

Curs 11:

Analiza seriilor de timp

Structura

- Motivație
- Pre-procesarea seriilor de timp
- Predicție
- Identificare șabloane
- Grupare și clasificare
- Detecție anomalii

Motivație

Problema: Se cunosc date săptămânale privind indexul Dow Jones și se dorește identificarea acțiunilor pentru care creșterea de profit va fi cea mai mare în săptămâna care urmează

Set date: Dow Jones Index (UCI Machine Learning - Brown, Pelosi & Dirska, 2013) - 750 înregistrări, 16 atribute

Exemple de companii cotate și pt care sunt înregistrate informații:

3M	MMM	Cisco Systems	CSCO
American Express	AXP	Coca-Cola	KO
Alcoa	AA	DuPont	DD
AT&T	T	ExxonMobil	XOM
Bank of America	BAC	General Electric	GE
Boeing	BA	Hewlett-Packard	HPQ
Caterpillar	CAT	The Home Depot	HD
Chevron	CVX	Intel	INTC
			IBM

Motivație

Problema: care acțiune va înregistra cea mai mare creștere în săptămâna care urmează?

Exemplu [Dow Jones Index from <http://archive.ics.uci.edu/ml/datasets.html>]

16 attribute

quarter: the yearly quarter (1 = Jan-Mar; 2 = Apr-Jun).

stock: the stock symbol (lista de pe slide-ul anterior)

date: the last business day of the work (de obicei e Vineri)

open: the price of the stock at the beginning of the week

high: the highest price of the stock during the week

low: the lowest price of the stock during the week

close: the price of the stock at the end of the week

volume: the number of shares of stock that traded hands in the week

percent_change_price: the percentage change in price throughout the week

percent_change_volume_over_last_wek: the percentage change in the number of shares of stock that traded hands for this week compared to the previous week

previous_weeks_volume: the number of shares of stock that traded hands in the previous week

Motivație

Problema: care acțiune va înregistra cea mai mare creștere în săptămâna care urmează?

Exemplu [Dow Jones Index from <http://archive.ics.uci.edu/ml/datasets.html>]
16 attribute

next_weeks_open: the opening price of the stock in the following week

next_weeks_close: the closing price of the stock in the following week

percent_change_next_weeks_price: the percentage change in price of the stock in the following week

days_to_next_dividend: the number of days until the next dividend

percent_return_next_dividend: the percentage of return on the next dividend

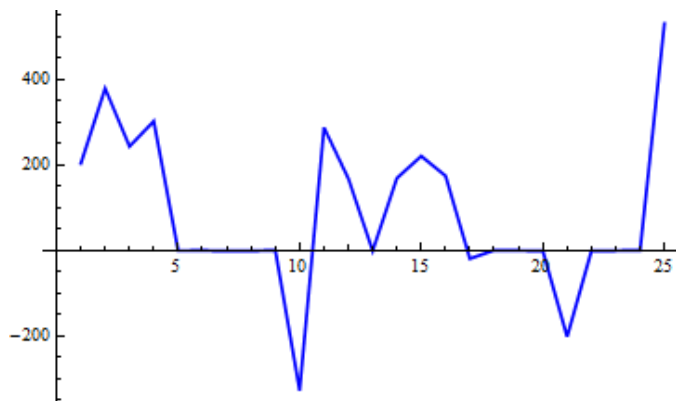
Motivație

Problema: care acțiune va înregistra cea mai mare creștere în săptămâna care urmează?

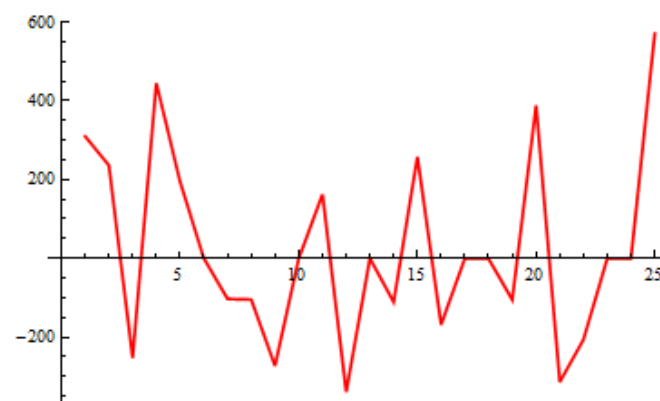
Exemplu [Dow Jones Index de la <http://archive.ics.uci.edu/ml/datasets.html>]

16 attribute

percent_change_next_weeks_price: the percentage change in price of the stock in the following week



IBM



HP

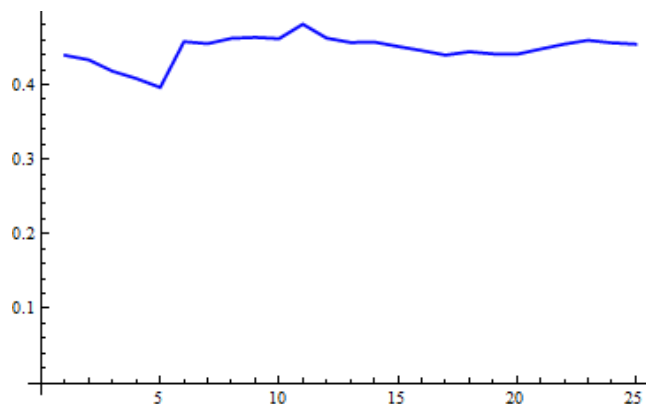
Motivație

Problema: care acțiune va înregistra cea mai mare creștere în săptămâna care urmează?

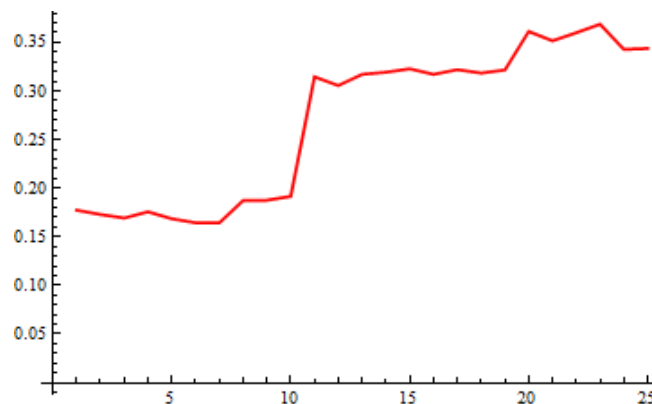
Exemplu [Dow Jones Index de la <http://archive.ics.uci.edu/ml/datasets.html>]

16 attribute

percent_return_next_dividend: the percentage of return on the next dividend



IBM



HP

Motivație

Pe lângă datele financiare există o mulțime de alte surse de serii de timp:

- **Senzori:**

- Date de mediu colectate prin intermediul diferitelor tipuri de senzori (temperatura, presiune, umiditate)

