{jt}^v)^{\alpha_{jt}^v + n_{jt}^v - 1 + m_{jt}^v} d\theta}
 \end{aligned}$$

Posterior W.verteilung im einfachen, multinomialen Fall
(denke dir die die in diesem Ausdruck konstanten j, v weg)

$$= \prod_{j=1}^n \frac{\alpha_{jw_j}^{w_{par}(V_j)} + n_{jw_j}^{w_{par}(V_j)}}{\alpha_j^{w_{par}(V_j)} + n_j^{w_{par}(V_j)}}$$

Parameteridentifikation

Ist $D = (x_1, \dots, x_m)$ ein Satz von Beobachtungen, und $D_l = (x_1, \dots, x_{l-1})$, $l=1, \dots, m$, so ist (bei gegebenem Netzwerk)

$$\begin{aligned} P(D) &= P(x_m | D_m) P(D_m) = \dots = \prod_{l=1}^m P(x_l | D_l) \\ &\vdots \\ &= \prod_{i=1}^n \prod_{v \in \{1..r\}^{q_j}} \frac{\Gamma(\alpha_j^v)}{\Gamma(\alpha_j^v + n_j^v)} \prod_{t=1}^r \frac{\Gamma(\alpha_{jt}^v + n_{jt}^v)}{\Gamma(\alpha_{jt}^v)} \end{aligned}$$

Cooper, Herskovits
(1992)

Erinnerung: α_{jt}^v sind die Prior Parameter, und n_{jt}^v ist die Anzahl der Beobachtungen, in denen $(\text{Knoten } j, \text{Eltern}(j)) = (t, v)$ gilt.

**Diese Wahrscheinlichkeit $P(D)$ wird
BD (Bayessche Dirichlet)-Metrik genannt.**

Lernen von Bayes-Netzen 2

Sei $D=(x_1, \dots, x_n)$ eine Menge von Beobachtungen, die gemäß einer unbekanntem W.verteilung $P(X)$ gezogen wurden.

Gelernt werden soll ein Bayes-Netz $B=B(D)$, welches $P(X)$ möglichst gut beschreibt.

Es gibt zwei Stellschrauben, die an die Daten angepasst werden können:

- Die topologische Struktur des Netzes
- Die lokalen W.verteilungen

Wir befassen uns daher mit

- Netzwerkrekonstruktion
("structure discovery")
- Parameteridentifikation

Netzwerkidentifikation

Bisher wurde nur die halbe Wahrheit gelernt: Ein Bayesnetz B ist erst durch seine Parameter Θ und durch seine Topologie Γ eindeutig bestimmt, $B=B(\Theta, \Gamma)$. Bisher wurde beschrieben, wie die Parameter gelernt werden können. Im folgenden geht es um das Auffinden einer „wahrscheinlichen“ Topologie.

Die Formel von Cooper und Herskovits beschreibt die Wahrscheinlichkeit $P(D|\Gamma)$, die Daten D zu beobachten bei festgehaltener Netzstruktur Γ . Wir können uns, genau wie bei der Parameteridentifikation, auf den Bayesschen Standpunkt stellen:

$$P(\Gamma | D) = \frac{P(D | \Gamma)P(\Gamma)}{P(D)} = \frac{P(D | \Gamma)P(\Gamma)}{\sum_{\gamma \in \text{DAGs}(V)} P(D | \gamma)P(\gamma)}$$

Wiederum gilt es, einen prior, den Strukturprior $P(\Gamma)$ festzulegen. Die Summation im Nenner ist normalerweise (mehr als 6 Knoten) nicht durchführbar.

