INTELLIGENTE DATENANALYSE IN MATLAB

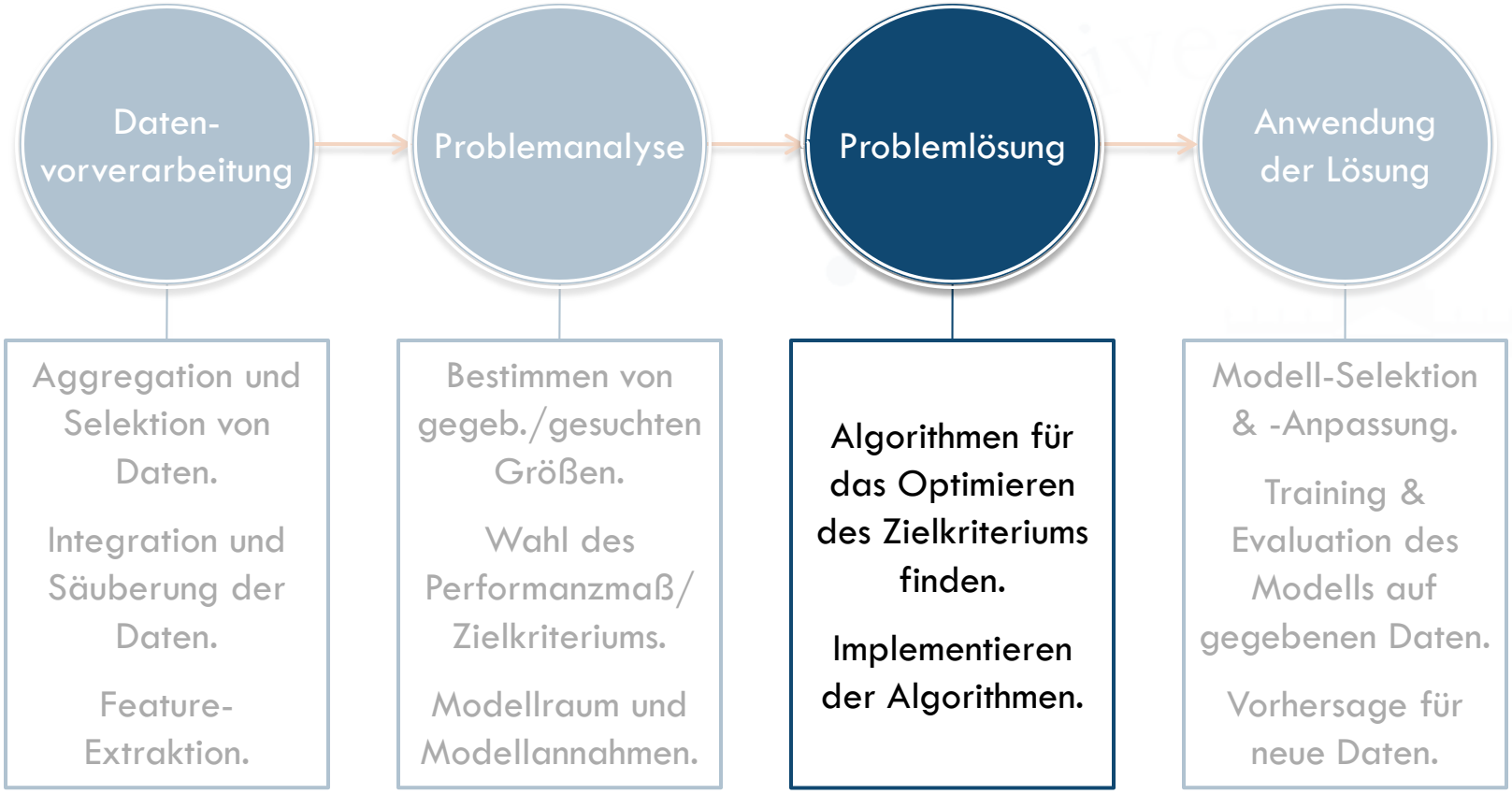
Sequenzanalyse

Literatur

- Klaus Neusser: Zeitreihenanalyse in den Wirtschafts-wissenschaften.
- Mike Hüftle: Stochastische Prozesse in der Zeitreihenanalyse.
  http://www.ivh.uni-hannover.de/optiv/Methoden/StochMet/StochMet.pdf

Überblick

- Schritte der Datenanalyse:



Überblick

- Lernen aus einer Sequenz.
  - Stochastischer Prozess.
  - Datentransformation.
  - Parameter eines stationären stochastischen Prozesses schätzen.
  - Prognose.
- Lernen aus mehreren Sequenzen.
  - Sequenz-Kernel.

Lernen aus einer Sequenz

Problemstellung

- Gegeben: Eine Sequenz $\{x_t \in \mathbb{R} : t = 1...n\}$ von *n* geordneten Datenpunkten (Beobachtungen, Messpunkte).
- Gesucht: Stochastisches Modell welches Sequenz gut erklärt.
- Ziele:
  - Beschreibung: Untersuchen der Charakteristika der Sequenz über Kennzahlen und Diagramme.
  - Modellierung: Verstehen des zugrunde liegenden, Daten-generierenden Modells.
  - Prognose: Vorhersage zukünftiger Werte aufgrund des gelernten Modells.

