3.6 Hidden Markov Modelle (HMM)

- Daten prinzipiell wie bei Markov-Ketten: $X_1, \ldots, X_T$ Realisationen eines stochastischen Prozesses $X = \{X_t, t \in \mathbb{N}\}$ mit diskretem Zustandsraum S.

- Die Annahme einer Markov-Kette wird ersetzt durch einen zweistufigen Mechanismus:
  - Den Daten unterliegt eine unbeobachtete (latente, hidden) Markov-Kette $C = \{C_t, t \in \mathbb{N}\}$ mit Zustandsraum $\{1, \ldots, m\}$.
  - Gegeben $C_t = i$ wird X_t generiert aus einer diskreten Verteilung mit

$$\pi_{xi} = P(X_t = x | C_t = i).$$

    Diese Verteilung kann parametrisch spezifiziert sein, z.B.

$$X_t | C_t = i \sim Po(\lambda_i)$$

    oder (für endliches S) nonparametrisch modelliert werden.
  - Zusätzliche Annahme: $X_t | C_t$ sind bedingt (gegeben C_t) unabhängig.

- Vorteil im Vergleich zu Markov-Ketten: Flexiblere Modellklasse, z.B. impliziert eine Markov-Kette eine (geometrische) Verteilung für die Rückkehrzeiten T_{ii}. In Anwendungen passt diese Verteilung häufig nicht zu den beobachteten Wartezeiten.

- $\{X_t\}$ hat im Allgemeinen marginal nicht die Markov-Eigenschaft!

  Beispiele:

(a) Regenfall-Daten:

$\{C_t\}$: MK mit 2 Zuständen; 0=gutes Wetter, 1= schlechtes Wetter

Übergangsmatrix: $\Gamma = \begin{pmatrix} \gamma_{00} & \gamma_{01} \\ \gamma_{10} & \gamma_{11} \end{pmatrix}$

$\{X_t\}$: Stochastischer Prozess mit 2 Zuständen; 0=kein Regen, 1=Regen

$$P(X_t = 1 \,|\, C_t = 1) = \pi_{11}$$
$$P(X_t = 1 \,|\, C_t = 0) = \pi_{10}$$
$$P(X_t = 0 \,|\, C_t = 1) = \pi_{01}$$
$$P(X_t = 0 \,|\, C_t = 1) = \pi_{00}$$

(b) Das gelegentlich unehrliche Casino:

2 Würfel: Würfel 1 fair; Würfel 2 unfair mit
$P(X_t = 6) = 0.5$, $P(X_t = 1) = \ldots = P(X_t = 5) = 0.1$
Casino wechselt zwischen Würfeln nach Übergangsmatrix

$$\Gamma = \begin{array}{c} \\ f \\ u \end{array} \begin{array}{c} \begin{array}{cc} f & u \end{array} \\ \begin{pmatrix} 0.95 & 0,05 \\ 0.1 & 0.9 \end{pmatrix} \end{array}$$

$\rightarrow$ Markovkette für Würfel

Insgesamt HMM mit $P(X_t = j \,|\, C_t = f) = \frac{1}{6} \quad \forall j$

$$P(X_t = j \,|\, C_t = u) = \begin{cases} \frac{1}{2}, & j = 6 \\ \frac{1}{10}, & \text{sonst} \end{cases}$$

(c) DNA-Sequenzen mit CpG-Inseln:

Bei Auftreten eines Paares CG (Notation: CpG) Mutation des C zu T mit hoher Wkt. $\Rightarrow$ CpG Paare seltener. Mutation in manchen Abschnitten (z.B. am Beginn von Genen) unterdrückt: 'CpG-Inseln'

HMM: $X_t \in \{A, C, G, T\}, \quad C_t \in \{0{=}\text{Insel}, 1{=}\text{keine Insel}\}$
Daten: $\{X_t, \ldots\}$ beobachteter Pfad
Gesucht: $\{C_t, \ldots\}$ latenter Pfad, enthält Inseln.

