Theoretische Biophysik
-
Statistische Physik

14. Vorlesung

Pawel Romanczuk

Wintersemester 2018

http://lab.romanczuk.de/teaching/

Zusammenfassung letzte VL

- Aktive Brownsche Bewegung
- Effektive Diffusion Aktive Brownsche Bewegung und Run & Tumble Dynamik
- Informationstheorie - Einleitung
- Shannon-Entropie als Maß der Unsicherheit

Shannon Entropie als Informationsmaß

$$H = -\lambda \sum_i p_i \ln p_i = -\sum_i p_i \log_2 p_i \qquad \lambda = \frac{1}{\ln 2}$$

Interpretation:
Maß der (beseitigten) Unsicherheit über den Systemzustand
→ Wie viele JA-Nein Fragen müssen im Mittel gestellt werden, bis wir volle Information über das System haben (den Mikrozustand des Systems komplett bestimmt haben)?

Frage: Besitzt die Shannon-Entropie alle Eigenschaften über die ein Informationsmaß verfügen sollte?

Positivität

Bevor ein Ereignis statt findet haben wir einen Informationsmangel, d.h. nach dem Eintreten des Erignisses bzw. Kennenlernen des Ergebnisses haben wir immer einen Informationsgewinn

→ **Ein Informationsmaß muß immer positiv sein.**

Für beliebige Wahrscheinlichkeitsverteilungen gilt:

$$p_i \leq 1 \quad \Rightarrow \quad -\log_2 p_i \geq 0$$

bzw. für die Shannon-Entropie:

$$H = -\sum_i p_i \log_2 p_i \geq 0$$

Shannon-Entropie hat ein Minimum: $H=0$

Wir nehmen an $p_m=1$ und alle $p_n=0$ $(n\neq m)$, d.h. das Ereignis E_m tritt mit absoluter Sicherheit ein.

→ **Kennenlernen eines sicheren Ergebnisses beseitigt keine Unsicherheit - Es kein Informationsgewinn.**

Logarithmus von 0 ist nicht definiert (divergiert gegen -∞), aber es gilt:

$$\lim_{p_n \to 0} p_n \log_2 p_n = 0$$

Daher erhalten wir für den Fall der speziellen Fall eines absolut sicheren Ergebnisses:

$$H = -\underbrace{\lim_{p_m \to 1} p_m \log_2 p_m}_{=0} - \underbrace{\sum_{n \neq m} \lim_{p_n \to 0} p_n \log_2 p_n}_{=0}$$

$$H = 0$$

Shannon-Entropie hat ein Maximum

Ein Ereignis kann am wenigsten vorausgesagt werden wenn Z mögliche Ergebnisse (Zustände) gleich wahrscheinlich sind.

→ **größter Informationsgewinn – Maximum von H!**

Bestimmung des Extremums der Shannon-Entropie mit Hilfe des Langrage‘schen Multiplikator μ (Analogie zur Entropiemaximierung ohne Energieerhaltung, siehe VL zu Boltzmannverteilung):

$$H' = -\lambda \sum p_i \ln p_i + \mu\left(-1 + \sum_i p_i\right)$$

$$\frac{\partial H'}{\partial p_j} = -\ln p_j - 1 + \mu = 0 \quad \Rightarrow \quad p_j = e^{\mu - 1}$$

$$\sum_{j=1}^{Z} p_j = \sum_{j=1}^{Z} e^{\mu-1} = Z e^{\mu-1} = 1 \qquad p_j = \frac{1}{Z}$$