- **Date medicale**

- Electrocardiograma (ECG)
- Electroencefalograma (EEG)
- Date de monitorizare în timp reali a pacienților de la terapie intensivă

- **Date de tip “web log” (clickstream data)**

- Secvențe indicând vizite ale unor pagini web

Motivație

Pe lângă datele financiare există o mulțime de alte surse de serii de timp:

▪Senzori:

- Date de mediu colectate prin intermediul diferitelor tipuri de senzori (temperatura, presiune, umiditate)

Task: **predicție valori viitoare**

▪Date medicale

- Electrocardiograma (ECG)
- Electroencefalograma (EEG)
- Date de monitorizare în timp reali a pacienților de la terapie intensivă

Task: **identificare comportament anormal**

▪Date de tip “web log” (clickstream data)

- Secvențe indicând vizite ale unor pagini web

Task: **identificare tipare de utilizare, profile de utilizatori**

Serii de timp

Exemplu 1 (percentage of return on the next dividend for first 10 weeks included in Dow Jones Index dataset)

0.177, 0.172, 0.169, 0.175, 0.168, 0.164, 0.164, 0.187, 0.187, 0.191

Momentul de timp nu apare ca variabilă explicită. Totuși valorile specificate trebuie interpretate în contextul unor momente de timp.

- Timpul este atribut **contextual**
- Valoarea înregistrată este atribut **comportamental**

Exemplu 2 (temperatura la prânz înregistrată în 7 zile consecutive)

21, 24, 23, 25, 22, 19, 20

Atributul contextual este timpul, cel comportamental este temperatura

Serii de timp

Există diferite tipuri de serii de timp (temporale)

In raport cu domeniul de timp:

- Continue (e.g. EEG)
- Discrete (denumite secvențe)

In raport cu attributele comportamentale

- Univariate (un atribut)
- Multivariate sau vectoriale (mai multe attribute)

Pre-procesarea seriilor de timp

Valori absente

Problema:

- Lipsesc valori corespunzătoare unor momente de timp (de exemplu din cauza unor defecte ale senzorilor)
- În special când sunt mai multe atribute comportamentale (colectate de senzori independenți) ar trebui asigurată sincronizarea între serii, în special prin completarea valorilor absente

Soluție:

- Valoarea absentă este estimată folosind interpolare
- Caz simplu: [interpolare liniară](#)

Pre-procesarea seriilor de timp

Imputarea valorilor absente prin interpolare liniară

Fie (y_1, y_2, \dots, y_n) o serie de timp corespunzătoare momentelor (t_1, t_2, \dots, t_n)

Presupunem ca lipsește valoarea corespunzătoare momentului t cuprins între t_i și t_{i+1} . Presupunând că atributul comportamental y variază liniar cu t pe intervalul $[t_i, t_{i+1}]$ se poate estima valoarea lui y

$$y = y_i + \frac{t - t_i}{t_{i+1} - t_i} (y_{i+1} - y_i)$$

Pre-procesarea seriilor de timp

Eliminarea zgomotului

Problema: dispozitivele utilizate pt colectarea datelor (senzorii) pot fi afectați de bruiaje, a.î. seria poate conține valori generate în procesul de colectare a datelor și care nu reflectă comportamentul real al atributului înregistrat

Modalități de tratare a zgomotului

- Impachetare (Binning)
- Netezire (Moving-Average Smoothing)

Pre-procesarea seriilor de timp

Binning

Idee:

- Intervalul de timp global $[t_1, t_n]$ corespunzător seriei (y_1, y_2, \dots, y_n) este divizat în m subintervale conținând fiecare câte k elemente ($m=n/k$)
- Fiecare subinterval va fi asociat unei valori calculate ca medie a valorilor din seria de timp ce corespunde momentelor incluse în subinterval

Observații:

- Se presupune că momentele de timp corespunzătoare seriei inițiale sunt egal distanțate
- Se reduce numărul de valori disponibile de k ori (este un tip de compresie cu pierdere de informație)

$$(t_1, t_2, \dots, t_n) \rightarrow ((t_1, \dots, t_k), (t_{k+1}, \dots, t_{2k}), \dots, (t_{(m-1)k+1}, \dots, t_{mk}))$$

$$(y_1, y_2, \dots, y_n) \rightarrow (z_1, z_2, \dots, z_m)$$

$$z_i = \frac{1}{k} \sum_{j=1}^k y_{(i-1)k+j}, i = \overline{1, m}$$

Pre-procesarea seriilor de timp

Moving average smoothing

Idee: se reduce pierderea de informație cauzată de binning folosind “ferestre” de mediere care se suprapun, adică media se calculează pentru elementele ce aparțin unei **ferestre mobile** (se deplasează de-a lungul seriei)

$$(t_1, t_2, \dots, t_n) \rightarrow ((t_1, \dots, t_k), (t_2, \dots, t_{k+1}), \dots, (t_{(m-1)k+1}, \dots, t_{mk}))$$

$$(y_1, y_2, \dots, y_n) \rightarrow (z_1, z_2, \dots, z_m)$$

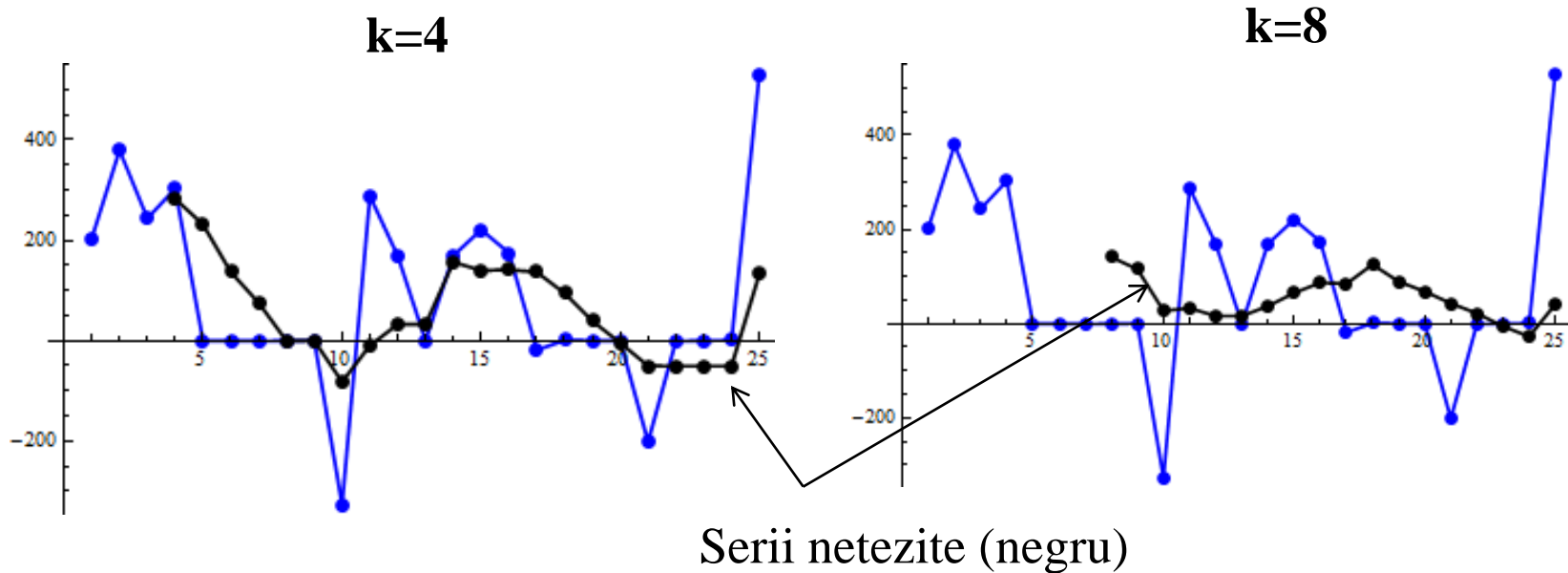
$$z_i = \frac{1}{k} \sum_{j=1}^{(m-1)k+1} y_{i+j-1}, i = \overline{1, m}$$