Lernen aus einer Sequenz

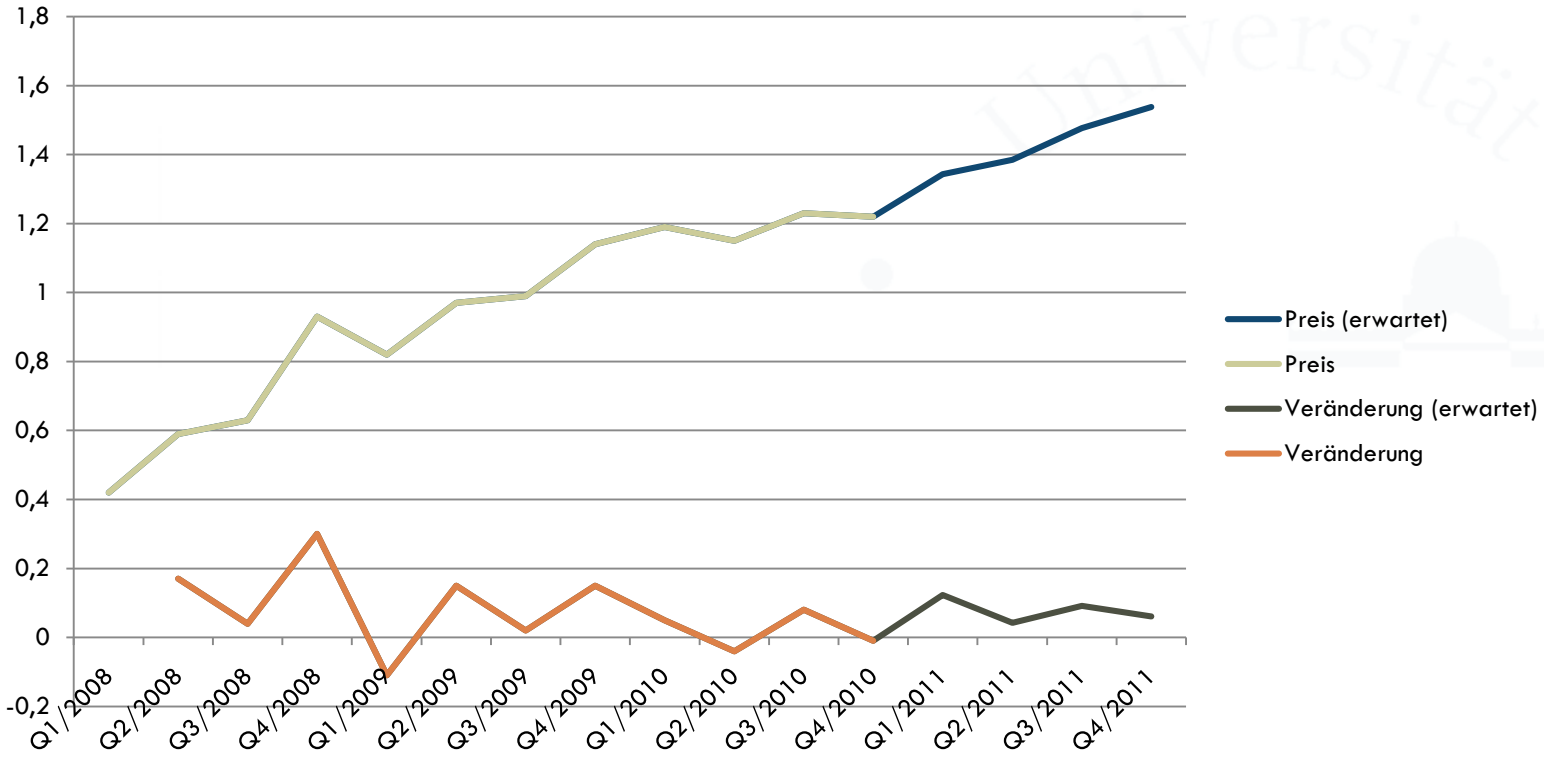
Gegeben Daten

- Sequenzen sind durch Paare (x_i, t_i) gegeben:
  - $x_i \in \mathbb{R}^m$ gibt den Wert der Sequenz an der Position t_i an.
  - Äquidistante Folge (gleiche Abstände $t_{i+1} - t_i$) $\{x_t : t = 1...n\}$
    $\Rightarrow$ diskreter stationärer Prozess.
  - Nicht-äquidistante Folge $\{x(t_i) : i = 1...n\}$
    $\Rightarrow$ stetiger stationärer Prozess.
- Betrachten im Weiteren nur univariate, äquidistante Zeitreihen, d.h. $\{x_t \in \mathbb{R} : t = 1...n\}$ wie beispielweise
  - Tägliche Messung der Temperatur an einem Ort,
  - Entwicklung der Einwohnerzahl eines Landes,
  - Aktienkurse an aufeinanderfolgenden Börsentagen.

Lernen aus einer Sequenz

Beispiel

- Vorhersage der Preisentwicklung eines Produktes:



Stochastischer Prozess

Definition

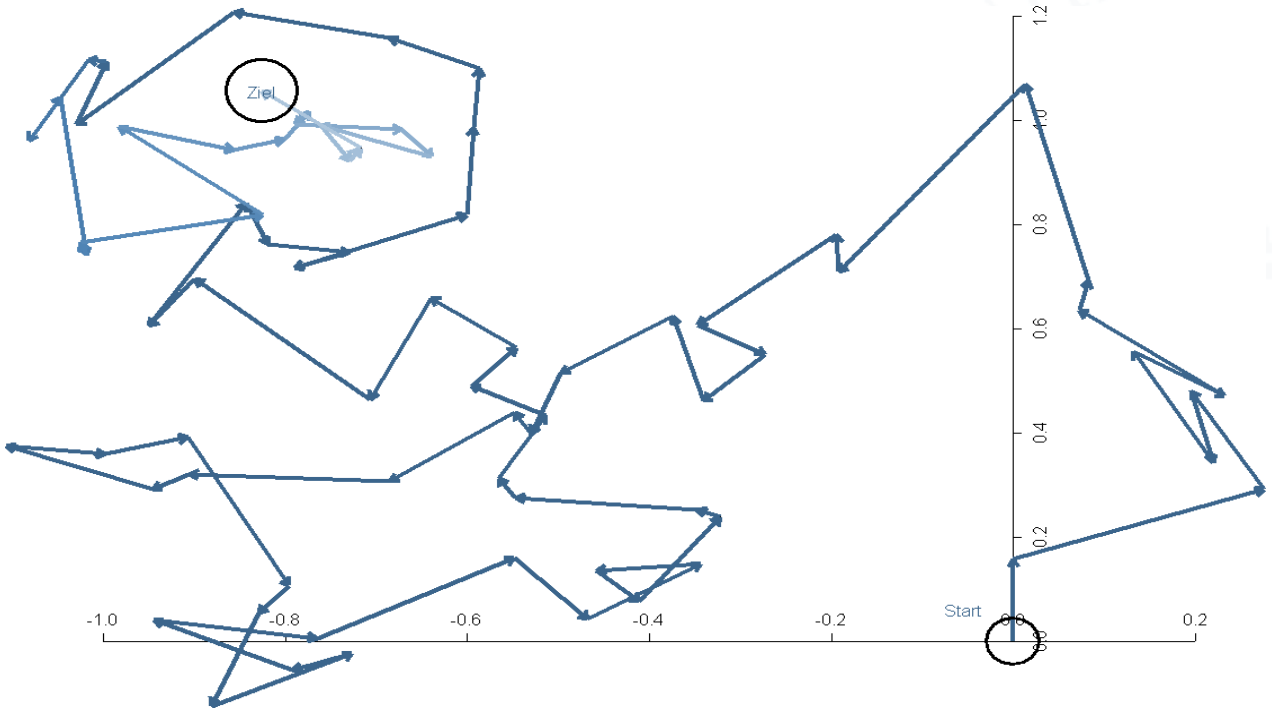
- Stochastischer Prozess:
  - Abbildung $X(\omega,t)$ aus $\Omega \times T$ auf Menge der reellen Zahlen die für jedes fixierte $t \in T$ eine Zufallsgröße X_t und für jedes fixierte $\omega \in \Omega$ eine gewöhnliche Funktion $x(t)$ darstellt.
  - Jede Zufallsgröße X_t nimmt für einen Versuchsausgang einen Wert (Realisierung) x_t an.

- Annahme: Gegebene Sequenz $\{x_t \in \mathbb{R} : t = 1...n\}$ ist Folge von Realisierungen eines stochastischen Prozesses.

Stochastischer Prozess

Beispiel

- Brown‘sche Bewegung:
  - Beschreibt zufällige Bewegung von Teilchen.



Stochastischer Prozess

Kenngrößen

- Mittelwertfunktion:
  - Erwartungswert $\mu_X(t) = \mathrm{E}[X_t]$ der Zufallsgröße X_t zum Zeitpunkt *t*.
  - Durchschnittliche Sequenz, um welche die tatsächlichen Beobachtungen des stochastischen Prozesses schwanken.
- Varianzfunktion:
  - Varianz $\sigma_X^2(t) = \mathrm{E}[(X_t - \mu_X(t))^2]$ der Zufallsgröße X_t zum Zeitpunkt *t*.
  - Mittlere quadratische Abweichung vom erwarteten Verlauf des stochastischen Prozesses.

Stochastischer Prozess

Kenngrößen

- Autokovarianzfunktion:
  - Autokovarianz $\gamma_X(s,t) = \mathrm{E}[(X_s - \mu_X(s))(X_t - \mu_X(t))]$ der Zufallsgrößen X_s und X_t zu den Zeitpunkten $s,t \in T$.
- Autokorrelationsfunktion:
  - Autokorrelation $\rho_X(s,t)$ der Zufallsgrößen X_s und X_t zu den Zeitpunkten $s,t \in T$ mit

$$\rho_X(s,t) = \frac{\gamma_X(s,t)}{\sigma_X(s)\sigma_X(t)}.$$

  - Gibt an ob die Zufallsgrößen eines stochastischer Prozesses linear korreliert sind, d.h. ob der Prozess zyklische Muster beschreibt.

Stochastischer Prozess

Eigenschaften

Ein stochastischer Prozess heißt

- *stationär* falls für $\forall r,s,t \in T$
  - Erwartungswert $\mu_X(t) = \mu_X$ konstant,
  - Varianz $\sigma_X^2(t) < \infty$,
  - Autokovarianz $\gamma_X(s,t) = \gamma_X(s+r,t+r)$, d.h. die Autokovarianzfunktion hängt nur vom zeitlichen Abstand $h = |s-t|$ ab; somit gilt $\gamma_X(s,t) = \gamma_X(t,s) = \gamma_X(h,0) = \gamma_X(-h,0)$.
- *diskret* falls die Zeitpunkte $t \in T$ abzählbar sind.
- *stetig* falls T ein Intervall der reellen Zahlen ist.

Stochastischer Prozess

Schätzer der Kenngrößen für stationäre Prozesse

- Schätzer für Mittelwertfunktion:

$$\hat{\mu}_X(t) = \frac{1}{t}\sum_{i=1}^{t} x_i$$

- Schätzer für Varianzfunktion:

$$\hat{\sigma}_X^2(t) = \frac{1}{t}\sum_{i=1}^{t}(x_i - \hat{\mu}_X(t))^2$$

- Schätzer für Autokovarianzfunktion:

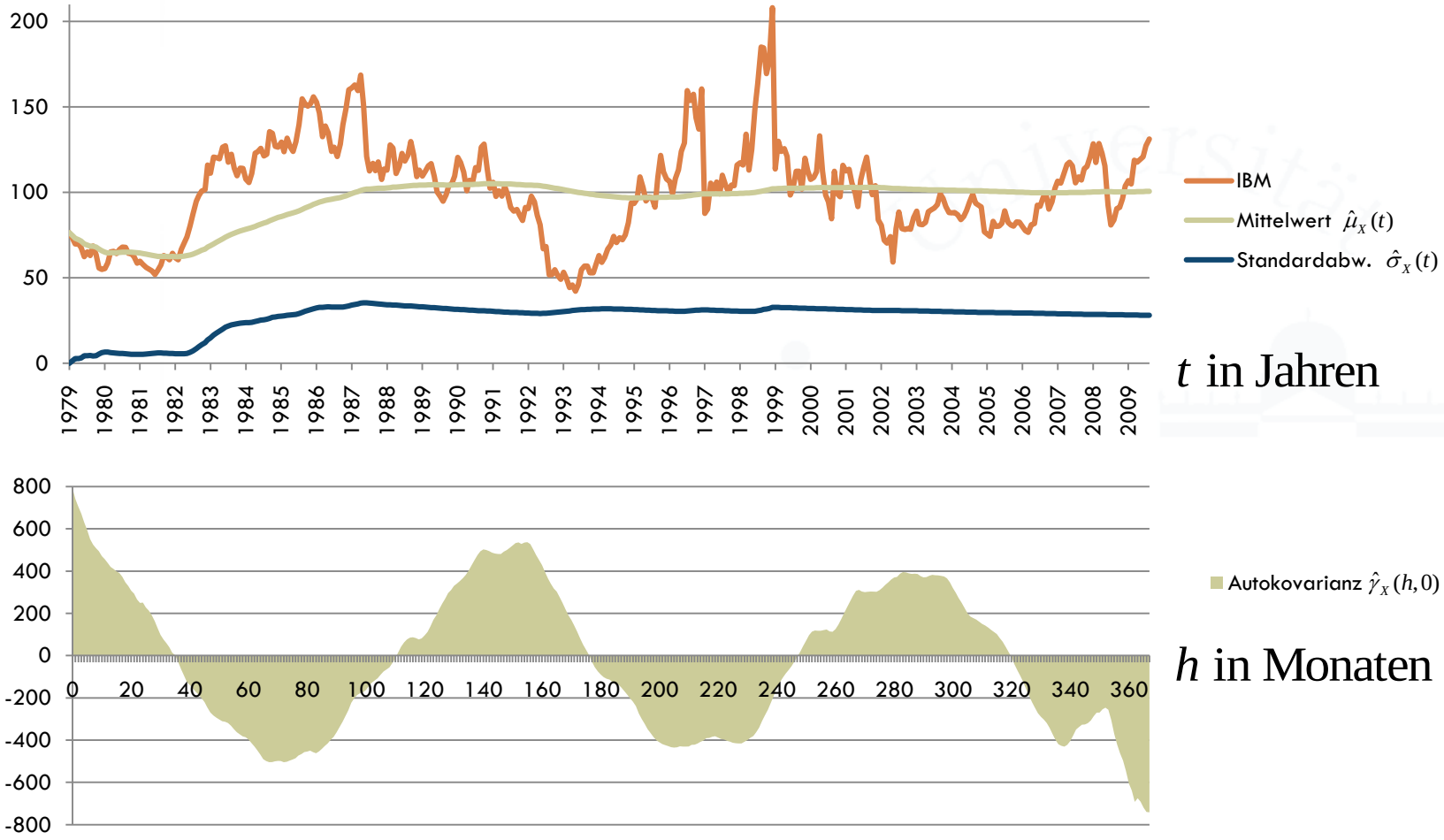
$$\hat{\gamma}_X(h,0) = \frac{1}{n}\sum_{i=1}^{n-h}(x_i - \hat{\mu}_X(n))(x_{i+h} - \hat{\mu}_X(n))$$

- Schätzer für Autokorrelationsfunktion

$$\hat{\rho}_X(h,0) = \frac{\hat{\gamma}_X(h,0)}{\hat{\gamma}_X(0,0)}$$

Stochastischer Prozess

Beispiel: Schätzer für Kenngrößen



Stochastischer Prozess

Lernen des Prozesses

- Stochastischer Prozess durch Mittelwert-, Varianz- und Autokovarianzfunktion hinreichend genau gegeben.
- Schätzung dieser Funktionen für nicht-stationäre Prozesse benötigt jedoch sehr viele Daten!

- Ansatz:
  1. Daten transformieren um Stationarität sicherzustellen (z.B. Mittelwert=Null sicherstellen).
  2. Stationären Prozess wählen und dessen Parameter schätzen.
  3. Gelerntes Modell anwenden (Prognose) und ursprüngliche Datentransformation umkehren.

