5 Statistische Grundlagen (Teil 2)

Die Kovarianz zweier Zufallsvariablen gibt an, ob sich diese tendenziell ähnlich verhalten bezüglich ihrer Abweichungen von ihrem jeweiligen Mittelwert. Die Korrelation ist die normierte Kovarianz. Kovarianzen können für Zufallsvariablen und für geordnete Stichproben (Zeitreihen) berechnet werden.
Die Autokovarianz einer Zeitreihe zum Abstand j gibt an, ob sich mit Abstand j aufeinanderfolgende Werte einer Zeitreihe tendenziell ähnlich verhalten. Die normierte Autokovarianz nennt sich Autokorrelation. Sie kann durch Messobjekte gemessen werden.

5.1 Kovarianz zweier Zufallsvariablen (LK 4.2)

- Definition: **Kovarianzfunktion**

$$\mathrm{Cov}[X,Y] = \mathrm{E}[(X - \mathrm{E}[X]) \cdot (Y - \mathrm{E}[Y])] = \mathrm{E}[X \cdot Y] - \mathrm{E}[X] \cdot \mathrm{E}[Y] \tag{2.25}$$

  - X, Y unabhängig $\Rightarrow$ Cov[X, Y]=0
    - Beweis (LK Problem 4.8)

$$\mathrm{Cov}[X,Y] = \mathrm{E}[X \cdot Y] - \mathrm{E}[X] \cdot \mathrm{E}[Y] =$$

$$\int_{-\infty}^{\infty} x \cdot y \cdot f(x,y)\,dx\,dy - \mathrm{E}[X] \cdot \mathrm{E}[Y] \overset{\text{Unabhängigkeit}}{=}$$

$$\int_{-\infty}^{\infty} x \cdot y \cdot f(x) \cdot f(y)\,dx\,dy - \mathrm{E}[X] \cdot \mathrm{E}[Y] =$$

$$\mathrm{E}[X] \cdot \mathrm{E}[Y] - \mathrm{E}[X] \cdot \mathrm{E}[Y] = 0$$

  - ABER nicht: Cov[X, Y]=0 $\Rightarrow$ X, Y unabhängig
    - Gegenbeispiel (LK Problem 4.9)
      Sei X diskrete ZV mit gleichwahrscheinlichen Ereignissen -2, -1, 1 und 2. Sei $Y=X^2$. Offensichtlich sind X und Y nicht unabhängig, aber sie sind unkorreliert:

$$\mathrm{E}[X] = 0$$

$$\mathrm{Cov}[X,Y] = \mathrm{E}[X \cdot Y] - \mathrm{E}[X] \cdot \mathrm{E}[Y] =$$

$$\frac{1}{4}(-2 \cdot 4 - 1 \cdot 1 + 1 \cdot 1 + 2 \cdot 4) - 0 \cdot \mathrm{E}[Y] = 0$$

5.2 Korrelation zweier Zufallsvariablen (LK 4.2)

- Definition: **Korrelationsfunktion**

$$Cor[X,Y] = \frac{Cov[X,Y]}{\sqrt{VAR[X] \cdot VAR[Y]}} \tag{2.26}$$

- Spezialfälle
  - Maximale positive Korrelation
    Y=X $\Rightarrow$ Cor[X, Y]=1
  - Maximale negative Korrelation
    Y=-X $\Rightarrow$ Cor[X, Y]=-1
- Die Korrelationsfunktion beschreibt auf normierte Weise, wie stark zwei Zufallsvariablen gleichzeitig von ihrem Mittelwert abweichen.
  - Cor[X,Y]>0: gleichzeitige Differenzen der beiden ZV X und Y von ihrem jeweiligen Mittelwert tendieren gleiches Vorzeichen zu haben.
  - Cor[X,Y]<0: gleichzeitige Differenzen der beiden ZV X und Y von ihrem jeweiligen Mittelwert tendieren unterschiedliches Vorzeichen zu haben.