Obs:

- Numărul de elemente din serie este redus de la n la $n-k+1$
- Variațiile pe termen scurt pot fi pierdute prin mediere
- Mediarea poate fi unidirecțională (se utilizează doar valori anterioare momentului curent) sau bidirecțională/ centrată (se utilizează atât valori anterioare cât și ulterioare)

Pre-procesarea seriilor de timp

Exemplu (Moving average smoothing)



Pre-procesarea seriilor de timp

Netezire exponențială

Idee: valoarea netezită se definește ca o combinație liniară a valorii curente și a valorii netezite anterioare

$$z_i = \alpha \cdot y_i + (1 - \alpha) \cdot z_{i-1}, \quad i = \overline{1, m}$$

$$z_0 = y_1$$

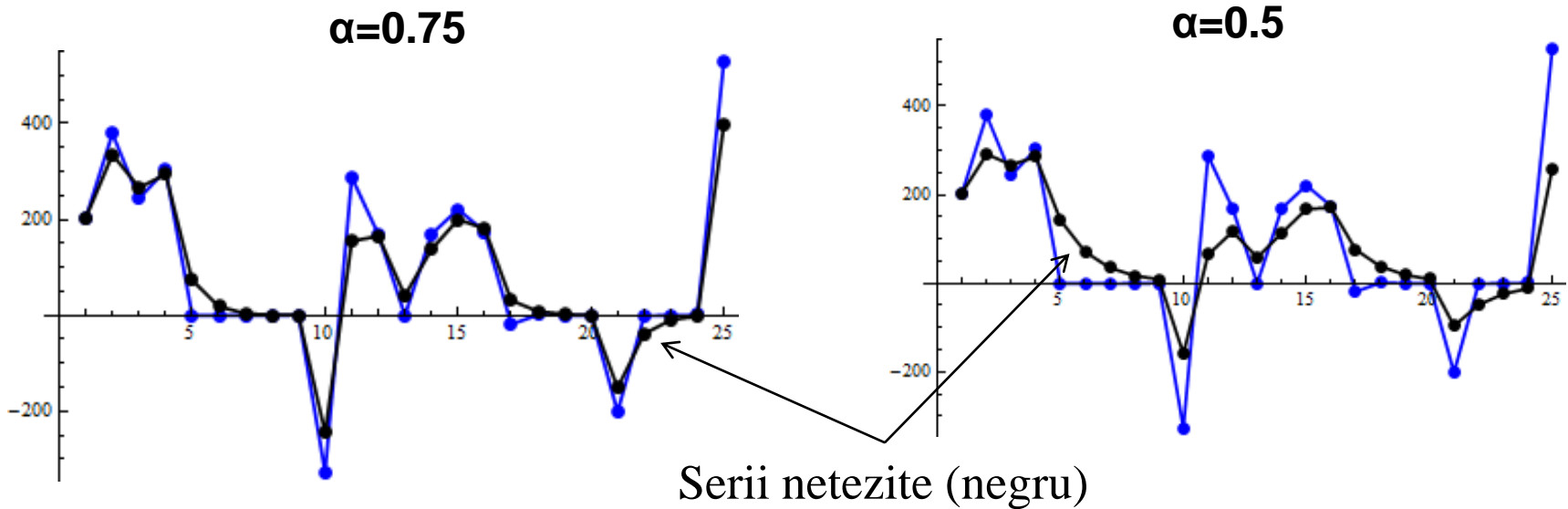
$$z_i = (1 - \alpha)^i z_0 + \alpha \sum_{j=1}^i y_j (1 - \alpha)^{i-j}, \quad i = \overline{1, m}$$

Obs:

- Dacă $\alpha=1$ atunci nu se aplică netezire; dacă $\alpha=1$ toată seria este netezită (va avea valoarea primului element)
- Netezirea exponențială se bazează pe ideea că valorile mai recente sunt mai importante iar cele mai “vechi” au o influență mai mică; influența valorilor anterioare este controlată prin α