Datentransformation

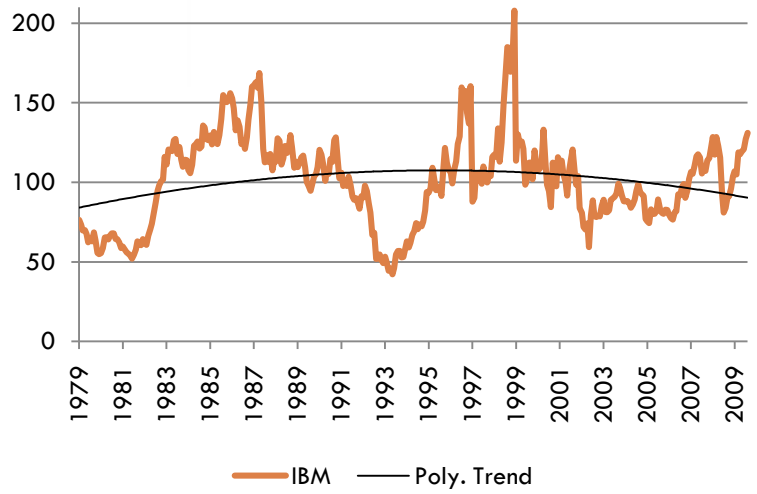
Trendbereinigung

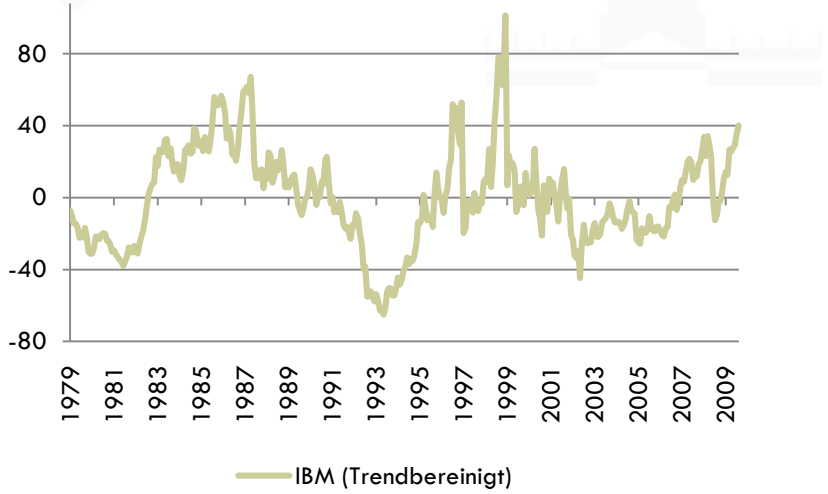
- In der Praxis Mittelwert oft nicht konstant über die Zeit.
- Modellierung des Mittelwerts als Funktion der Zeit, z.B.
  - Linearer Trend: $\hat{\mu}_X(t) = \beta_0 + \beta_1 t$
  - Polynomieller Trend: $\hat{\mu}_X(t) = \sum_{i=0}^{p} \beta_i t^i$
  - Exponentieller Trend: $\hat{\mu}_X(t) = \sum_{i=0}^{p} e^{\beta_i t}$
  - Potenzieller Trend: $\hat{\mu}_X(t) = \sum_{i=0}^{p} t^{\beta_i}$

Datentransformation

Trendbereinigung

- Schätzung der Parameter β_i bspw. durch Minimierung des quadratischen Fehlers $\min_{\beta_i} \sum_{t=1}^{n} (\hat{\mu}_X(t) - x_t)^2$.
- Bereinigte Sequenz $x'_t = x_t - \hat{\mu}_X(t), t = 1...n$ ist (approximativ) stationär bzgl. des Mittelwertes mit $\hat{\mu}'_X = 0$.





Datentransformation

Filter

- Entfernung nicht-stationärer Komponenten durch
  - Anwenden des Differenzenfilters:

$$x'_t = \Delta x_t = x_{t+1} - x_t$$

Entfernt lineare Trends

  - *d*-maliges Anwenden des Differenzenfilters:

$$x'_t = \Delta^d x_t = \Delta^{d-1} x_{t+1} - \Delta^{d-1} x_t$$

Entfernt poly. Trends

  - Anwenden des Differenzenfilters mit Fenstergröße Δt:

$$x'_t = \Delta_{\Delta t} x_t = x_{t+\Delta t} - x_t$$

Entfernt saisonale Trends

  - Glättung (gleitender Durschnitt mit Fenstergröße q):

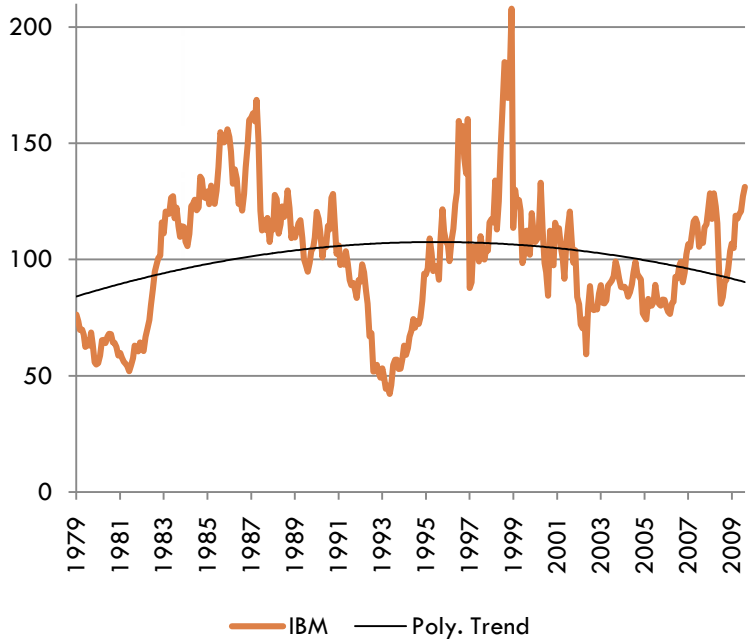
$$x'_t = \frac{1}{2q+1} \sum_{i=-q}^{q} x_{t+i}$$

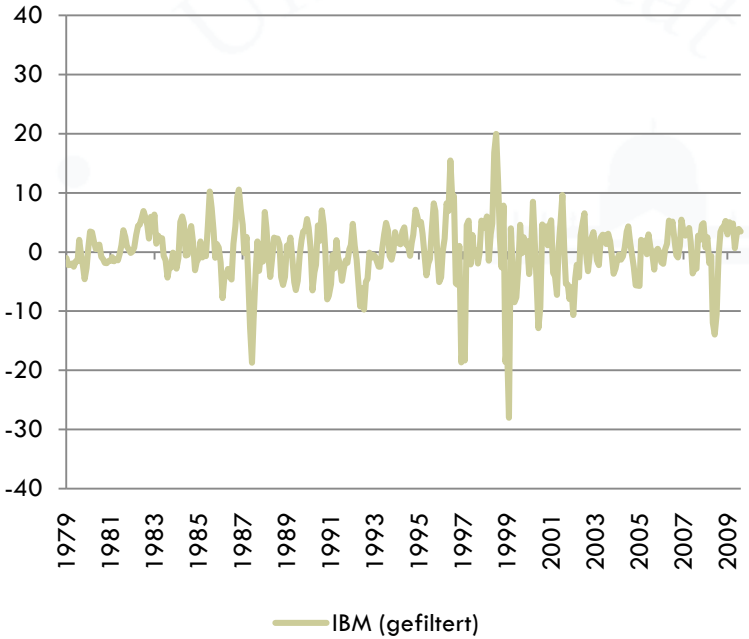
Entfernt lokale Trends

Datentransformation

Filter

- Beispiel: (Einmaliges) Anwenden des Differenzenfilters und Glättung mit $q = 1$.