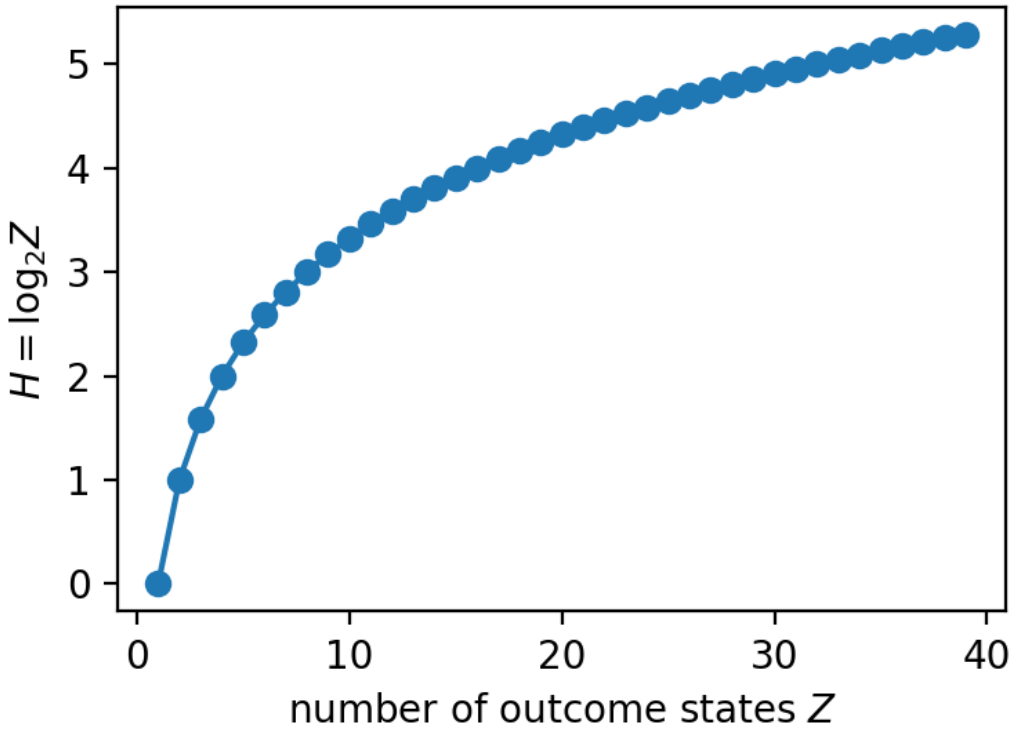
→ **Gleichverteilung!**

Maximale Entropie H wächst monoton mit Z

Die Unsicherheit steigt mit der Zahl der möglichen Ergebnisse Z (p_i>0) an.

→ **Gewonnene Information pro Beobachtung steigt mit Z**

Gleichverteilung: $p_i = \frac{1}{Z}$ $H = -Z\frac{1}{Z}\log_2\frac{1}{Z} = \log_2(Z)$



Shannon-Entropie ist additiv

Wir betrachten zwei voneinander unabhängige Ereignisse aus zwei Ereignismengen:

$$x_i = \{1, 2, \ldots, m\} \qquad y_i = \{1, 2, \ldots, n\}$$

(z.B. gleichzeitiges Werfen von einer Münze und einem Würfel)

Wahrscheinlichkeit für gleichzeitige Auftreten des Ereignisses i (aus x) und j (aus y) falls x und y unabhängig (Definition):

$$p(i, j) = p_i p_j$$

Verbundinformation („Verbundentropie“, engl.: joint entropy)

$$\begin{aligned} H(x, y) &= -\sum_{ij} p(i, j) \log_2 p(i, j) \\ &= -\sum_i \sum_j p_i p_j \log_2 (p_i p_j) \end{aligned}$$

Shannon-Entropie ist additiv

$$
\begin{aligned}
H(x,y) &= -\sum_i \sum_j p_i p_j \log_2(p_i p_j) \\
&= -\sum_i \sum_j p_i p_j (\log_2 p_i + \log_2 p_j) \\
&= -\sum_i \sum_j p_i p_j \log_2 p_i - \sum_i \sum_j p_i p_j \log_2 p_j \\
&= -\sum_j p_j \sum_i p_i \log_2 p_i - \sum_i p_i \sum_j p_j \log_2 p_j \\
&= -\sum_i p_i \log_2 p_i - \sum_j p_j \log_2 p_j \\
&= H(x) + H(y)
\end{aligned}
$$

Axiomatische Herleitung der Shannon-Entropie

Man kann die Shannon-Entropie H auch direkt aus der Forderung herleiten, dass ein Informationsmaß die obigen Axiome erfüllen muss:

- Positivität
- Minimum $H=0$ bei absolut sicherem Ergebnis
- Maximum bei gleichverteilten Wahrscheinlichkeiten der Möglichkeiten Z und monoton wachsend mit Z
- Additivität für unabhängige Variablen bzw. erweiterte Formulierung für abhängige Variablen (siehe Folgendes).