5.3 Empirische Kovarianz und Korrelation zweier Zeitreihen

Das Konzept der Kovarianz (Korrelation) Cov[X, Y] wurde für zwei Zufallsvariablen definiert und die Formeln entsprechen wie bei den Momenten auch einer Auswertung über den Wertebereich ihres Verbundereignisses (X, Y). Bei der empirischen Kovarianz wird anstelle des Wertebereiches die Zeit als Grundlage für die Auswertung genommen.

a) Zeitdiskreter Fall

$$\overline{COV}[X,Y] = \frac{1}{n-1} \cdot \sum_{0 \le i < n} \left(X_i - \overline{X}(n)\right) \cdot \left(Y_i - \overline{Y}(n)\right) \tag{2.32}$$

b) Zeitkontinuierlicher Fall

$$\overline{COV}[X,Y] = \frac{1}{T} \cdot \int_0^T \left(X(t) - \overline{X}\right) \cdot \left(Y(t) - \overline{Y}\right) dt \tag{2.33}$$

Beweis

Idee des Beweises: wir berechnen den Erwartungswert der empirischen Kovarianz $E\left[\overline{COV}[X,Y]\right] = \frac{1}{n-1} \cdot \sum_{0 \le i < n} E\left[\left[X_i - \overline{X}(n)\right] \cdot \left[Y_i - \overline{Y}(n)\right]\right]$, die sich als Funktion der Zufallsvariablen X_i und Y_j berechnet, und zeigen, dass dieser gleich der Kovarianz von X und Y ist, wenn alle X_i und Y_j nach X und Y verteilt sind.

Es gilt:

$$E\left[X_i \cdot \overline{Y}(n)\right] = E\left[X_i \cdot \frac{1}{n} \cdot \sum_{0 \le j < n} Y_j\right] = E\left[\frac{1}{n} \cdot X_i \cdot Y_i + \frac{1}{n} \cdot X_i \cdot \sum_{0 \le j < n,\, j \ne i} Y_j\right] =$$

$$= \frac{1}{n} \cdot E[X_i \cdot Y_i] + \frac{1}{n} \cdot E\left[E[X_i] \cdot \sum_{0 \le j < n,\, j \ne i} E[Y_j]\right] = \frac{1}{n} \cdot E[X \cdot Y] + \frac{n-1}{n} \cdot E[X] \cdot E[Y]$$

Hinweis: Bei dieser Rechnung wurden folgende Rechenregeln verwendet:

1. X_i und Y_j werden nicht gleichzeitig realisiert und sind deswegen unabhängige Zufallsvariablen. Darum darf beim nachfolgenden Ausdruck der Erwartungswert auch auf die Zufallsvariablen der Stichproben ein zweites Mal angewendet werden:
   $$E\left[X_i \cdot \sum_{j \ne i} Y_j\right] = E\left[E[X_i] \cdot \sum_{j \ne i} E[Y_j]\right]$$
2. Die Stichprobenwerte X_i (auch Y_i) sind identisch verteilt: $E[X_i] = E[X]$
3. Die Erwartungswerte von Erwartungswerten sind wieder die Erwartungswerte: $E[E[X]] = E[X]$

Weiter gilt:

$$E\left[\overline{X}(n)\cdot\overline{Y}(n)\right]=E\left[\left(\frac{1}{n}\cdot\sum_{0\le i<n}X_i\right)\cdot\left(\frac{1}{n}\cdot\sum_{0\le i<n}Y_i\right)\right]=$$

$$\frac{1}{n^2}\cdot E\left[\sum_{0\le i<n}X_i\cdot Y_i+\sum_{0\le i,j<n,i\neq j}X_i\cdot Y_j\right]=\frac{1}{n}\cdot E[X\cdot Y]+\frac{n-1}{n}\cdot E[X]\cdot E[Y]$$