Stationärer stochastischer Prozess

Beispiel für stationäre Prozesse

- Starke stationäre Prozesse:
  - White-Noise-Prozess (Rauschen).
- Lineare stationäre Prozesse:
  - Moving-Average-Prozess q-ter Ordnung (MA_q).
- Autoregressive Prozesse:
  - Autoregressiver Prozesse p-ter Ordnung (AR_p).
  - Autoregressiver Moving-Average-Prozess ($ARMA_{p,q}$).
- Zählprozesse.
- Poisson-Prozesse.
- (Hidden) Markov-Prozesse & Markov-Ketten.
- Wiener-Prozesse.

Starke stationäre Prozesse

White-Noise-Prozess

- Annahmen:
  - Aufenthaltsort $x_t = \varepsilon_t$ zum Zeitpunkt *t* ist unabhängig vom Aufenthaltsort zu einem vorherigen Zeitpunkt.
- Eigenschaften:
  - Mittelwertfunktion: $\mu_X(t) = 0$
  - Varianzfunktion: $\sigma_X^2(t) = \sigma_X^2$
  - Autokovarianzfunktion: $\gamma_X(h,0) = 0$
- Beispiel:
  - Unabhängig normalverteilte Beobachtungen: $X_t \sim N(0, \sigma_X^2)$

Lineare stationäre Prozesse

Moving-Average-Prozess

- Beispiel: Ein Eisverkäufer möchte Verkaufszahlen von Eiscreme vorhersagen.
  - Nur bisherige Verkaufszahlen x_t gegeben.
- Annahme: Kunden essen nur Eis wenn in den letzten Tagen die Sonne oft zu sehen war.
- Modell:
  - (Unbeobachtete) Rauschwerte ε_t sind die Sonnenstunden pro Tag normiert auf Mittelwert 0.
  - Beispiel für Modell mit Modellparametern:

$$x_t = \varepsilon_t + 0{,}8 \cdot \varepsilon_{t-1} + 0{,}5 \cdot \varepsilon_{t-2} + 0{,}2 \cdot \varepsilon_{t-3}$$

Lineare stationäre Prozesse

Moving-Average-Prozess

- Annahmen:
  - Aufenthaltsort x zum Zeitpunkt t ist das gewichtete Mittel aus $q+1$ Rauschwerten zu den Zeitpunkten $t-q,\ldots,t$:

$$x_t = \varepsilon_t + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j}$$

    mit Rauschwerten ε_i wobei $\mathrm{E}[\varepsilon_i] = 0$ und $\mathrm{Var}[\varepsilon_i] = \sigma_\varepsilon^2$.
- Ziel: Berechnung der Gewichte α_i.

Lineare stationäre Prozesse

Moving-Average-Prozess

- Berechnung der Gewichte α_i am Beispiel eines MA_1-Prozesses: $x_t = \varepsilon_t + \alpha_1 \varepsilon_{t-1}$
- Umformung des Prozesses ergibt:

$$\begin{aligned} \varepsilon_j &= x_j - \alpha_1 \varepsilon_{j-1} \\ \varepsilon_{j-1} &= x_{j-1} - \alpha_1 \varepsilon_{j-2} \\ &\vdots \\ \varepsilon_1 &= x_1 \end{aligned} \Rightarrow \begin{aligned} \varepsilon_j &= x_j - \alpha_1 \varepsilon_{j-1} \\ &= x_j - \alpha_1 (x_{j-1} - \alpha_1 \varepsilon_{j-2}) \\ &= x_j + (-\alpha_1)^1 x_{j-1} + (-\alpha_1)^2 \varepsilon_{j-2} \\ &\vdots \end{aligned} \Rightarrow \varepsilon_j(\boldsymbol{\alpha}) = \sum_{i=0}^{j-1} (-\alpha_1)^i x_{j-i}$$

- Minimieren des mittleren quadratischen Fehlers zwischen Modell und Sequenz mit $\mathrm{E}[\varepsilon_t] = \mu_\varepsilon = 0$:

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} \sum_{j=1}^{n} (\varepsilon_j(\boldsymbol{\alpha}) - \mu_\varepsilon)^2 = \arg\min_{\boldsymbol{\alpha}} \sum_{j=1}^{n} \varepsilon_j(\boldsymbol{\alpha})^2$$

Autoregressive Prozesse

Einfacher Autoregressiver Prozess

- Beispiel: Ein Eisverkäufer möchte Verkaufszahlen von Eiscreme vorhersagen.
  - Nur bisherige Verkaufszahlen x_t gegeben.
- Annahme: Kunden essen nur Eis wenn sie in den letzten Tagen wenig Eis gegessen haben.
- Modell:
  - Verkaufszahl x_t ist nur abhängig von den vorherigen Verkaufszahlen und einem Rauschwert ε_t (Laune der Kunden).
  - Beispiel für Modell mit Modellparametern:

$$x_t = \varepsilon_t - 0,7 \cdot x_{t-1} - 0,6 \cdot x_{t-2} + 0,1 \cdot x_{t-3}$$

Autoregressive Prozesse

Einfacher Autoregressiver Prozess

- Annahmen:
  - Aufenthaltsort *x* zum Zeitpunkt *t* ist die Summe aus dem Rauschwert zum Zeitpunkt *t* und dem gewichteten Mittel der *p* Aufenthaltsorte zu den Zeitpunkten $t-p,\ldots,t-1$:

$$x_t = \varepsilon_t + \sum_{j=1}^{p} \beta_j x_{t-j}$$

  mit Rauschwerten ε_t wobei $\mathrm{E}[\varepsilon_t] = 0$, $\mathrm{Var}[\varepsilon_t] = \sigma_\varepsilon^2$ und $|\beta_j| \neq 1$.
- Ziel: Berechnung der Gewichte β_j.