Die einzige Funktion die diese Axiom erfüllt hat die allgemeine Form:

$$H = -C \sum_i p_i \ln p_i$$

Beispiel Gesprochene Sprache

Ereignisse: Auftreten von Buchstaben (Symbole) in einer Folge.

Zahl der Symbole Z=27 (26 Buchstaben + 1 Leerzeichen)

Falls gleich wahrscheinlich:

$$H = -\sum_i p_i \log_2 p_i = \log_2 27 = 4.76\,\mathrm{bit/Zeichen}$$

Allerdings ist die Wahrscheinlichkeit des Auftretens eines Buchstaben nicht gleich verteilt. Zum Beispiel in Englisch:

Symbol	Leerzeichen	E	T	O	A	N	I	...	Q
Wahrscheinlichkeit	0.2	0.105	0.072	0.065	0.063	0.059	0.055	...	0.001

$$H = 4.04\,\mathrm{bit/Zeichen}$$

Korrelationen zwischen Buchstaben

Allerdings ist der tatsächliche Wert für H in Englisch deutlich kleiner als 4.04 bits/Zeichen.

Der Grund dafür ist, dass das Auftreten von Buchstaben nicht völlig zufällig ist. Es gibt Korrelationen zwischen Buchstaben.

Beispiel: Im Deutschen folgt nach „q“ niemals ein „m“ aber dafür sehr häufig ein „u“

Korrelationen sind sehr Sprachabhängig: Im Deutschen folgt nach „c“ sehr selten ein „z“, im Polnischen dafür aber häufig, da „cz“ als Buchstaben-kombination ist, die einem bestimmten Laut entspricht (gesprochen wie ein kurzes „tsch“)

Informationsmaß ist neutral bezogen auf Inhalt

Das Informationsmaß H ist komplett unabhängig vom Inhalt, d.h. es hat nichts mit der Bedeutung der Information zu tun. Es bezieht sich alleine auf die statistische Vorhersagbarkeit eines Ereignisses.

→ Je seltener ein Ereignis, desto größer dessen Informationsgehalt.

Vergleiche:

1) Die Textnachricht „Tante gestorben“ kann je nach Empfänger sehr unterschiedliche Bedeutung / Information haben obwohl H gleich:

- ein Empfänger empfindet große Trauer.
- ein anderer Empfänger große Freude, wegen der Erwartung eines großen Erbes

2) Ziehen einer Karte vom Stapel von 32 Karten (→ $H=5$ bit, wenn gut gemischt)

- Bedeutung für das Spiel sehr unterschiedlich je nach Spiel und Karte

Informationsgehalt bei korrelierten Informationen

Bei korrelierten (nicht unabhängigen) Informationen scheint sich die Verbundentropie zu verringern. Um dies genauer zu untersuchen benötigen wir einen **Hilfssatz:**

Sei p_i die Wahrscheinlichkeitsverteilung, die uns interessiert, q_i sein eine andere Wahrscheinlichkeitsverteilung.

$$\sum_i p_i = 1 \qquad \sum_i q_i = 1$$

Wir betrachten die Funktion: $\varphi = \sum_i p_i \ln q_i$

Dann gilt:

$$\sum p_i \ln q_i \leq \sum_i p_i \ln p_i$$

(Gleicheit gilt nur falls alle $q_i = p_i$)

Beweis des Hilfssatzes:

Sei $q_i = p_i + u_i$ dann gilt, wegen $\sum_i p_i = \sum_i q_i = 1$, $\sum_i u_i = 0$

$$\varphi = \sum_i p_i \ln q_i = \sum_i p_i \ln(p_i + u_i) = \sum_i p_i \ln \left[p_i (1 + \frac{u_i}{p_i}) \right]$$

$$\varphi = \sum_i p_i \ln p_i + \sum_i p_i \ln \left(1 + \frac{u_i}{p_i} \right)$$