Netzwerkidentifikation

Daher benutzt man zur Suche nach plausiblen Topologien relative Scores:

$$\frac{P(\Gamma_1 | D)}{P(\Gamma_2 | D)} = \frac{P(D | \Gamma_1)P(\Gamma_1)}{P(D | \Gamma_2)P(\Gamma_2)}$$

Die aus den Daten D gelernte Wahrscheinlichkeitsverteilung ergibt sich ganz analog zum Fall der Parameteridentifikation als

$$\begin{aligned} P(w | D) &= \sum_{\Gamma \in \text{DAGs}} P(w, \Gamma | D) = \sum_{\Gamma \in \text{DAGs}} \overbrace{P(w | D, \Gamma)}^{\text{bek.}} \overbrace{P(\Gamma | D)}^{\text{bek.}} \\ &= \sum_{\Gamma \in \text{DAGs}} P(\Gamma | D) \int_{\Theta_\Gamma} P(w | D, \Gamma, \theta) P(\theta | D, \Gamma) d\theta \end{aligned}$$

Wiederum ist die Summation unmöglich, weswegen man zwischen zwei Näherungen wählt:

Netzwerkidentifikation

- **Modellselektion**: Suche ein Modell mit maximalem (hohem) Posterior $P(\Gamma|D)$ und nimm an, Γ sei die wahre Netzstruktur
- **Modellmittelung**: Ziehe eine möglichst große Zahl an zufälligen Samples Γ aus der Verteilung $P(\Gamma|D)$ und nähere $P(w|D)$ an durch

$$P(w|D) = \sum_{\Gamma \in \text{DAGs}} P(w|D, \Gamma) P(\Gamma|D) \approx \sum_{\Gamma \in \text{Samples}} P(w|D, \Gamma)$$

Rejection Sampling

Unsere Aufgabe:

$$P(w|D) = \sum_{\Gamma \in \text{DAGs}} P(w|D, \Gamma) P(\Gamma|D) \approx \sum_{\Gamma \in \text{Samples}} P(w|D, \Gamma)$$

Allgemein (diskreter Fall):

$$P(w) = \sum_{\Gamma \in G} P(w|\Gamma) P(\Gamma) \approx \sum_{\Gamma \in \text{Samples}} P(w|\Gamma)$$

⇒ Frage: Wie verschafft man sich Stichproben aus einer Zufallsverteilung wie z.B. hier $P(\Gamma)$?

Rejection Sampling

Annahme: Es sei möglich, aus einer Verteilung $Q(x)$ zu ziehen, und es gebe ein c mit $cQ(x) > P(x)$ für alle x

Ziehen eines Zufallssamples aus $P(x)$ erfolgt durch:

Erzeuge ein Zufallselement x aus der Verteilung Q .

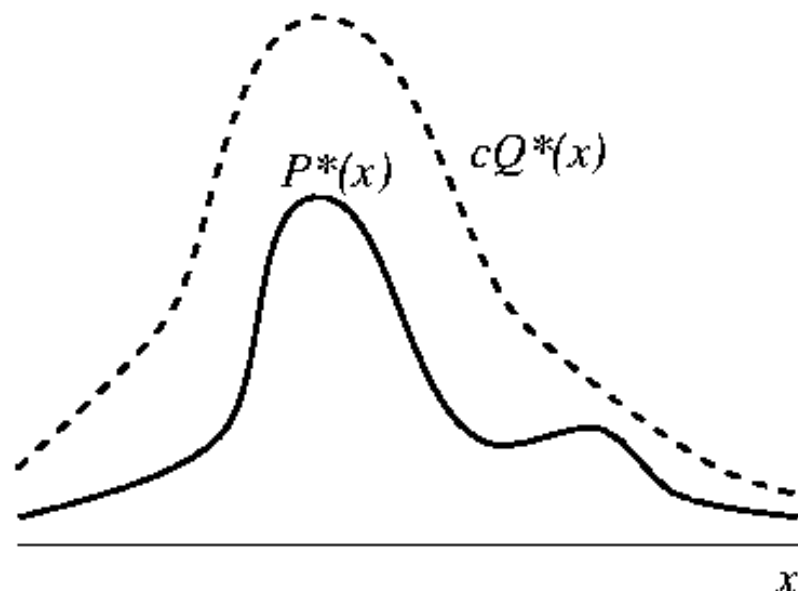
- Ziehe eine Zufallszahl u aus $U[0,1]$
- Akzeptiere x , falls $cQ(x) \cdot u \leq P(x)$, ansonsten wiederhole die ersten beiden Schritte

Die so realisierte Zufallsvariable ist verteilt wie $P(x)$.