Inferenzproblem:

* Schätze ÜM Γ der latenten MK C
* Schätze bedingte Wkt. $P(X_t = j \,|\, C_t = k)$
* Schätze Pfad $\{C_t, t = 1, \ldots, T\}$ der verborgenen MK ("Viterbi-Algorithmus").

- Unbekannte Modellparameter:
  - Startverteilung $\delta = p(0)$ der latenten Markov-Kette.
  - Übergangsmatrix $\Gamma = (\gamma_{ij}) = P$ der latenten Markov-Kette.
  - $\Pi = (\pi_{si})$ Matrix der Wahrscheinlichkeitsverteilungen $P(X_t = s|C_t = i)$ für nonparametrische Hidden Markov Modelle bzw. $\theta = (\theta_i)$ Vektor der Modellparameter der parametrischen Verteilungen.

- Die beobachteten Größen X_t sind unabhängig für
  - $\pi_{si} \equiv \pi_s$,
  - $m = 1$,
  - $\gamma_{ij} \equiv \gamma_j$.

- Likelihood in Hidden Markov-Modellen:

$$\begin{aligned} L_T &= P(X_0 = x_0, \ldots, X_T = x_T) \\ &= \sum_{i_0=1}^{m} \ldots \sum_{i_T=1}^{m} P(X_0 = x_0, \ldots, X_T = x_T, C_0 = i_0, \ldots, C_T = i_T) \\ &= \sum_{i_0=1}^{m} \ldots \sum_{i_T=1}^{m} P(X_0 = x_0, \ldots, X_T = x_T \,|\, C_0 = i_0, \ldots, C_T = i_T) \\ &\qquad \cdot P(C_0 = i_0, \ldots, C_T = i_T) \end{aligned}$$

$$= \sum_{i_0=1}^{m} \cdots \sum_{i_T=1}^{m} (\underbrace{\pi_{x_0 i_0} \cdot \ldots \cdot \pi_{x_T i_T}}_{P(X_0=x_0|\ldots)\cdot\cdots\cdot P(X_T=x_T|\ldots)}) (\underbrace{\delta_{i_0} \gamma_{i_0 i_1} \cdot \ldots \cdot \gamma_{i_{T-1} i_T}}_{P(C_0=i_0,\ldots,C_T=i_T)}).$$

- Umschreiben ergibt einen numerisch günstigeren Ausdruck:

$$L_T = \delta \Lambda(x_0) \left(\prod_{t=1}^{T} B_t \right) \mathbb{1}',$$

mit

$$\begin{aligned} B_t &= \Gamma \Lambda(x_t), \\ \Lambda(x) &= \operatorname{diag}(\pi_{x1}, \ldots, \pi_{xm}), \\ \delta &= (\delta_1, \ldots, \delta_m). \end{aligned}$$

- Wenige Parameter (insbesondere parametrische Modelle) $\Rightarrow$ direkte Maximierung über numerische Verfahren.

- Für nonparametrische Modelle mit vielen Parametern: Baum-Welsh-Algorithmus (EM-Algorithmus).

- Vorhersagewahrscheinlichkeiten:

$$P(X_{T+1} = x_{t+1} | X_T = x_T, \ldots, X_0 = x_0) = \frac{L_{T+1}}{L_T}.$$

- Schätzung des (wahrscheinlichsten) Pfads der latenten Markov-Kette:

$$\max_{i_0,\ldots,i_T} P(C_0 = i_0, \ldots, C_T = i_T | X_0 = x_0, \ldots, X_T = x_T)$$

bzw. äquivalent dazu

$$\max_{i_0,\ldots,i_T} P(C_0 = i_0, \ldots, C_T = i_T, X_0 = x_0, \ldots, X_T = x_T).$$

Lösung: Viterbi-Algorithmus. Definiere:

$$\begin{aligned}
\zeta_{0i} &= P(C_0 = i \,|\, X_0 = x_0) \\
\zeta_{ti} &= \max_{i_0,\ldots,i_{t-1}} P(C_0 = i_0, \ldots, C_t = i \,|\, X_0 = x_0, \ldots, X_t = x_t) \\
&= \left[\max_k \zeta_{t-1,k}\gamma_{k,i}\right] \cdot \pi_{x_t i} \\
&= \text{Wkt. des wahrscheinlichsten Pfades zu } C_t = i
\end{aligned}$$

$\Rightarrow$ Die $T \times m$ Matrix der ζ_{ti} lässt sich in $O(Tm^2)$ Schritten aus den π_{si}, γ_{ij} berechnen.