Damit ist der Erwartungswert der empirischen Varianz erwartungstreu:

$$E\left[\overline{COV}[X,Y]\right]=\frac{1}{n-1}\cdot\sum_{0\le i<n}E\left[\left[X_i-\overline{X}(n)\right]\cdot\left[Y_i-\overline{Y}(n)\right]\right]=$$

$$=\frac{1}{n-1}\cdot\sum_{0\le i<n}\left(E[X_i\cdot Y_i]-E\left[X_i\cdot\overline{Y}(n)\right]-E\left[Y_i\cdot\overline{X}(n)\right]+E\left[\overline{X}(n)\cdot\overline{Y}(n)\right]\right)=$$

$$=\frac{1}{n-1}\cdot\sum_{0\le i<n}\left(\begin{array}{l}E[X\cdot Y]-\frac{1}{n}\cdot E[X\cdot Y]-\frac{n-1}{n}\cdot E[X]\cdot E[Y]-\frac{1}{n}\cdot E[X\cdot Y]-\\-\frac{n-1}{n}\cdot E[X]\cdot E[Y]+\frac{1}{n}\cdot E[X\cdot Y]+\frac{n-1}{n}\cdot E[X]\cdot E[Y]\end{array}\right)=$$

$$=\frac{1}{n-1}\cdot\sum_{0\le i<n}\left(\frac{n-1}{n}\cdot E[X\cdot Y]-\frac{n-1}{n}\cdot E[X]\cdot E[Y]\right)=$$

$$=\frac{1}{n}\cdot\sum_{0\le i<n}(E[X\cdot Y]-E[X]\cdot E[Y])=E[X\cdot Y]-E[X]\cdot E[Y]=COV[X,Y]$$

Insbesondere ist damit auch die Gültigkeit von (2.30) gezeigt.

Die empirische Korrelation zweier Zeitreihen erhält man durch Normierung ihrer Kovarianz mit der empirischen Varianz:

$$\overline{COR}[X,Y]=\frac{\overline{COV}[X,Y]}{\sqrt{S^2[X]\cdot S^2[Y]}}$$

5.4 Autokovarianz und Autokorrelation einer Zeitreihe (LK Gl. 4.9)

Wir betrachten zwei Zeitreihen:
X_i : 1, 2, 2, 3, 4, 4, 4, 5, 5, 4, 3, 3, 3, 2, 2, 3, 3, 2, 2, 1, 1, 1, 2, 2, 3
sowie
Y_i : 1, 4, 2, 5, 3, 4, 2, 1, 5, 2, 4, 1, 4, 2, 5, 4, 1, 3, 5, 4, 1, 3, 4, 1, 5
Bei Zeitreihe X_i sind sich aufeinanderfolgende Zahlen ähnlich, während sich bei Y_i aufeinanderfolgende Zahlen eher unähnlich sind. Die Autokorrelation ist ein Maß für die Ähnlichkeit von innerhalb einer Zeitreihe aufeinanderfolgender Zufallsvariablen. Der Abstand der betrachteten Zufallsvariablen kann vorgegeben werden und wird im Englischen als „lag“ bezeichnet.
Die Autokorrelation ist die normierte Autokovarianz.
Die Autokovarianz berechnet sich als die Kovarianz der betrachteten Zeitreihe und der um das lag verschobenen betrachteten Zeitreihe.

Definition:
Die Autokovarianz C_j eines stochastischen Prozesses X_i mit Abstand j ist definiert durch $\overline{COV}[X,Y]$ mit $Y_i=X_{i+j}$.

Berechnung der empirischen Autokovarianz:

$$\hat{C}_j(n) = \frac{1}{(n-j)} \sum_{0\le i<n-j} \left[X_i - \overline{X}\right]\left[X_{i+j} - \overline{X}\right] =$$

$$\frac{1}{n-j} \cdot \left(\sum_{j\le i<n} \left(X_i \cdot X_{i-j}\right) - \overline{X}(n) \cdot \left(2 \cdot \sum_{0\le i<n} X_i - \sum_{0\le i<j} X_i - \sum_{n-j\le i<n} X_i \right) \right) + \left(\overline{X}(n)\right)^2 \qquad (2.34)$$

Es gibt auch eine ähnliche Definition mit einem anderen Vorfaktor, die bessere analytische Eigenschaften hat.

$$\hat{C}_j(n) = \frac{1}{n} \sum_{0\le i<n-j} \left[X_i - \overline{X}\right]\left[X_{i+j} - \overline{X}\right] \qquad (2.35)$$