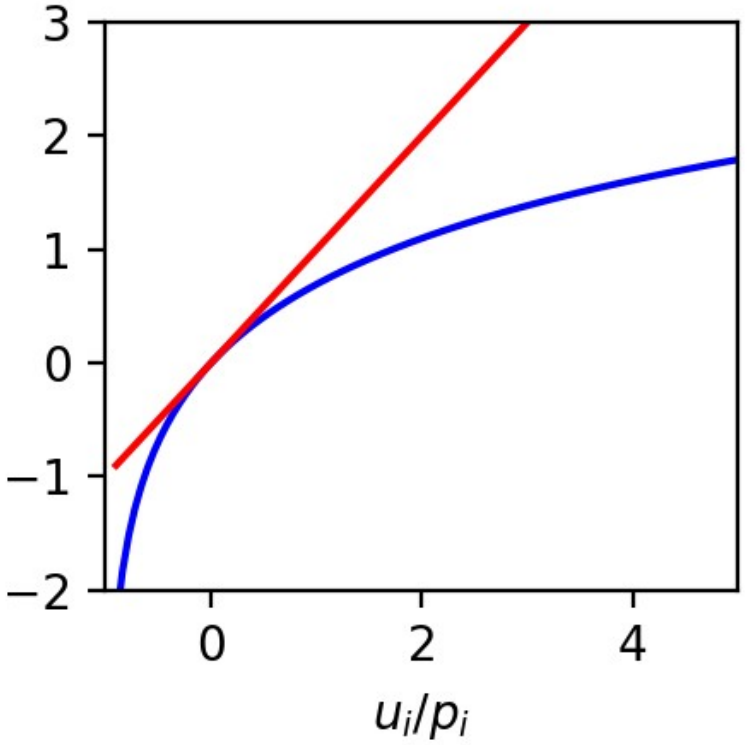
Für $u_i = 0$ gilt:

$$\ln \left(1 + \frac{u_i}{p_i} \right) = 0$$

Beweis des Hilfssatzes:

Für $u_i \neq 0$ gilt:

$$\ln\left(1+\frac{u_i}{p_i}\right) \leq \frac{u_i}{p_i}$$



$$\sum_i p_i \ln q_i = \sum_i p_i \ln p_i + \sum_i p_i \ln\left(1+\frac{u_i}{p_i}\right) \leq \sum_i p_i \ln p_i + \underbrace{\sum_i p_i \frac{u_i}{p_i}}_{=0}$$

$$\sum p_i \ln q_i \leq \sum_i p_i \ln p_i$$

Allgemeiner Betrachtung

Es gilt also: $$\sum p_i \log_2 q_i \leq \sum_i p_i \log_2 p_i$$

Betrachten wir jetzt wieder zwei Ereignismengen (korreliert oder unkorreliert):

$$x_i = \{1, 2, \ldots, m\} \qquad y_i = \{1, 2, \ldots, n\}$$

p(i,j): Verbundwahrscheinlichkeit für das simultane Auftreten von *x=i* und *y=j*.

Es gilt: 1. $$\sum_{i,j} p(i,j) = 1$$

2. $$p_i = \sum p(i,j) \quad \rightarrow \text{ unabhängig von } j \text{ (in } y\text{)}$$

$$p_j = \sum_i p(i,j) \quad \rightarrow \text{ unabhängig von } i \text{ (in } x\text{)}$$

3. $$\sum_i p_i = 1, \sum_j p_j = 1 \Rightarrow \sum_i p_i \sum_j p_j = \sum_{i,j} p_i p_j = 1$$

Verbundinformation und „Teilinformation“

Angenommen die Ereignisse aus x und y korrelieren. Was passiert wenn wir nur allein die Ereignisse aus x und alleine die Ereignisse aus y registrieren?

Erwartung: Wir verzichten auf mögliche Kenntnisse die in Korrelationen enthalten sein können, daher ist der Informationsgewinn größer (da Kenntnisse geringer).