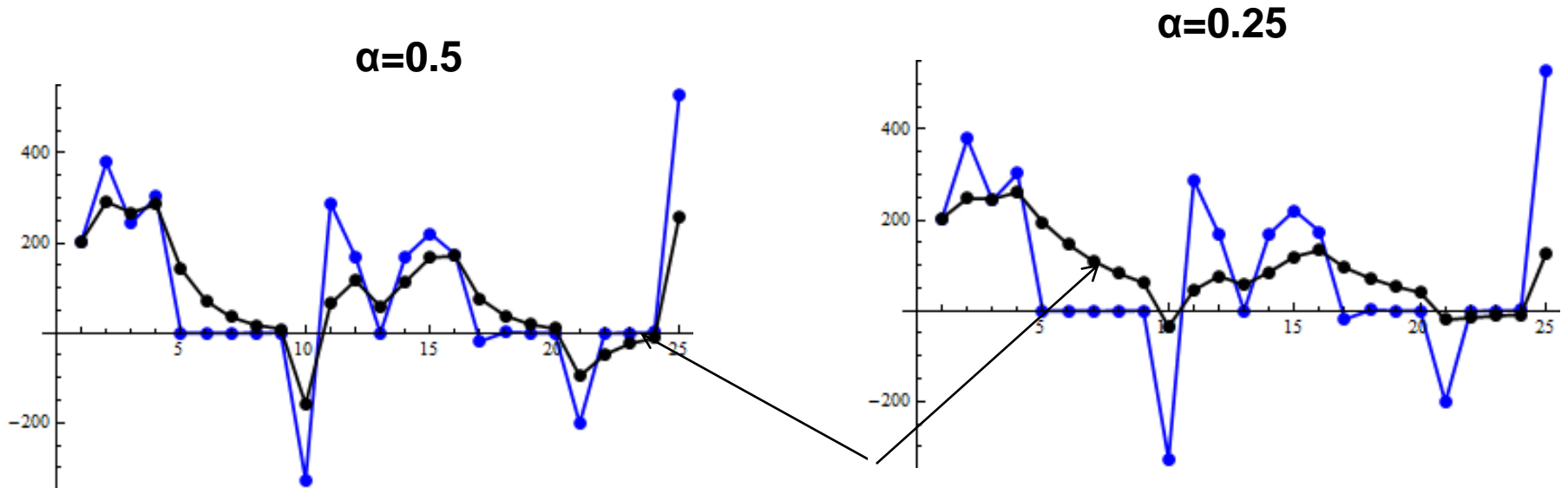
Pre-procesarea seriilor de timp

Exemplu (exponential smoothing)



Pre-procesarea seriilor de timp

Exemplu (exponential smoothing)



Serii netezite (negru)

Pre-procesarea seriilor de timp

Normalizare (scalare) – este utilă în special când se prelucrează mai multe serii de timp)

Variante: Normalizare bazata pe domeniu

$$z_i = \frac{y_i - \min(y)}{\max(y) - \min(y)}, \quad i = \overline{1, m}$$

Standardizare

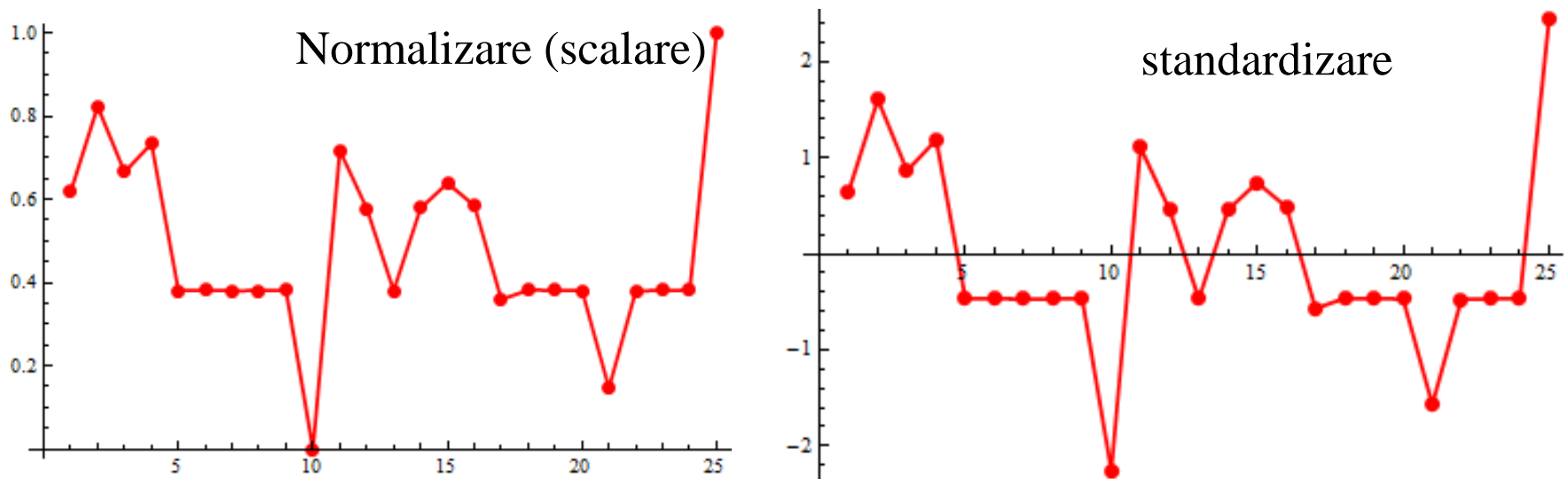
$$z_i = \frac{y_i - \text{mean}(y)}{\text{stdev}(y)}, \quad i = \overline{1, m}$$

Obs:

- $\min(y)$ și $\max(y)$ reprezintă valoarea minimă respectiv cea maximă din serie
- $\text{mean}(y)$ și $\text{stdev}(y)$ sunt valoarea medie respectiv abaterea standard

Pre-procesarea seriilor de timp

Exemplu



Obs:

- Normalizarea (prin scalare) și standardizarea conservă forma seriei dar schimbă domeniul de valori

Predicție

Scop:

- Estimarea prețului viitor al unei acțiuni, predicția vremii, estimarea evoluției unor indicatori economici etc

Predicție:

- **Intrare:** una sau mai multe serii de timp
- **Ieșire:** valori viitoare ale seriei

Cum poate fi abordată problema?

- Ca o problemă de regresie – se estimează explicit dependența dintre atributele comportamentale și timp
- Utilizând modele care exprimă relația dintre valori curente și valori anterioare ale seriei (**modele autoregresive**)

Predicție

Obs: modelele de predicție funcționează bine pentru seriile **staționare**

Intuitiv, o serie staționară se caracterizează prin faptul că proprietățile sale statistice (medie, varianță, autocorelație) sunt constante în timp

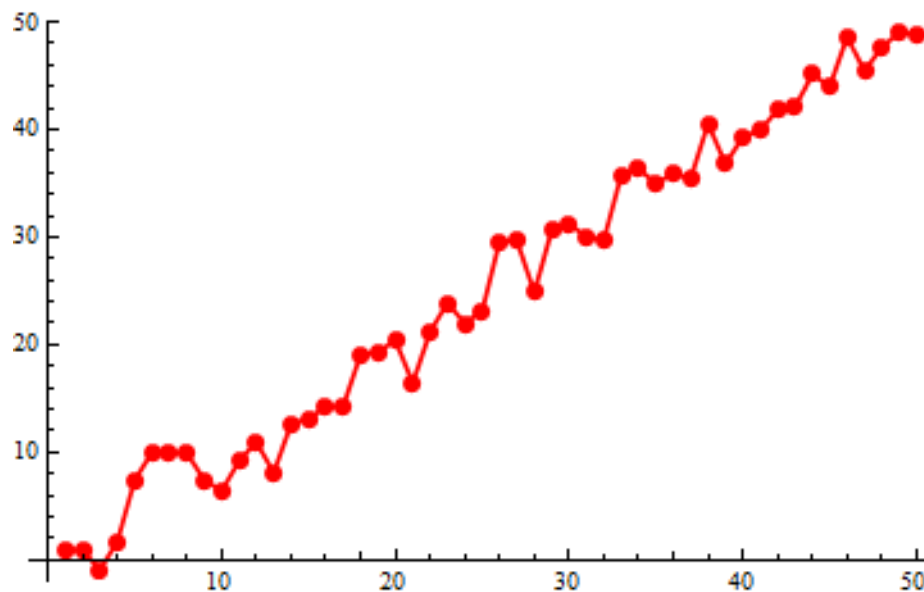
Staționaritate strictă: distribuția de probabilitate a valorilor din orice interval de timp $[a,b]$ este identică cu distribuția de probabilitate a valorilor din intervalul shiftat $[a+h, b+h]$ (pentru un $h>0$ arbitrar)

Obs:

- Proprietățile statistice bazate pe ferestre de timp pot fi estimate și se obțin valori similare pt ferestre diferite
- În cazul seriilor nestaționare acest lucru nu mai este adevărat, deci înainte de a aplica o tehnică de predicție autoregresivă ar fi util ca o serie nestaționară să fie transformată într-una staționară sau se folosește o tehnică care ține cont de acest lucru.