Beweis: Sei $Y \sim Q(x)$

$$\begin{aligned} P(\text{"akzeptiere } x\text{"}) &= P(cQ(x) \leq P(x), Y = x) \\ &= P(cQ(x) \leq P(x) \mid Y = x) \cdot P(Y = x) \\ &= (P(x) / cQ(x)) \cdot Q(x) = P(x) / c \end{aligned}$$

Rejection Sampling



Probleme:

- Wie kann man sich eine brauchbare Verteilung Q beschaffen?
- In hochdimensionalen Räumen tendiert die Konstante c dazu, sehr groß zu werden. Damit wird auch $ucQ(x)$ i.d.R. sehr groß, d.h., die Akzeptanzrate wird extrem gering, was das Verfahren dann unpraktikabel macht.

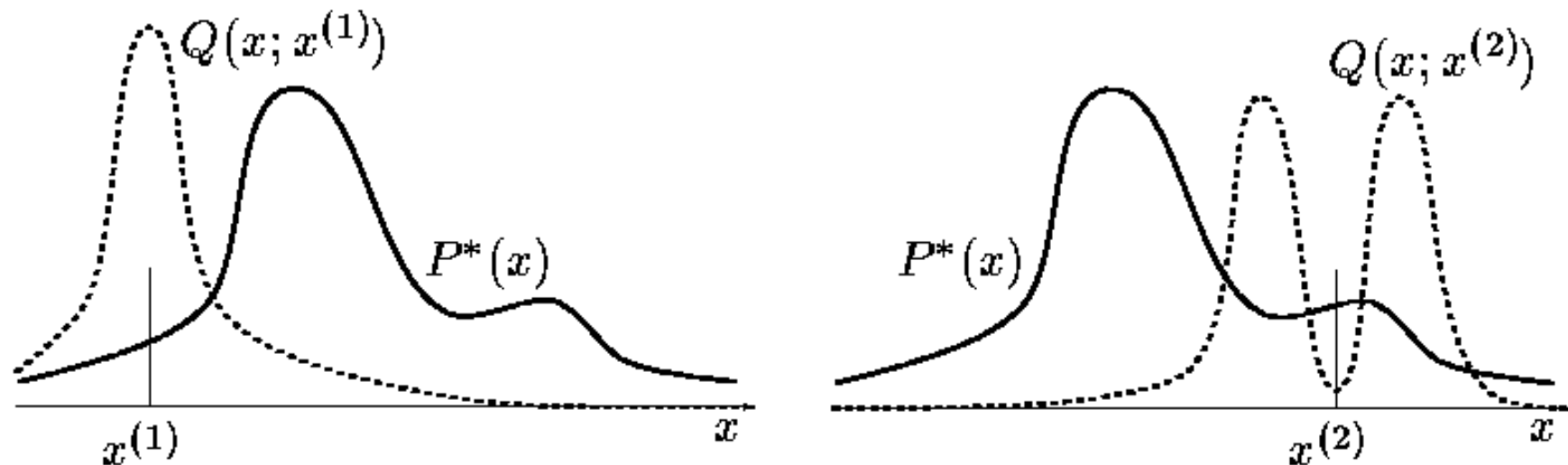
Markov Chain Monte Carlo

Der Metropolis Algorithmus benutzt eine Vorschlagsfunktion $Q(x'|x_t)$ welche vom aktuellen x_t abhängig ist:

- Starte mit einem beliebigen x_0
- Gegeben ein x_t , schlage ein x' aus der Verteilung $Q(x', x_t)$ vor
- Ziehe eine Zufallszahl u aus $U[0, 1]$
- Lehne ab, falls
$$u \leq \frac{P(x')Q(x_t | x')}{P(x_t)Q(x' | x_t)}$$
- Im Ablehnungsfalle setze $x_{t+1}=x_t$, ansonsten setze $x_{t+1}=x'$

Markov Chain Monte Carlo

- Der Metropolis Algorithmus produziert eine Folge von abhängigen Samples x_0, x_1, \dots
- Für $t \rightarrow \infty$ geht die empirische Verteilung der x_0, \dots, x_t gegen die Verteilung $P(x)$.