Autoregressive Prozesse

Einfacher Autoregressiver Prozess

- Berechnung der Gewichte β_i eines AR_p-Prozesses:

$$x_t = \varepsilon_t + \sum_{j=1}^{p} \beta_j x_{t-j}$$

- Idee: Minimieren des mittleren quadratischen Fehlers.

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \frac{1}{n-p-1} \sum_{i=p+1}^{n} \left(x_i - \varepsilon_i - \sum_{j=1}^{p} \beta_j x_{i-j} \right)^2 = \arg\min_{\boldsymbol{\beta}} \sum_{i=p+1}^{n} \left(x_i - \sum_{j=1}^{p} \beta_j x_{i-j} \right)^2$$

Erwartungswert Null & konstante Varianz

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \left(\begin{bmatrix} x_{p+1} \\ x_{p+2} \\ \vdots \\ x_n \end{bmatrix} - \begin{bmatrix} x_p & x_{p-1} & \cdots & x_1 \\ x_{p+1} & x_p & \cdots & x_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{n-1} & x_{n-2} & \cdots & x_{n-p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \right)^2 \Rightarrow \boldsymbol{\beta}^* = \mathbf{A}^+ \mathbf{y}$$

$\mathbf{y}$

$\mathbf{A}$

Pseudo-Inverse von $\mathbf{A}$

Autoregressive Prozesse

Autoregressiver Moving-Average-Prozess

- Motivation:
  - Fast jeder endliche Datensatz gut an MA- oder AR-Modell mit hoher Ordnung anpassbar.
  - Je größer die Ordnung, desto mehr Parameter.
  - Ziel: Modell mit möglichst wenigen Parametern welches zudem reale Bedeutung (Interpretation) hat.
- Idee: Kombination von MA- und AR-Modellen zu $ARMA_{p,q}$-Prozess.

$$x_t = \varepsilon_t + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j} + \sum_{j=1}^{p} \beta_j x_{t-j}$$

Autoregressive Prozesse

Autoregressiver Moving-Average-Prozess

- Beispiel: Vorhersage der Werbeausgaben zweier konkurrierender Unternehmen.
  - Nur bisherige Werbeausgaben x_t und y_t gegeben.
- Modell:
  - Werbeausgabe x_t ist abhängig von der vorherigen Werbeausgabe des Konkurrenten y_{t-1} und einem Rauschwert.
  - Werbeausgabe y_t ist abhängig von der vorherigen Werbeausgabe des Konkurrenten x_{t-1} und einem Rauschwert.

$$\begin{aligned} x_t &= \varepsilon_t + \beta_1 y_{t-1} \\ y_t &= \varepsilon'_t + \beta'_1 x_{t-1} \end{aligned} \quad \Rightarrow \quad x_t = \varepsilon_t + \beta_1(\varepsilon'_{t-1} + \beta'_1 x_{t-2}) = \varepsilon_t + \alpha_1 \varepsilon_{t-1} + \beta_2 x_{t-2}$$

$ARMA_{2,1}$-Prozess

Autoregressive Prozesse

Autoregressiver Moving-Average-Prozess

- Überlagerungssatz von ARMA-Prozessen:
  - x_t und y_t seien zwei unabhängige ARMA-Prozesse der Ordnung (p_1,q_1) und (p_2,q_2).
  - Summe $z_t = x_t + y_t$ ist wieder ein ARMA-Prozess der Ordnung (p,q).
  - Für AR-Ordung gilt: $p \leq p_1 + p_2$
  - Für MA-Ordung gilt: $q \leq \max(p_1 + q_2, p_2 + q_1)$
- Folgerung:
  - Summe zweier MA-Prozesse ergibt wieder MA-Prozess.
  - Summe zweier AR-Prozesse ergibt ARMA-Prozess.

Autoregressive Prozesse

Autoregressiver Moving-Average-Prozess

- Berechnung der Gewichte α_i und β_i eines ARMA-Prozesses analog zu MA-Prozessen.
- Abhängig von der Datentransformation unterscheidet man verschiedene ARMA-Prozesse, z.B.:
  - ARIMA: Ist x_t nach d-maligem Anwenden des Differenzenfilters ein stationärer $ARMA_{p,q}$-Prozess, so bezeichnet man den Prozess x_t als $ARIMA_{p,d,q}$-Prozess.
  - ARMAX: Nach vorherige Trendbereinigung sind die Residuen (bereinigte Werte x_t') Realisierungen eines ARMA-Prozesses.

Prognose

Autoregressiver Moving-Average-Prozess

- Vorhersage durch gelernten stochastischen Prozess (Prognose für x_{t+h}):
  - x_t, x_{t-1}, … x_1 entsprechen den tatsächlichen Beobachtungen.
  - x_{t+h}, … x_{t+1} werden durch ihre Prognosen ersetzt.
  - Störterme ε_t, ε_{t-1}, … ε_1 entsprechen den Prognosefehlern der 1-Schrittprognosen in der Vergangenheit.
  - Störungen ε_{t+h}, … ε_{t+1} werden durch ihren Erwartungswert Null ersetzt.
- Datentransformation (Trendbereinigung, Filter etc.) für prognostizierte Werte x_{t+h} umkehren.

Lernen aus mehreren Sequenzen

Problemstellung

- Gegeben: Mehrere Sequenzen mit bekannten Zielattributen (gelabelte Daten).
- Gesucht: Modell $f : \mathbf{x} \mapsto y$.
- Ansatz: Verwendung von Kernel-Modellen und Sequenz-Kernel.
  - Quadratische Euklidische Distanz (RBF-Kernel).
  - Dynamic Time Warping (DTW-Kernel).
  - Editierdistanz.