Um das zu quantifizieren vergleichen wir die **Verbundinformation**

$$H(x,y) = -\sum_{i,j} p(i,j) \log_2 p(i,j)$$

die Information für ein Ereignis nur aus x:

$$H(x) = -\sum_{i} p_i \log_2 p_i = -\sum_{i} \sum_{j} p(i,j) \log_2 p_i$$

und die Information für ein Ereignis nur aus y:

$$H(y) = -\sum_{j} p_j \log_2 p_j = -\sum_{i} \sum_{j} p(i,j) \log_2 p_j$$

Verbundinformation und „Teilinformation“

Wir vergleichen die Summe $H(x)$ und $H(y)$ mit $H(x,y)$:

$$H(x) + H(y) = -\sum_i \sum_j p(i,j)[\log_2 p_i + \log_2 p_j]$$

$$= -\sum_i \sum_j \underbrace{p(i,j)}_{=p} \log_2 (\underbrace{p_i p_j}_{=q})$$

Für nicht unabhängige, also korrelierte, Ereignisse gilt $p(i,j) \neq p_i p_j$

Setzt man also $p_i p_j = q$ so folgt mit dem obigen Hilfssatz:

$$\sum_i \sum_j p(i,j) \log_2 (p_i p_j) \geq \sum_i \sum_j p(i,j) \log_2 p(i,j)$$

$$H(x) + H(y) \geq H(x,y)$$

Berücksichtigung von Korrelationen → geringerer Informationsgewinn
(Gleichheit nur bei unabhängigen Prozessen!)

Bedingte Information

Betrachte bedingte Information für das gemeinsame Auftreten der Ereignisse i und j aus den Mengen x und y.

Bedingte Wahrscheinlichkeit (Kurzschreibweise): $p_i(j) := p(j|i)$

$$p(i,j) = p_i \cdot p_i(j)$$

Wahrscheinlichkeit, dass sowohl i als auch j eintritt

Wahrscheinlichkeit, dass i eintritt

Wahrscheinlichkeit, dass j eintritt unter der Bedingung dass i eingetreten ist.

Es gilt also: $$p_i(j) = \frac{p(i,j)}{p_i}$$

Wir betrachten wieder die Verbundinformation:

$$H(x,y) = \sum_{i,j} p(i,j) \log_2 p(i,j)$$

Bedingte Information

Ersetzen im Logarithmus von $p(i,j)$ durch $p_i p_i(j)$:

$$H(x,y) = -\sum_i \sum_j p(i,j) \log_2 (p_i p_i(j))$$

$$= -\sum_i \underbrace{\sum_j p(i,j) \log_2 p_i}_{= p_i} - \sum_i \sum_j p(i,j) \log_2 p_i(j)$$

$$= -\sum_i p_i \log_2 p_i \; - \sum_i \sum_j p(i,j) \log_2 p_i(j)$$

Bedingte Information: $$H_x(y) = -\sum_i \sum_j p(i,j) \log_2 p_i(j)$$

$$H(x,y) = H(x) + H_x(y)$$

Informationsgewinn beim Messen von x und y = Informationsgewinn beim Messen von x alleine + Informationsgewinn beim Messen von y bei **bereits bekannten** x

Bedingte Information

Wenn x und y nicht unabhängig sind dann liefert das Auftreten von einem bestimmten Ereignis aus x bereits Informationen über möglichen Ereignisse aus y!

Mit der zuvor abgeleiteten Ungleichung gilt:

$$H(x) + H(y) \geq H(x, y) = H(x) + H_x(y)$$

bzw.

$$\boxed{H(y) \geq H_x(y)}$$

→ Wenn x und y nicht unabhängig sind, dann ist der Informationsgewinn wenn wir nur y messen größer, als wenn wir y bei bereits bekannten x messen.

Korrelation und Redundanz in der Sprache

Das Auftreten „bedingter Wahrscheinlichkeiten“ ist typisch für Wahrnehmung der Sprache (sowohl gesprochen wie geschrieben).

Für Sprache (wie auch für dynamische Prozesse) existiert eine sequentielle Ordnung bzw. zeitliche Reihenfolge:

1. Zeichen, 2. Zeichen, 3. Zeichen, usw.

Beim Hören (Lesen) des m-ten Zeichens ist das (m-1)-te bereits bekannt.

Frage: Wie viel Informationen enthält also überhaupt noch das m-te Zeichen?

→ **Im Extremfall gar keine!**

Korrelation und Redundanz in der Sprache

In jedem Fall verringert sich der Informationsgehalt durch Betrachtung bedingter Wahrscheinlichkeiten, z.B.:

auf „Q“ folgt in der Regel „U“; auf „C“ folgt oft „H“, etc.