Predicție

Exemplu: serie artificial construită: $y_i = i + z_{\text{gomot}}$ (zgomotul este generat folosind o distribuție normală de medie 0 și abatere standard 2)



Predicție

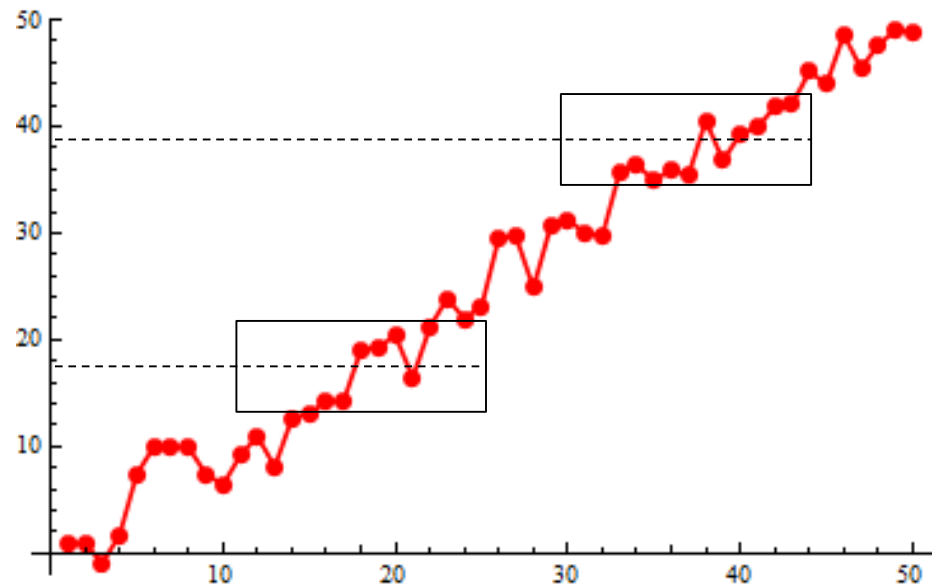
Exemplu: serie artificial construită: $y_i = i + \text{zgomot}$ (zgomotul este generat folosind o distribuție normală de medie 0 și abatere standard 2)

Obs:

▪ Aceasta este o **serie nestaționară** întrucât mediile valorilor corespunzând unor ferestre de timp diferite sunt diferite

Medie 2
(a doua fereastră)

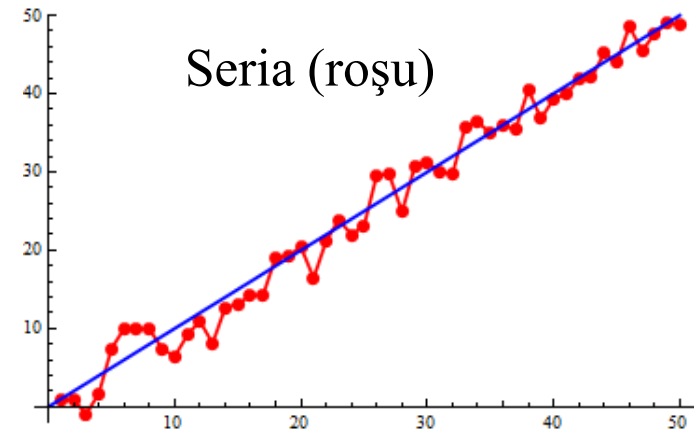
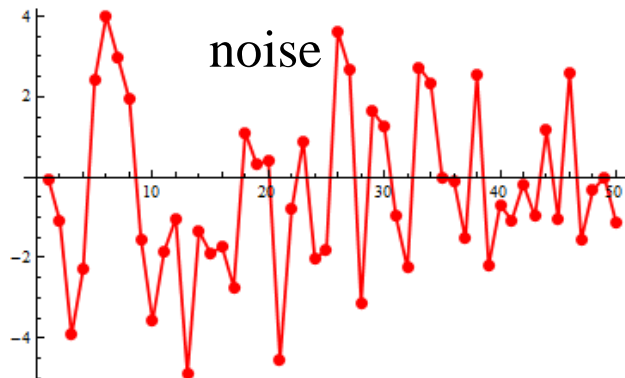
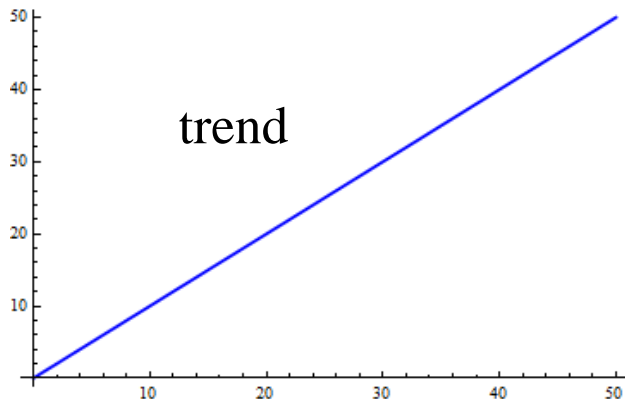
Medie 1
(prima fereastră)



Predicție

Exemplu: serie artificial construită: $y_i = i + z_{gomot}$ (zgomotul este generat folosind o distribuție normală de medie 0 și abatere standard 2)