Rückwärts auflösen ergibt

$$\begin{aligned}
\hat{i}_T &= \operatorname*{argmax}_{1 \le i \le m} \quad \zeta_{Ti} \\
&= i, \text{das am Ende des wahrscheinlichsten Pfades steht.}
\end{aligned}$$

$$\hat{i}_t = \underset{1 \le i \le m}{\operatorname{argmax}} \quad (\zeta_{ti} \gamma_{i,\hat{i}_{t+1}})$$

$= i$, das am Vorgängerpunkt des wahrscheinlichsten Pfades steht.

3.7 Allgemeine Markov-Ketten

- Definition: Markov-Kette (mit allgemeinem Zustandsraum S)

  Eine Markov-Kette ist ein stochastischer Prozess $\{X_t,\ t \in \mathbb{N}_0\}$ mit Zustandsraum $(S, \mathcal{S})$, der die Markov-Eigenschaft

$$P(X_{t+1} \in A | X_t = x, X_{t-1} \in A_{t-1}, \ldots, X_0 \in A_0) = P(X_{t+1} \in A | X_t = x)$$

  für beliebige $x \in S$ und $A, A_{t-1}, \ldots, A_0 \in \mathcal{S}$ erfüllt.

  Die MK heißt homogen, falls die Übergangswahrscheinlichkeiten nicht von t abhängen.

$$P(x, A) := P(X_{t+1} \in A | X_t = x) = P(X_1 \in A | X_0 = x)$$

  heißt dann Übergangskern der homogenen MK.

- Bemerkungen:

(a) Für Markov-Ketten mit diskretem Zustandsraum ist $P(x, A)$ durch die Übergangsmatrix

$$(p_{ij})_{i,j\in S} = [P(X_{t+1} = j|X_t = i)]_{i,j\in S}$$

bestimmt.

(b) Für $S = \mathbb{R}$ ist $P(x, A)$ durch eine Übergangsdichte $p(x, y)$ mit

$$P(x, A) = \int_A p(x, y)dy$$

bestimmt.

Beispiele

– Gauss-Irrfahrt:

$p(x, y)$ oder $f(y|x)$ Dichte einer $N(x, \sigma^2)$-Verteilung:

$$X_{t+1} = X_t + \epsilon_t, \quad \epsilon_t \overset{iid}{\sim} N(0, \sigma^2), \quad X_0 = 0 \Rightarrow X_{t+1}|X_t \sim N(x_t, \sigma^2)$$

$$\Rightarrow P(x, A) = P(X_{t+1} \in A \,|\, X_t = x) = \int_A \phi(y \,|\, x, \sigma^2) dy$$

$$\text{mit } f(y|x) = \phi(y \,|\, x, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right) = p(x, y)$$

– AR(1)-Prozess:

$$X_{t+1} = aX_t + \epsilon_t, \quad \epsilon_t \overset{iid}{\sim} N(0, \sigma^2), \quad p(x, y) = \phi(y \,|\, ax, \sigma^2)$$

– analog: mehrdimensionale Irrfahrten bzw. AR-Prozesse, $S = \mathbb{R}^p$:

$$X_{t+1} = AX_t + \epsilon_t, \; A \; (p \times p)\text{-Matrix}, \; \epsilon_t \overset{iid}{\sim} N_p(0, \Sigma).$$