Das sind Paarkorrelationen. Beim Berücksichtigen weiterer, langreichweitiger Korrelationen, sinkt der Informationsgewinn weiter: Wissen wir z.B. bereits „WOHNUN“ können wir sehr leicht ergänzen zu „WOHNUNG“.

Kürzere Zeichenketten normalerweise mehr mögliche Ergänzungen, z.B.:

„UN“ → „UND“

„UNO“

„UNI“

usw.

Korrelation und Redundanz

Beispiel: Englische Sprache.

1) nur 27 Zeichen (Annahme der Gleichverteilung, ohne Berücksichtigung der tatsächlichen Wahrscheinlichkeiten):

$$H_0 = 4.76\mathrm{bit/Zeichen}$$

2) Berücksichtigung der tatsächlichen Wahrscheinlichkeitsverteilung:

$$H_1 = 4.04\mathrm{bit/Zeichen}$$

3) Berücksichtigung der Paarkorrelationen (erfordert Berechnung / Bestimmung der bedingten Wahrscheinlichkeiten):

$$H_2 = 3.32\mathrm{bit/Zeichen}$$

Korrelation und Redundanz

Das Konzept kann verallgemeinert werden, indem man Zeichenketten betrachtet („Blöcke“ der Länge N)

$b_i(N-1)$: ein Block von Zeichen der Länge *(N-1)*, mit dem Index i

$N=4$: „SCH“, „URS“, “WOI“, ...

$i=1$ $i=2$ $i=3$ ...

$p\big(b_i(N-1)\big)$: Wahrscheinlichkeit, dass dieser Block in einem („langen“) Text auftritt.

Korrelation und Redundanz

Wir ergänzen beim Lesen einen Block aus N-*1* Zeichen durch das N-te Zeichen

$b_{ij}(N) = (b_i(N-1), j)$: ein Block von N Zeichen zusammengesetzt aus dem Block von N-*1* Zeichen mit dem Index i, ergänzt durch ein weiteres Zeichen *j*

Anwendung des Konzeptes der bedingten Wahrscheinlichkeiten:

$$p\big(b_{ij}(N)\big) = p\big(b_i(N-1), j\big) = p\big(b_i(N-1)\big) \cdot p_{b_i(N-1)}(j)$$

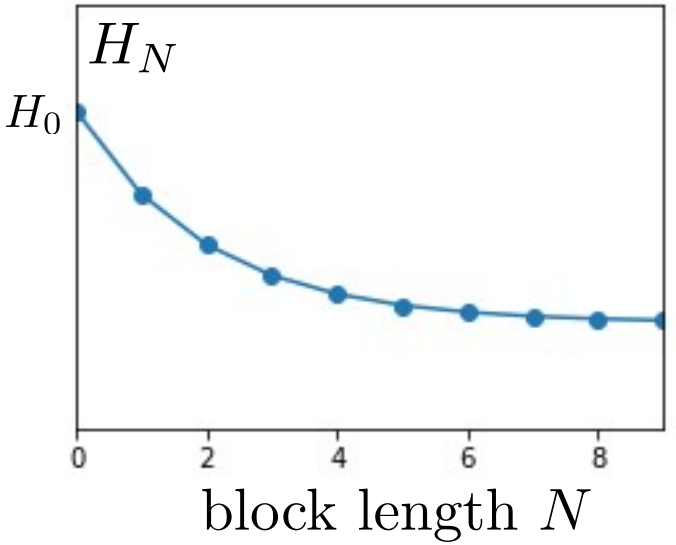
Damit können wir den durchschnittlichen Informationsgehalt eines Zeichens bei Kenntnis der vorangehenden N-*1* Zeichen berechnen:

$$H_N = -\sum_{i,j} p\big(b_i(N-1), j\big) \log_2 p_{b_i(N-1)}(j)$$

Korrelation und Redundanz

H_N ist monoton fallend mit N; siehe Beispiel englische Sprache:

$$H_0 = 4.76\text{bit/Zeichen}$$
$$H_1 = 4.04\text{bit/Zeichen}$$
$$H_2 = 3.32\text{bit/Zeichen}$$
$$H_3 = 3.1\text{bit/Zeichen}$$



Der Grenzfall $H = \lim_{N\to\infty} H_N$ ist schwer zu Berechnen da man die Häufigkeiten langer Zeichenketten in noch längeren Texten bestimmen muss.