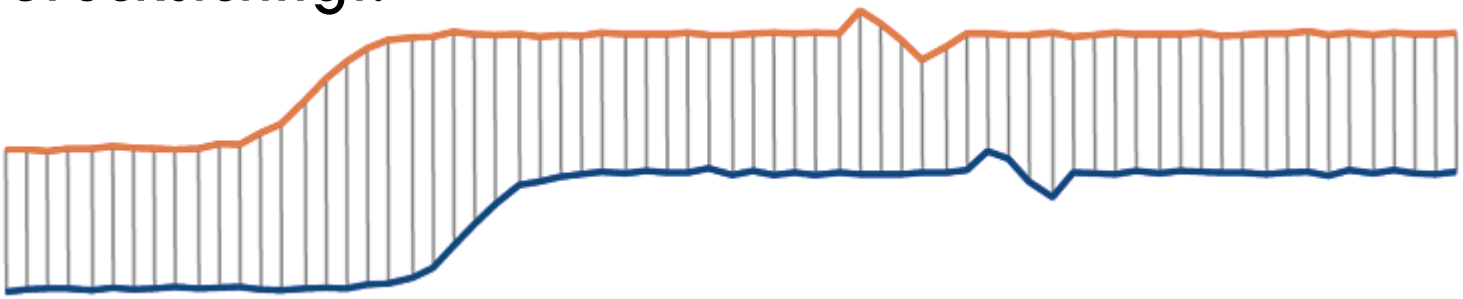
Lernen aus mehreren Sequenzen

Quadratische Euklidische Distanz

- Für numerische Attribute.
- Sequenzen als Vektoren auffassen und quadratischen euklidischen Abstand bestimmen:

$$D_{RBF}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{t}(x_i - y_i)^2 \quad \Rightarrow \quad k_{RBF}(\mathbf{x},\mathbf{y}) = e^{-\lambda D_{RBF}(\mathbf{x},\mathbf{y})}$$

- Verschobene/gedehnte Motive werden nicht berücksichtigt.

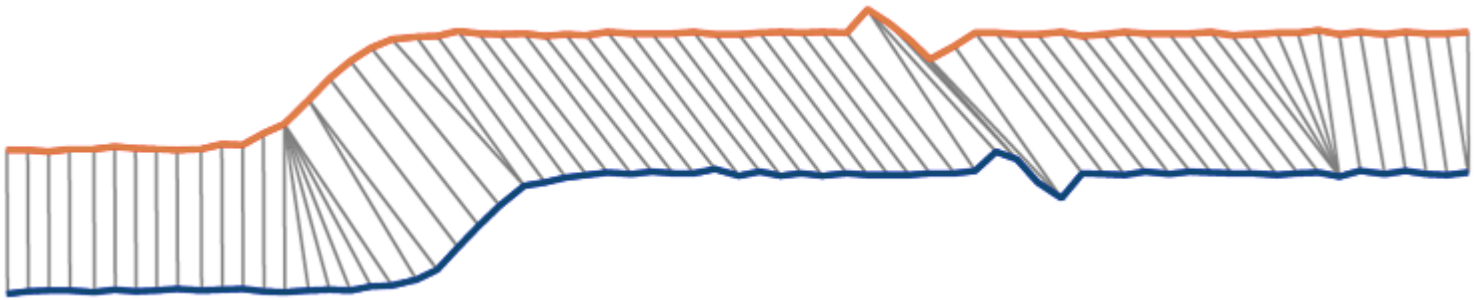


Lernen aus mehreren Sequenzen

Dynamic Time Warping

- Für numerische Attribute.
- Zuordungsfunktionen $\pi_x(i) \in [1, t_x]$ und $\pi_y(i) \in [1, t_y]$.
- DTW-Distanz ist Minimum des verschobenen (quadratischen euklidischen) Abstands:

$$D_{DTW}(\mathbf{x}, \mathbf{y}) = \min_{\pi_x, \pi_y} \sum_{i=1}^{t} (x_{\pi_x(i)} - y_{\pi_y(i)})^2 \Rightarrow k_{DTW}(\mathbf{x}, \mathbf{y}) = e^{-\lambda D_{DTW}(\mathbf{x}, \mathbf{y})}$$



Lernen aus mehreren Sequenzen

Dynamic Time Warping

- Berechnung mit dynamischer Programmierung, rekursive Definition:
  - Sei $\gamma(i, j)$ minimale (verschobene) quadrierte euklidische Distanz bis zu den Zeitpunkten *i* und *j*:

$$\gamma(i, j) = (x_i - y_j)^2 + \min(\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1))$$

- Algorithmus:

DTW(*Sequenzen* **x** *und* **y**)

Setze $\gamma(0,0) = 0,\ \forall i, j\ \gamma(i,0) = \infty,\ \gamma(0, j) = \infty$

FOR i = 1…t_x

FOR j = 1…t_y

$$\gamma(i, j) = (x_i - y_j)^2 + \min(\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1))$$

RETURN $\gamma(t_x, t_y)$

Lernen aus mehreren Sequenzen

Editierdistanz

- Für nominale und ordinale Attribute (Sonderform von Dynamic Time Warping).
- Editierdistanz ist minimale Anzahl an Operationen (Einfügen, Löschen, Ersetzen) um zwei Sequenzen ineinander zu überführen.
- Berechnung mit dynamischer Programmierung, rekursive Definition:
  - Sei $\gamma(i, j)$ minimale Editierdistanz zu den Zeitpunkten *i* und *j*:

$$\gamma(i, j) = \left[x_i \neq y_j \right] + \min(\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1))$$

Zusammenfassung

- Idee: Sequenzen sind Realisierungen eines stochastischen Prozesses (z.B. MA-, AR-, ARMA-Prozess).
- Lernen aus einer Sequenz:
  - Bereinigung der Daten von nicht-stationären Komponenten.
  - Schätzen der Prozess-Parameter.
  - Prognose neuer Werte und Umkehrung der Transformation.
- Lernen aus mehreren Sequenzen:
  - Sequenz-Kernel berechnen (z.B. DTW).
  - Kernel-Modell lernen & anwenden.