Netzwerke Sampeln mit Markov Chain Monte Carlo

Wir brauchen eine Vorschlagsfunktion $Q(\Gamma', \Gamma)$ für den Term

$$\frac{P(\Gamma' | D)Q(\Gamma_t | \Gamma', D)}{P(\Gamma_t | D)Q(\Gamma' | \Gamma_t, D)}$$

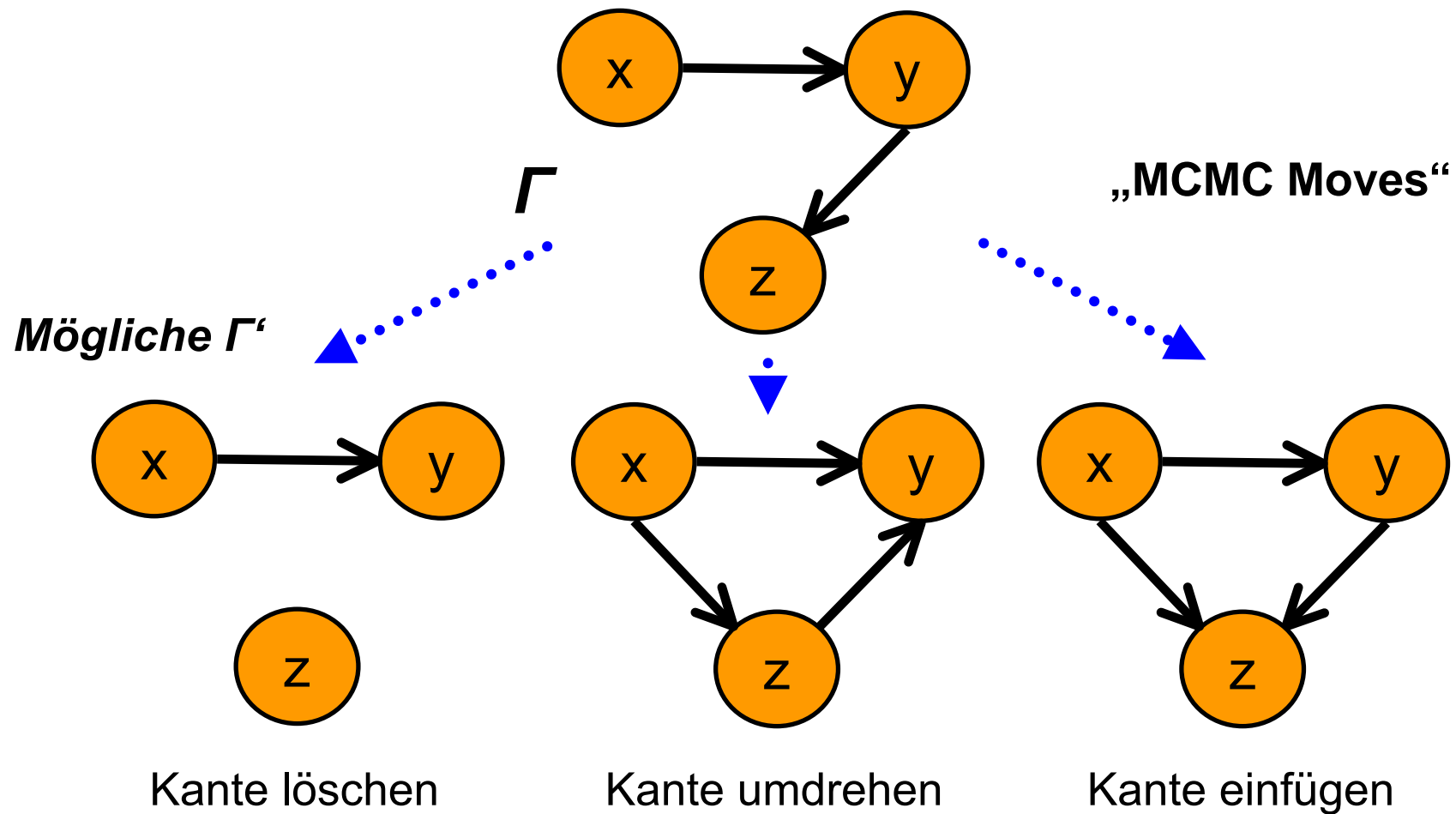
Wichtig ist hierbei, dass der gesamte Suchraum in wenigen Schritten $\Gamma \rightarrow \Gamma'$ durchschritten werden kann (dann ist die Gefahr geringer, „am Rande der Verteilung gefangen zu bleiben“, und die Korrelation der aufeinander folgenden Samples nimmt mit schneller ab).