Abschätzungen ergeben: $H = 2.14\text{bit/Zeichen}$

Redundanz (Abfall der Information bei Berücksichtigung der Korrelationen):

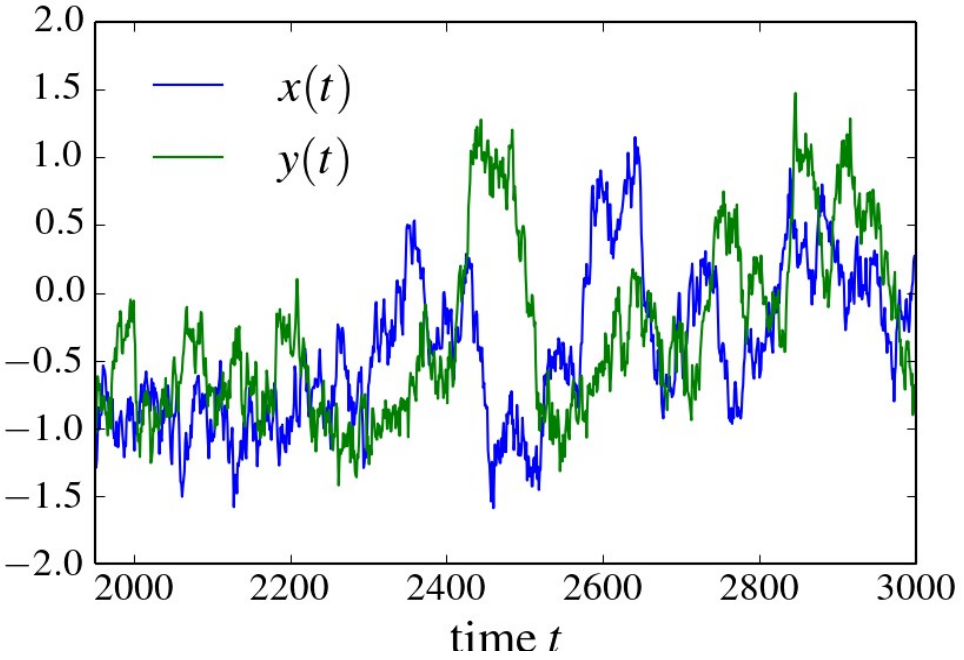
$$R = \frac{\Delta H}{H_0} = \frac{H_0 - H}{H_0} = 1 - \frac{H}{H_0}$$

hier: $R \approx 0.55$

Detektion Kausaler Kopplung zwischen Zeitreihen

- Wie nehmen an wir haben zwei Zeitreihen aus der Beobachtung eines komplexen Systems:

$$X: \quad x_1, x_2, x_3, \ldots x_n$$
$$Y: \quad y_1, y_2, y_3, \ldots y_n$$



- Können wir die Richtung, Stärke und eventuelle Verzögerung einer möglichen Kopplung zwischen X und Y bestimmen?

Korrelation ist nicht Kausalität!

Transfer-Entropie Schreiber, Phys Rev Lett 85 (2000)

- TE: parameterloses Maß eines gerichteten „Informationstransfers“ zwischen zwei stochastischen Prozessen:

$$T_{X\to Y} = H(y_{t+1}|y_t^{(k)}) - H(y_{t+1}|y_t^{(k)}, x_t^{(l)})$$

Wert von Y zum nächsten Zeitpunkt

Werte von Y für k Zeitpunkte in der Vergangenheit

Werte von X für k Zeitpunkte in der Vergangenheit

$$T_{X\to Y} = \sum_t p(y_{t+1}, y_t^{(k)}, x_t^{(l)}) \log \frac{p(y_{t+1}|y_t^{(k)}, x_t^{(l)})}{p(y_{t+1}|y_t^{(k)})}$$

- Wie viel Information liefert uns die Vergangenheit von X über die Zukunft von Y wenn wir gleichzeitig die Vergangenheit von Y in Betracht ziehen?