Obs. Sunt 2 componente: **tendința (trend)** și **zgomot (noise)**

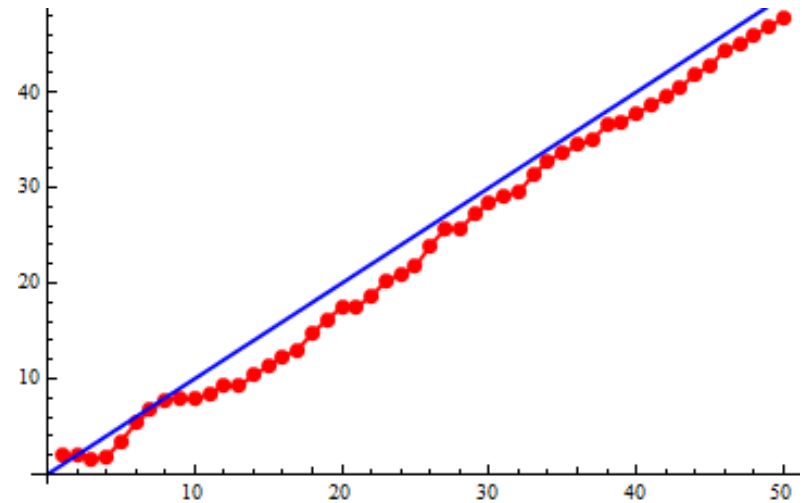
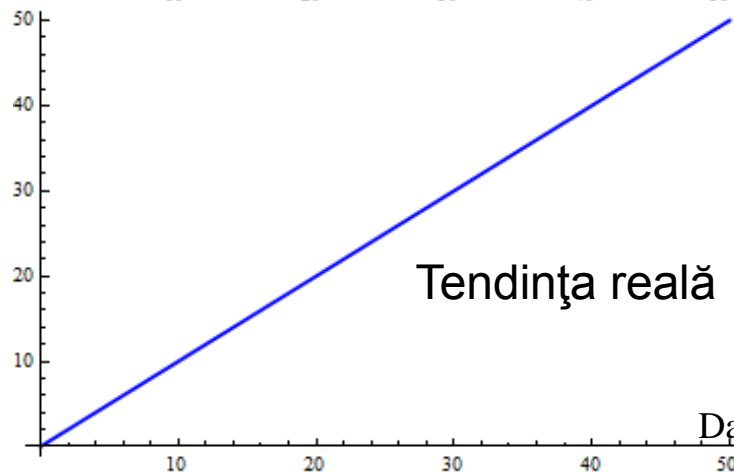
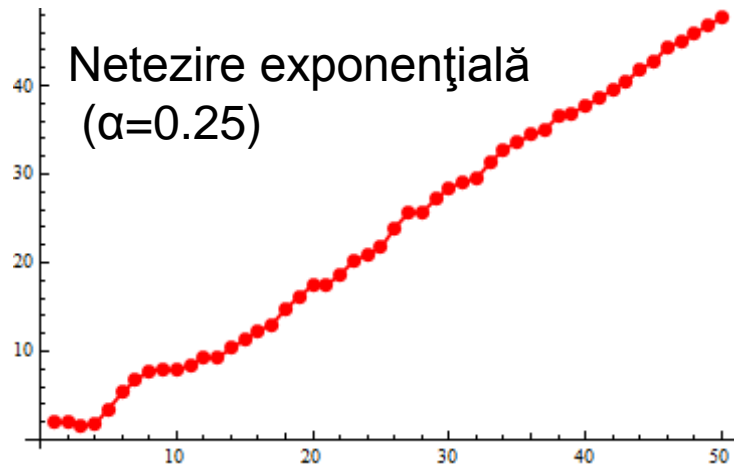


Cum pot fi extrase cele două componente din seria inițială?

Predicție

Extragerea tendinței = eliminarea zgomotului

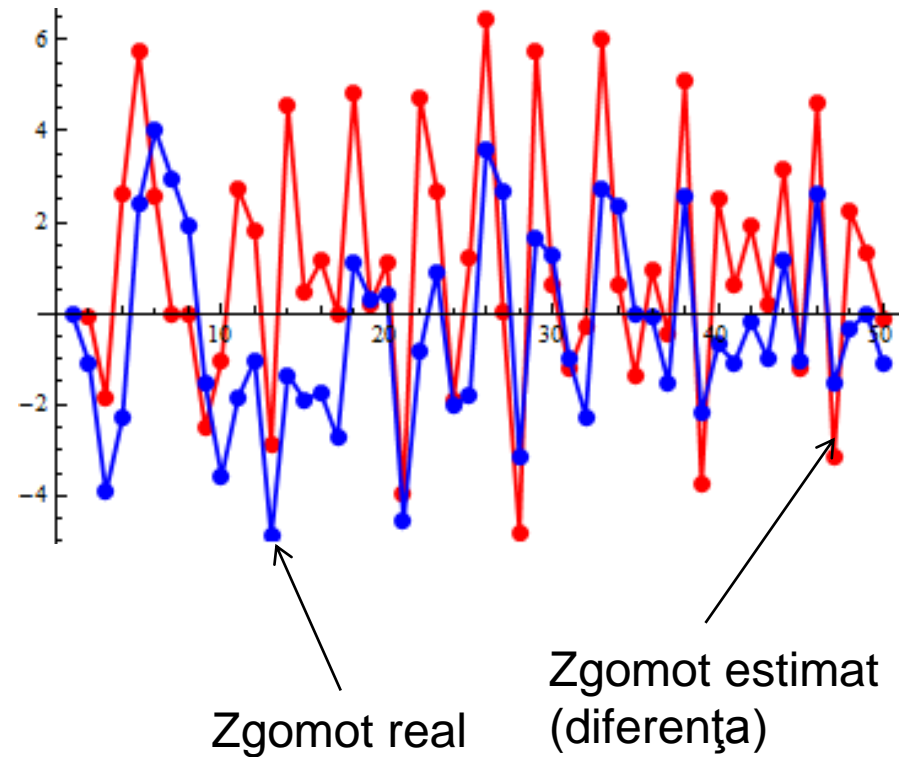
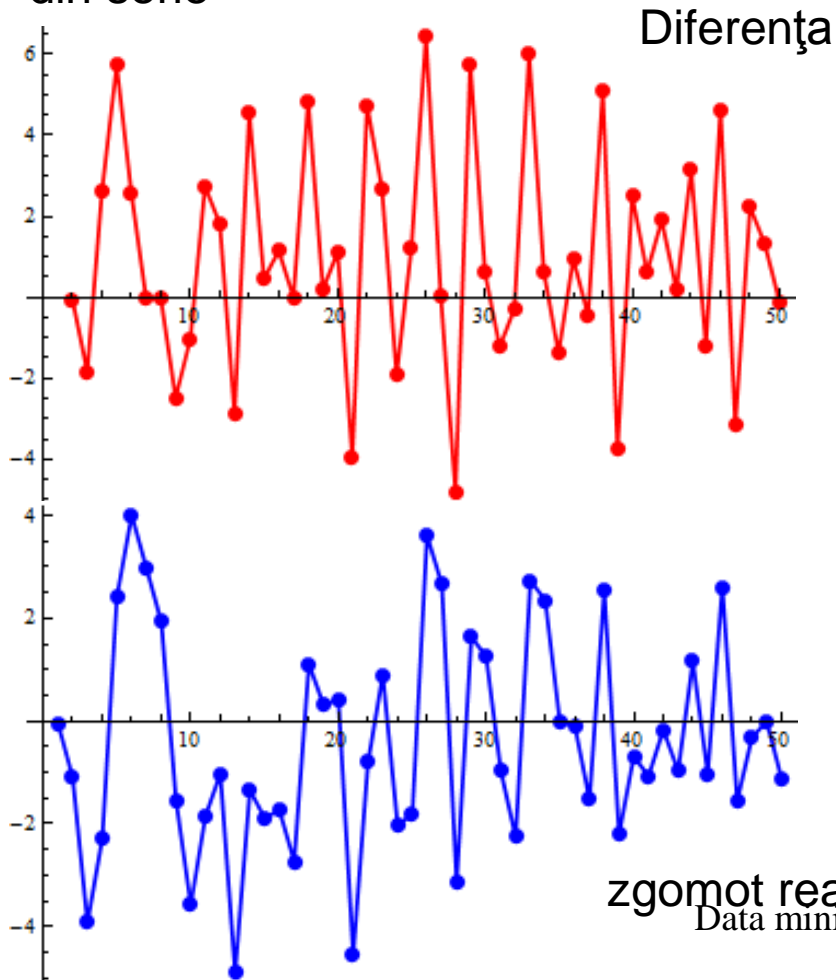
Cum poate fi realizată? prin netezire