Idee: Zu gegebenem Γ definiere eine Menge $M(\Gamma)$ von „benachbarten Graphen“, deren Kardinalität $|M(\Gamma)|$ relativ leicht bestimmt werden kann, und aus der uniform gezogen werden kann. Ziehe nun Γ' uniform aus $M(\Gamma)$ und setze

$$Q(\Gamma', \Gamma) = 1 / |M(\Gamma)|$$

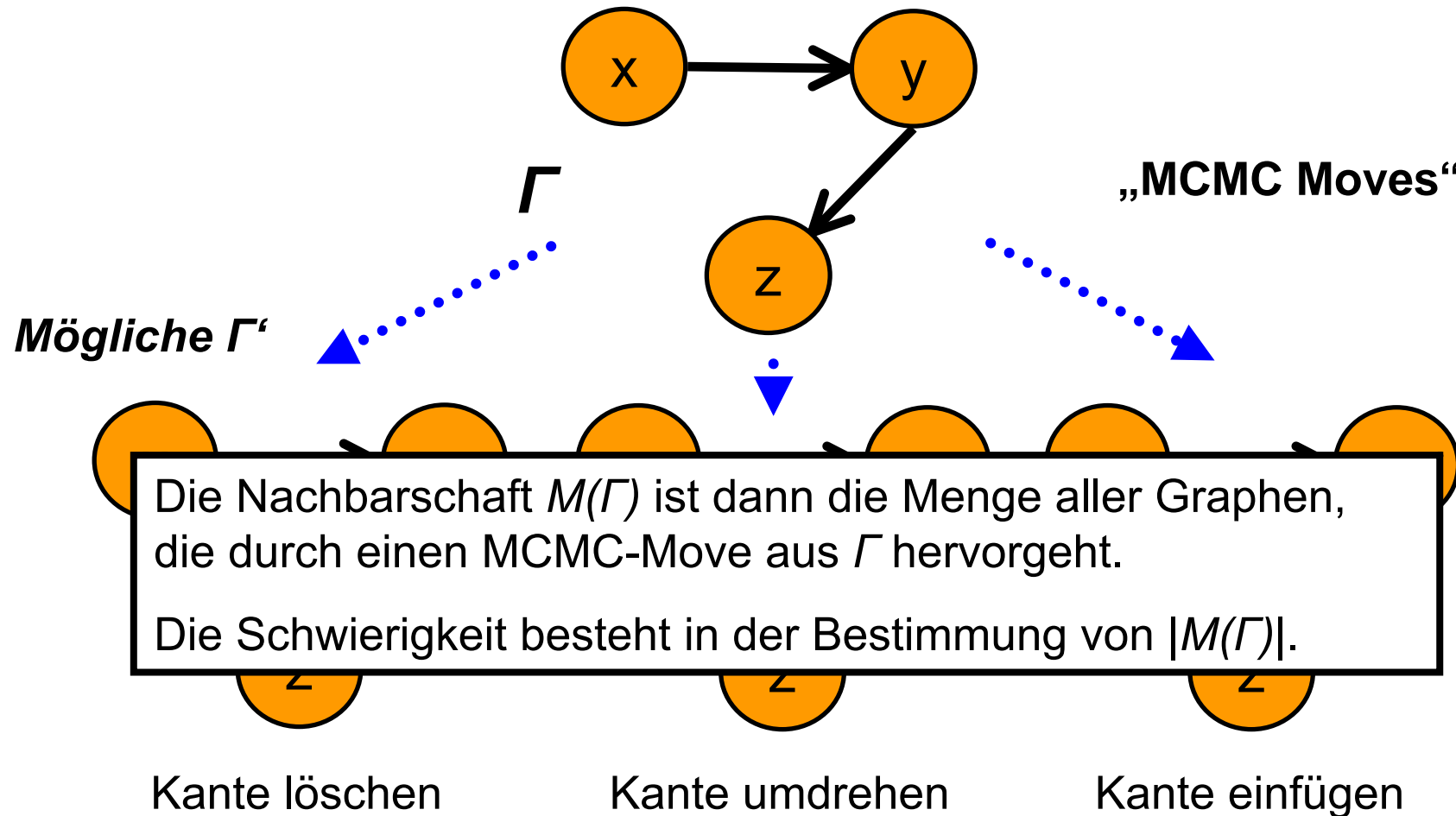
Netzwerke Sampeln mit Markov Chain Monte Carlo