Transfer-Entropie Schreiber, Phys Rev Lett 85 (2000)

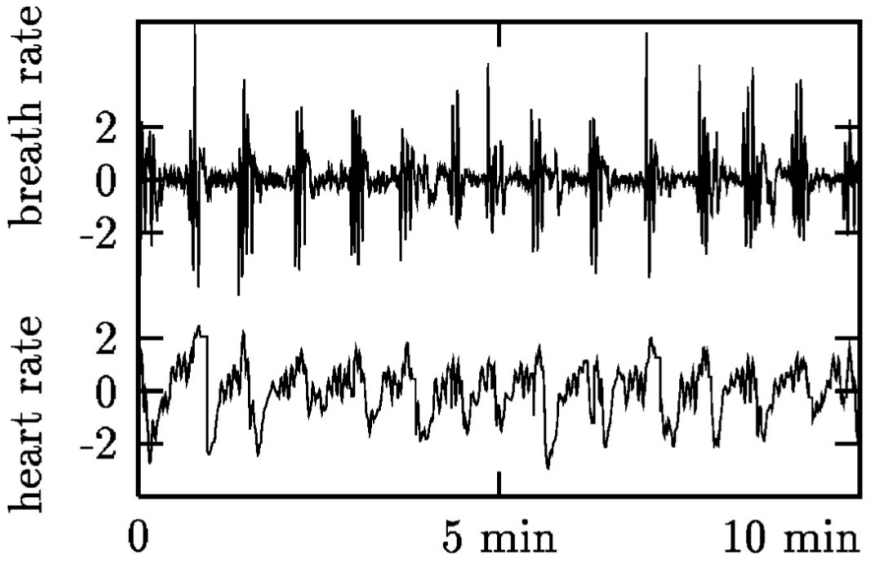


FIG. 3. Bivariate time series of the breath rate (upper) and instantaneous heart rate (lower) of a sleeping human. The data is sampled at 2 Hz. Both traces have been normalized to zero mean and unit variance.

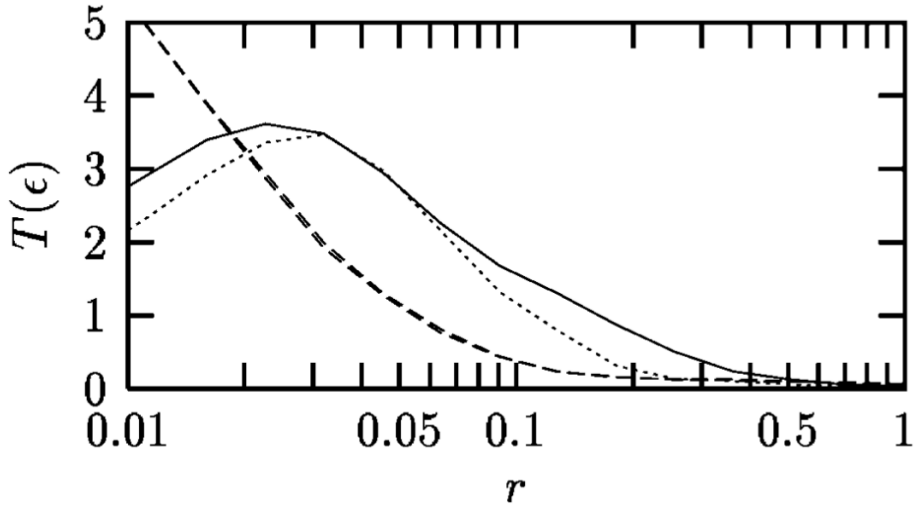


FIG. 4. Transfer entropies $T(\text{heart} \rightarrow \text{breath})$ (solid line), $T(\text{breath} \rightarrow \text{heart})$ (dotted line), and time delayed mutual information $M(\tau = 0.5\ \text{s})$ (directions indistinguishable, dashed line) for the physiological time series shown in Fig. 3.

- Anwendung von TE auf reale physiologische Zeitserien, bestimmt die Kopplungsrichtung und Verzögerung, die nicht mit zeitverzögerter Transinformation (engl. „time-lagged mutual information“) bestimmt werden kann.