Eine analoge Definition gibt es auch für den zeitkontinuierlichen Fall:

$$\hat{C}_u(T) = \frac{1}{T} \cdot \int_0^{T-u} \left(X(t+u) - \overline{X}\right) \cdot \left(X(t) - \overline{X}\right) dt \qquad (2.36)$$

Keiner der Schätzer 2.34, 2.35 und 2.36 ist erwartungstreu (LK p.253). Die Schätzergebnisse werden schlechter, je größer der betrachtete Abstand ist. Außerdem haben sie eine hohe Varianz v.a. bei kurzen Messreihen und sind untereinander korreliert, d.h. $\mathrm{COV}\left(\hat{C}_j, \hat{C}_k\right) \neq 0$. Für unsere Zwecke nutzen wir den Schätzer 2.34.

5.5 Empirische Autokorrelation einer Zeitreihe (LK Gl. 4.9)

Die Autokovarianz wird durch Normierung zur Autokorrelation:

$$\rho_j(n) = \frac{\hat{C}_j(n)}{S^2(n)} \tag{2.37}$$

Auch der Schätzer für die Autokorrelation ist nicht gut, insbesondere für große Abstände j. Siehe Beispiel in Kapitel 5.6!

5.6 Erfassungsobjekte für Autokovarianzen und Autokorrelationen von zeitdiskreten Messwerten

- Erweiterung des „discrete counters“
- Spezifikation des max. Abstandes j für die Autokovarianz bzw. Autokorrelation
- Interne Datenhaltung
  - j Variablen um die ersten Werte $X_0, \ldots, X_{j-1}$ zu speichern
  - j Variablen um die letzten Werte zu $X_{n-j}, \ldots, X_{n-1}$ zu speichern
  - 1 Zähler für die Summe $X_\Sigma = \sum_{0 \le i < n} X_i$
  - j+1 Zähler für die Summen $X^2_\Sigma(k) = \sum_{k \le i < n} X_{i-k} \cdot X_i$ für $0 \le k \le j$
- Berechnung der Autokorrelations- und Autokovarianzfunktion gemäß Gleichung (2.34) und (2.37).
- Achtung: die Werte sind nur gut für j<<n! Beispiele:
  - X=0, 1, 1, 2 ⇒ $\rho_0(4)$=0.75, $\rho_1(4)$=0.00, $\rho_2(4)$=0.00, $\rho_3(4)$=-1.50
  - X=0, 1, 2, 3, 4, 5, 6, 7, 8, 9 ⇒ $\rho_0(10)$=0.90, $\rho_1(10)$=0.70, $\rho_2(10)$=0.46, $\rho_3(10)$=0.19, $\rho_4(10)$=-0.12, $\rho_5(10)$=-0.46, $\rho_6(10)$=-0.85, $\rho_7(10)$=-1.26, $\rho_8(10)$=-1.72, $\rho_9(10)$=-2.21

Betragswerte größer 1 sind offensichtlich keine gültigen Korrelationswerte.

5.7 Wichtige Begriffe auf Deutsch und Englisch

Deutscher Begriff	Englischer Begriff
Verteilung	distribution
Verteilungsfunktion	distribution function cumulative distribution function (cdf)
Verteilungsdichtefunktion	probability density function (pdf)
p-Quantil	p-percentile, p-quantile
Faltung	convolution
Erwartungswert	mean
Moment	moment
Zentrales Moment	central moment
Varianz	variance
Standardabweichung	standard deviation
Variationskoeffizient	coefficient of variation
Schiefe	skewness
Kovarianz	covariance
Korrelation	correlation
Empirisches Moment Stichprobenmoment	sample moment
Empirischer Mittelwert Stichprobenmittelwert	sample mean
Empirische Varianz Stichprobenvarianz	sample variance
Empirische Schiefe Stichprobenschiefe	sample skewness
Empirische Kovarianz Stichprobenkovarianz	sample covariance
Empirische Autokovarianz Stichprobenautokovarianz	sample autocovariance
Empirische Autokorrelation Stichprobenautokorrelation	sample autocorrelation