Wir brauchen eine Vorschlagsfunktion $Q(\Gamma', \Gamma)$



Netzwerke Sampeln mit Markov Chain Monte Carlo

Wir brauchen eine Vorschlagsfunktion $Q(\Gamma', \Gamma)$



Kausalität vs. Korrelation

Ein Bayes-Netz legt eine gemeinsame Verteilung von Zufallsvariablen fest. Die Topologie des Netzes legt insbesondere fest, welche der Zufallsvariablen in sämtlichen Parametrisierungen (bedingt) unabhängig sind, und welche nicht.

Frage: Legt die Menge der Unabhängigkeiten auch umgekehrt die Topologie des Bayes-Netzes fest?

Antwort: Nein! Die Relation

$$B_1 \sim B_2$$

wenn B_1 die gleichen Unabhängigkeiten festlegt wie B_2

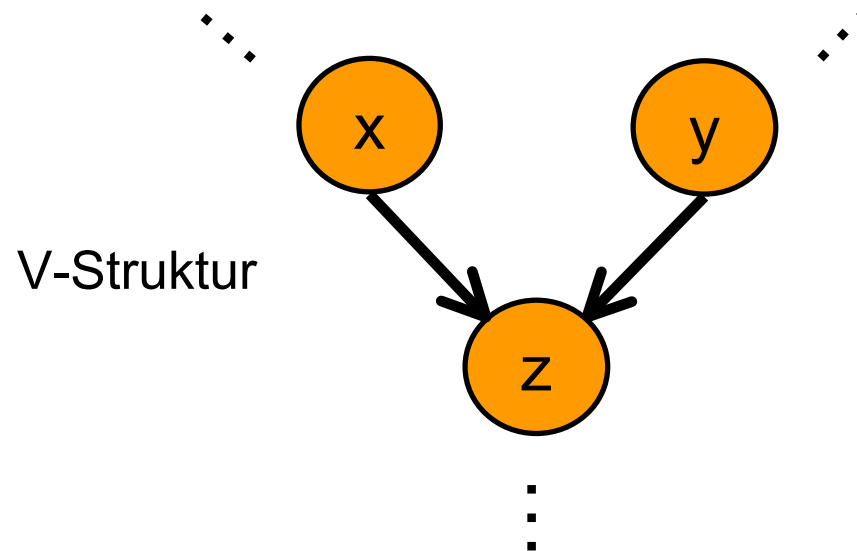
Ist eine Äquivalenzrelation. Bayes-Netze einer Äquivalenzklasse sind hinsichtlich der induzierten Unabhängigkeiten somit prinzipiell ununterscheidbar.