- Der t-schrittigen Übergangsmatrix $(p_{ij}^{(t)}) = P^t$ entspricht die t-te Iteration des Übergangskerns (Wahrscheinlichkeit, dass die Kette nach t Schritten einen Zustand in A erreicht, ausgehend vom Startwert x):

$$\begin{aligned} P^t(x, A) &= P(X_t \in A \mid X_0 = x) \\ &= \int P^{t-1}(y, A) P(x, dy) \end{aligned}$$

- Definition: Invariante Verteilung

  Eine Verteilung π^* auf $(S, \mathcal{S})$ mit Dichte π (bezüglich des dominierenden Maßes) heißt invariante Verteilung für den Übergangskern $P(x, A)$:$\Leftrightarrow$ Für alle $A \in \mathcal{S}$ gilt

$$\pi^*(A) = \int P(x,\, A)\pi(x)dx.$$

- Definition: Irreduzible Markov-Ketten

  Eine MK mit invarianter Verteilung π^* heißt irreduzibel, wenn sie positive Wahrscheinlichkeiten besitzt, von einem beliebigen Startpunkt x_0 aus jede Menge A zu erreichen, für die $\pi^*(A) > 0$ gilt:

  $\pi^*(A) > 0 \Rightarrow$ für jedes $x_0 \in S$ gibt es ein $t \geq 1$, so dass $P^t(x_0, A) > 0$.

- Definition: (A-)periodische Markov-Ketten

  Eine Markov-Kette heißt periodisch, wenn sie Teile des Zustandsraums nur in einer bestimmten Reihenfolge erreichen kann, andernfalls heißt die Markov-Kette aperiodisch.

- Grenzwertsatz

  Sei $\{X_t, t \in \mathbb{N}_0\}$ eine irreduzible, aperiodische Markov-Kette mit Übergangskern P und invarianter Verteilung π^*. Dann ist π^* eindeutig bestimmt und es gilt

$$||P^t(x,.) - \pi^*|| \to 0 \text{ für } t \to \infty,$$

  wobei

$$||P^t(x,.) - \pi^*|| = \sup_{A \in \mathcal{S}} |P^t(x,A) - \pi^*(A)|.$$

3.8 MCMC und Metropolis-Hastings-Algorithmus

- Bayesianische Inferenz:

  Bezeichne $D = (d_1, \ldots, d_n)$ Daten, x unbekannte Parameter in der Likelihood

  $$L(x|D) = f(D|x),$$

  f die gemeinsame Dichte der Beobachtungen $D = (d_1, \ldots, d_n)$ und $p(x)$ die Dichte der Priori-Verteilung der Parameter x.

  Dann ist nach dem Satz von Bayes

  $$\pi(x) := p(x|D) = \frac{L(x|D)p(x)}{\int L(x|D)p(x)dx} = \frac{L(x|D)p(x)}{L(D)}$$

  die Dichte der Posteriori-Verteilung der unbekannten Parameter.

  Im Allgemeinen ist $\pi(x)$ wegen der Integration im Nenner nicht analytisch zugänglich. Dagegen ist der Zähler bekannt und berechenbar.

- Könnten Zufallszahlen $x^{(t)}$, $t = 1, 2, \ldots, N$ aus der Posteriori-Verteilung gezogen werden, so könnten Momente der Posteriori-Verteilung und andere interessierende Größen durch entsprechende Analoga der empirischen Verteilung approximiert werden, z.B.

$$E_\pi(g(x)) = \int g(x)\pi(dx) \approx \frac{1}{N}\sum_{t=1}^{N} g(x^{(t)}),$$

für geeignete Funktionen g, oder

$$\pi^*(A) \approx \frac{1}{N}\sum_{t=1}^{N} I_{\{x^{(t)}\in A\}}.$$

Dabei kann x auch hochdimensional sein.

- Ziel von Markov Chain Monte Carlo (MCMC) Techniken ist die Konstruktion einer Markov-Kette, deren Übergangskern P gegen die gesuchte Posteriori-Verteilung π^* konvergiert (d.h. π^* soll die invariante Verteilung der Markov-Kette sein).