Predicție

Extragerea zgomotului = eliminarea tendinței

Cum poate fi realizată? prin calculul diferenței dintre elementele succesive din serie



Predicție

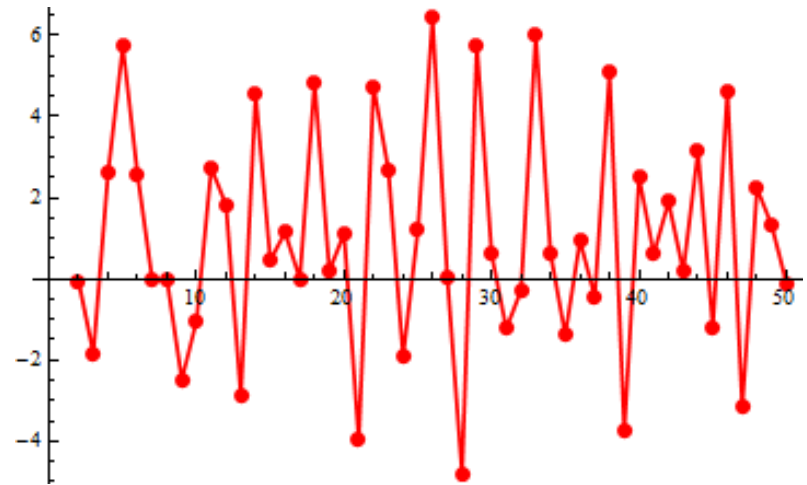
Extragerea zgomotului = eliminarea tendinței

Cum poate fi realizată? prin calculul diferenței dintre elementele succesive din serie

Transformare prin calculul diferenței: $z_i = y_i - y_{i-1}$

Obs:

- Seria obținută prin diferențiere este staționară



Diferența

Predicție

Extragerea zgomotului = eliminarea tendinței

Cum poate fi realizată? prin calculul diferenței dintre elementele succesive din serie

Alte variante:

▪ Eliminarea efectului sezonier $z_i = y_i - y_{i-P}$

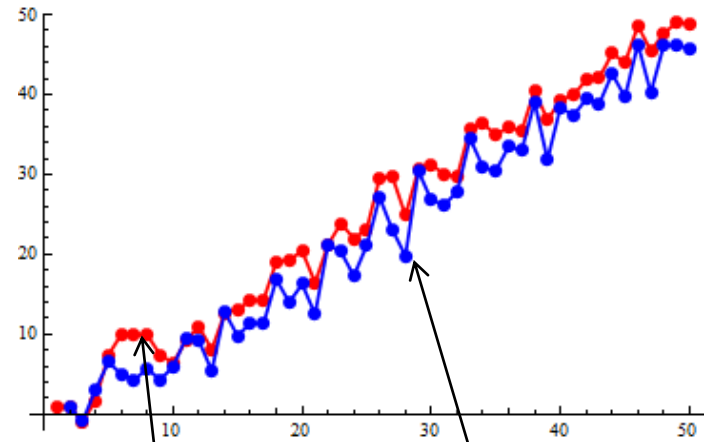
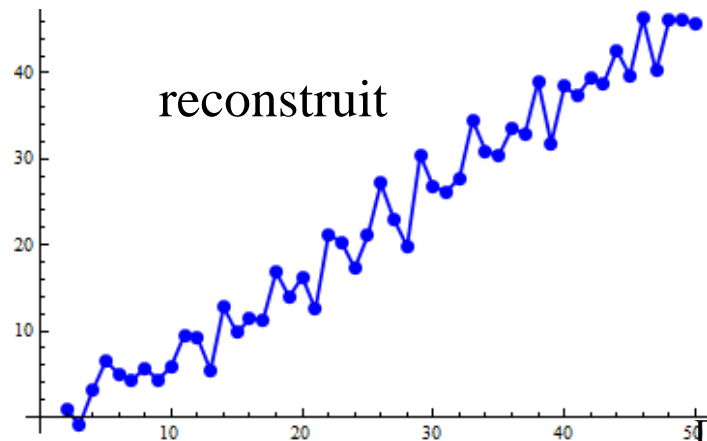
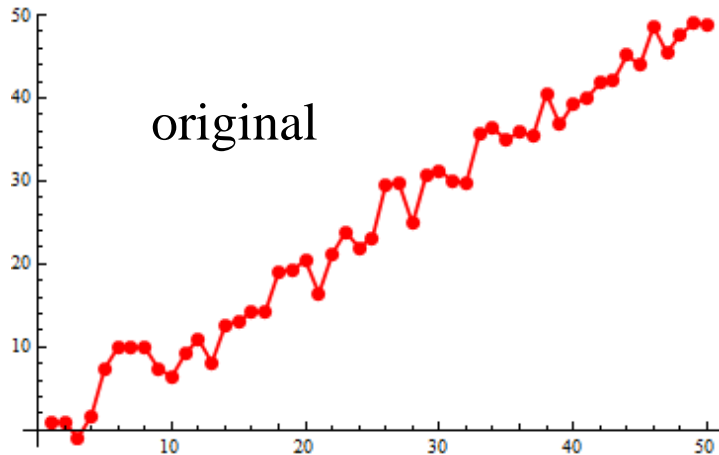
▪ La seriile cu creștere geometrică (de exemplu serii de prețuri în care factorul de inflație e constant) poate fi utilă **logaritmare** înainte de calculul diferențelor

Întrebare:

• Poate fi reconstruită seria inițială pornind de la estimările tendinței și zgomotului?

Tendința și zgomot

Reconstruire: suma dintre estimarea tendinței și estimarea zgomotului



Predicție

Cum poate fi estimată (prezisă) o nouă valoare din serie?