(Viele Netzidentifikationsalgorithmen nutzen Unabhängigkeitstests, um Kanten zu konstruieren oder zu löschen)

Kausalität vs. Korrelation

Verma und Pearl (1990) charakterisierten die Äquivalenzklassen wie folgt:

Satz: Zwei Graphen sind äquivalent, wenn ihr zu Grunde liegender ungerichteter Graph gleich ist, und sie die gleiche „V-Struktur“ haben.



Folgerung: Die in einem Bayes-Netz vorhandenen gerichteten Kanten sind **nicht notwendig kausale Beziehungen**.

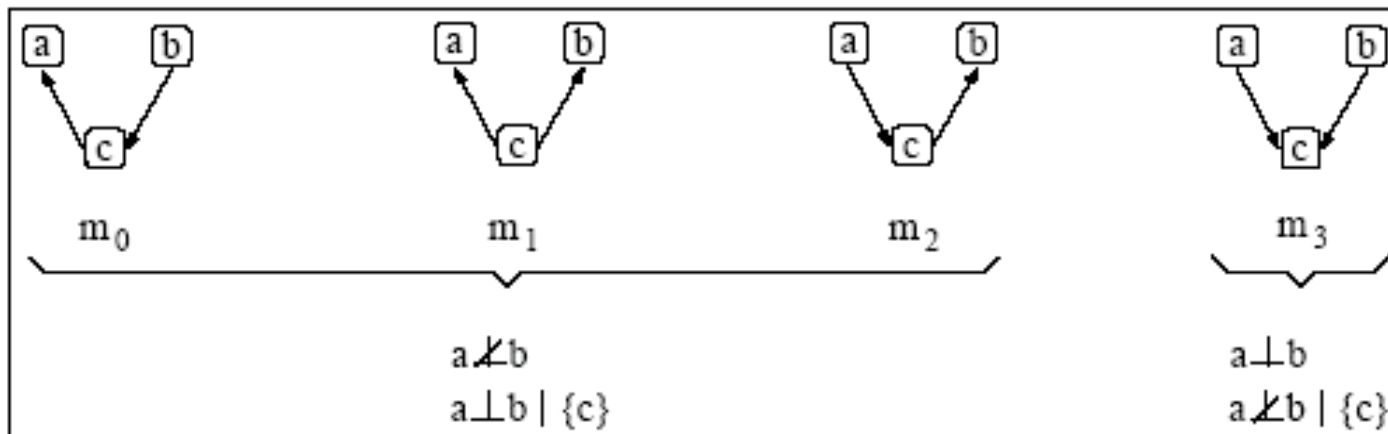
Kausalität vs. Korrelation

Beispiele für äquivalente und nicht äquivalente Graphen

$$P(a,b,c) = \underbrace{P(a|c)P(c|b)}_{m_0} P(b) = \underbrace{P(a|c)P(b|c)}_{m_1} P(c) = \underbrace{P(c|a)P(b|c)}_{m_2} p(a)$$

$$P(a,b,c) = \underbrace{P(c|a,b)}_{m_3} P(a)P(b)$$

Jede gemeinsame Verteilung $P(a,b,c)$, die mit Netz m_0 modelliert werden kann, kann auch mit m_1 bzw. m_2 modelliert werden und umgekehrt. Es gibt jedoch Verteilungen von (a,b,c) , die mit m_3 , nicht jedoch mit m_0 modelliert werden können.



Danksagung

David Heckerman, David J. MacKay

Literatur, Links

- **Constraint-Based Structural Learning in Bayesian Networks using Finite Data Sets**, Harald Steck, PhD thesis, TU München (2001)
- **Learning Bayesian Networks: The Combination of Knowledge and Statistical Data**, David Heckerman, <http://research.microsoft.com/~heckerman/>
- **Equivalence and synthesis of causal models**, Verma, T., Pearl, J., Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence, Boston (1990)
- **Introduction to Monte Carlo Methods**, David MacKay, bisnu.ac.kr/SEMINAR/GM/gm5.ppt