- Bisher war die Übergangsmatrix P bzw. der Übergangskern P gegeben. Daraus kann die (eindeutige) invariante stationäre Verteilung π bzw. π^* bestimmt werden. Jetzt betrachtet man das umgekehrte Problem: Finde zu gegebener Verteilung π^* eine Markov-Kette mit Übergangskern P und π^* als invarianter Verteilung. Dabei ist P nicht eindeutig!

  z.B. binäre MK mit $P = \begin{pmatrix} 1-p_0 & p_0 \\ p_1 & 1-p_1 \end{pmatrix}$

$$\Rightarrow \pi_0 = \frac{p_1}{p_0 + p_1}, \quad \pi_1 = \frac{p_0}{p_0 + p_1}$$

$\pi = \left(\frac{2}{3}, \frac{3}{5}\right)$ ergibt sich z.B. für $p_0 = \frac{3}{10}$, $p_1 = \frac{2}{10}$,

$p_0 = \frac{3}{7}$, $p_1 = \frac{2}{7}$ etc.

- Sei $q(x, y)$ eine sogenannte Vorschlagsdichte, mit der ein neuer Zustand Y der Markovkette vorgeschlagen wird. Sei weiterhin

$$\alpha(x, y) = \min\left\{\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, 1\right\}$$

die Akzeptanzwahrscheinlichkeit, mit der $Y = y$ als neuer Zustand akzeptiert wird. Definiere

$$p(x, y) = \begin{cases} q(x, y)\alpha(x, y) & x \neq y \\ 0 & \text{sonst} \end{cases}$$

und

$$r(x) = 1 - \int p(x, y)dy$$

(Wahrscheinlichkeit, dass der alte Zustand beibehalten wird)

und betrachte den Übergangskern

$$P(x, A) = \int_A p(x, y)dy + r(x)\delta_x(A)$$

mit

$$\delta_x(A) = 1 \Leftrightarrow x \in A.$$

Eine Markovkette mit dem so definierten Übergangskern besitzt π^* als invariante Verteilung. Somit konvergieren die Iterationen des Kerns gegen π^*, falls die Kette zusätzlich irreduzibel und aperiodisch ist.

Beweis der Invarianzeigenschaft:

Zu zeigen ist

$$\pi^*(A) = \int P(x,\, A)\pi(x)dx$$

für den gegebenen Kern $P(x, A)$.

Zunächst gilt die Umkehrbarkeitsbedingung $\pi(x)p(x,\, y) = \pi(y)p(y,\, x)$ für $x \neq y$:

$$\begin{aligned}
\pi(x)p(x,\, y) &= \pi(x)q(x,\, y)\alpha(x,\, y) \\
&= \pi(x)q(x,\, y)\min\left\{\frac{\pi(y)q(y,\, x)}{\pi(x)q(x,\, y)}, 1\right\} \\
&= \min\left\{\pi(y)q(y,\, x),\, \pi(x)q(x,\, y)\right\} \\
&= \pi(y)q(y,\, x)\min\left\{1,\, \frac{\pi(x)q(x,\, y)}{\pi(y)q(y,\, x)}\right\} \\
&= \pi(y)q(y,\, x)\alpha(y, x) \\
&= \pi(y)p(y,\, x)
\end{aligned}$$

Die Umkehrbarkeitsbedingung ist hinreichend für die Invarianzeigenschaft, denn

$$\begin{aligned}
\int P(x,\, A)\pi(x)dx &= \int\left(\int_A p(x,\, y)dy\right)\pi(x)dx + \int r(x)\delta_x(A)\pi(x)dx \\
&= \int_A\left(\int p(x,\, y)\pi(x)dx\right)dy + \int_A r(x)\pi(x)dx
\end{aligned}$$

$$\begin{aligned} &\overset{\text{Umk.bed.}}{=} && \int_A \left(\int p(y,\, x)\pi(y)dx \right) dy + \int_A r(x)\pi(x)dx \\ &= && \int_A (1 - r(y))\, \pi(y)dy + \int_A r(x)\pi(x)dx \\ &= && \int_A \pi(y)dy = \pi^*(A) \end{aligned}$$

- Metropolis-Hastings-Algorithmus:

  i) Wähle Startwerte x_0 und die Länge N der Kette. Setze $t = 1$.

  ii) Ziehe eine Zufallszahl Y aus $q(x_{t-1}, y)$und akzeptiere diese als neuen Zustand x_t mit Wahrscheinlichkeit $\alpha(x_{t-1}, y)$, andernfalls setze $x_t = x_{t-1}$.

  iii) Falls $t = N$ beende den Algorithmus, andernfalls setze $t = t + 1$ und fahre fort mit ii).

- Bemerkungen:
  - Nach einer gewissen Konvergenzphase (Burn In) können die gezogenen Zufallszahlen $x_0, x_1, \ldots, x_N$ als Realisierungen aus π^* angesehen werden. Eigenschaften von π^* kann man nun mit Hilfe der Zufallszahlen schätzen, z.B. den Erwartungswert durch $\bar{X}$, wobei üblicherweise die Burn In Phase unberücksichtigt bleibt.
  - Falls $q(x, y)$ symmetrisch ist, d.h $q(x, y) = q(y, x)$, vereinfacht sich die Akzeptanzwahrscheinlichkeit zu

$$\alpha(x, y) = \min\left\{\frac{\pi(y)}{\pi(x)}, 1\right\}.$$

    In diesem Fall spricht man vom sogenannten Metropolis-Algorithmus.
  - Die Vorschlagsdichte $q(x, y)$ sollte so gewählt werden, dass man leicht Zufallszahlen aus ihr ziehen kann und die Akzeptanzraten nicht zu klein werden (z.B. multivariate Normalverteilung).

- Beispiele für die Konstruktion von Vorschlagsdichten:
  - Unabhängige Ketten
    $$q(x, y) = q(y).$$
  - Random Walk Kette: Ziehe eine Zufallszahl Z aus einer Dichte $f(z)$ und setze $Y = x_{n-1} + Z$ d.h.
    $$q(x_{n-1}, y) = f(y - x_{n-1}).$$
  - Ist $x = (x_1', \ldots, x_p')'$ multivariat mit Teilvektoren (oder Skalaren) $x_1, \ldots, x_P$, so kann man den MH-Algorithmus auch komponentenweise anwenden.

- MCMC-Verfahren sind besonders nützlich bei hochdimensionalem Parametervektor x, der sich in Blöcke zerlegen lässt, die Teile des Modells beschreiben: $x = (x_1', \ldots, x_p')'$. Sukzessive Anwendung des MCMC-Algorithmus für die Blöcke:
  Sei $x'_{-j} = (x_1', \ldots, x'_{j-1}, x'_{j+1}, \ldots, x'_p)$ der Parametervektor ohne x_j. Update von x_j, d.h. $y_{-j} = x_{-j}^{(t)}$.

$$\Rightarrow \frac{\pi(y)}{\pi(x^{(t)})} = \frac{\pi(y_j, x_{-j}^{(t)})}{\pi(x_j^{(t)}, x_{-j}^{(t)})} = \frac{\pi(y_j \mid x_{-j}^{(t)})}{\pi(x_j^{(t)} \mid x_{-j}^{(t)})}$$

Sind bedingte Dichten bekannt, d.h. man kann aus ihnen ziehen, so wählt man als Vorschlagsdichte $q(x, y) = \pi(y_j \mid x_{-j}^{(t)}) \Rightarrow \alpha(x, y) = 1$ ('Gibbs sampler')

- Damit Iterationen des Übergangskerns tatsächlich gegen π^* konvergieren, ist neben der Invarianzeigenschaft zusätzlich Irreduzibilität und Aperiodizität erforderlich. In der Praxis sind diese Eigenschaften fast immer erfüllt, können aber theoretisch schwer nachgewiesen werden. Die Konvergenz wird daher in der Regel durch Analyse des MCMC Outputs überprüft, z.B. durch
  - grafische Darstellung der Samplingpfade
  - Berechnung der Autokorrelationsfunktion der gezogenen Zufallszahlen.