- estimarea unei noi valori cf modelului de tendință (trend)
- generarea unei noi valori cf modelului de zgomot
- adunarea valorilor

Problema principală:

- Este necesară construirea unui model de trend (e.g. prin regresie)
- Este necesară identificarea unui model pt zgomot; dacă este zgomot alb (valorile asociate unor momente diferite de timp sunt generate de variabile aleatoare independente cu distribuție normală și medie nulă) atunci valorile parametrilor (media și abaterea standard) pot fi ușor estimată

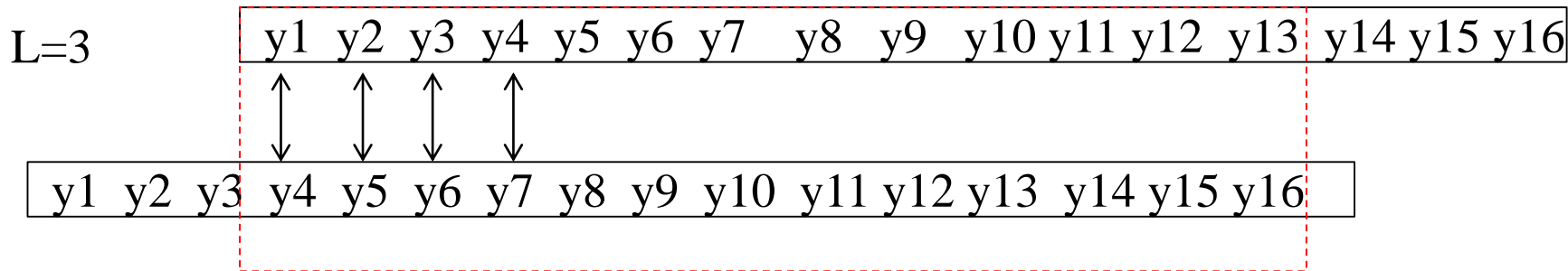
Altă abordare: se utilizează **autocorelația** = corelația dintre valorile corespunzătoare unor momente de timp învecinate

Modele autoregresive

Ideea de bază: dacă valoarea **autocorelației** este mare (în valoare absolută) atunci valoarea corespunzătoare unui moment poate fi estimată pe baza valorilor din vecinătate

Autocorelație pt o serie staționară, (y_1, y_2, \dots, y_n) = corelația dintre valori separate prin întârzierea L

$$\text{Autocorrelation}(L) = \frac{1}{n-L} \frac{\sum_{i=1}^{n-L} (y_i - \text{avg}(Y))(y_{i+L} - \text{avg}(Y))}{\text{var}(Y)}$$



Modele autoregresive

Forma generală a unui model autoregresiv de ordin p : AR(p)

$$y_t = \sum_{i=1}^p a_i y_{t-i} + c + \varepsilon_t$$

Obs:

- p este ordinul modelului și poate fi ales analizând diferite valori posibile ale întârzierii L :
 - p se alege ca fiind prima valoare L (pornind cu $L=1$) pt care valoarea absolută a auto-corelației este suficient de mică
- a_1, a_2, \dots, a_p și c sunt parametri ai modelului și se estimează folosind date de antrenare și metoda celor mai mici pătrate (similar cu modelele de regresie liniară)
- ε_t reprezintă zgomotul

Modele autoregresive

Modele de tip medie mobilă (Moving Average): MA(q)

Motivație:

- Modelele autoregresive simple nu pot explica toate variațiile (în mod particular schimbările bruște, de tipul șocurilor)

Idee:

- Modelele de tip MA prezic valorile următoare pe baza deviațiilor anterioare ale valorilor reale față de cele prezise

$$y_t = \sum_{i=1}^q b_i \varepsilon_{t-i} + c + \varepsilon_t$$

Obs:

- Presupunând că seria este staționară și zgomotul are medie 0, valoarea lui c este de fapt media valorilor din serie
- Parametrii b_1, b_2, \dots, b_q se estimează din date

Modele autoregresive

Modele autoregresive combinate: ARMA(p,q)

Motivație:

- Se combină capacitatea de predicție a modelelor autoregresive și a celor bazate pe medie mobilă:

$$y_t = \sum_{i=1}^p a_i y_{t-i} + \sum_{i=1}^q b_i \varepsilon_{t-i} + c + \varepsilon_t$$

Obs:

- Un aspect important este alegerea valorilor p și q: ar trebui alese cele mai mici valori care asigură o bună aproximare a datelor – nu este ușor de identificat

Descoperirea șabloanelor

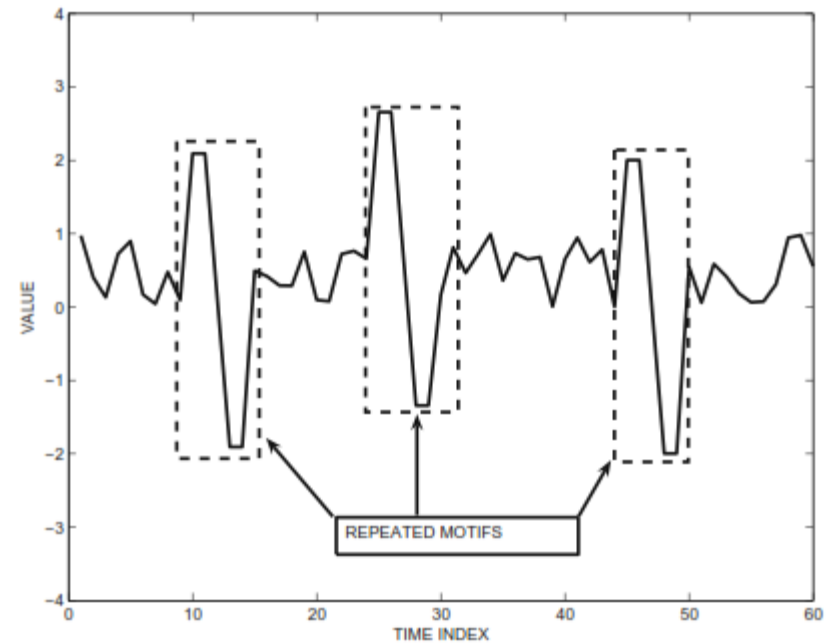
Șablon (motiv) = structură ce apare frecvent în serie

Procesul de descoperire

Intrare:

- Cel puțin o serie
- Lungimea L a șablonului
- Măsură de similaritate/ disimilaritate
- Prag pentru similaritate/ disimilaritate

Ieșire: subsecvență de lungime L ce apare frecvent în serie



C. Aggarwal, Data Mining – the Textbook, 2015

Descoperirea șabloanelor

Șablon (motiv) = structură ce apare frecvent în serie

Exemplu: Algoritm de tip forță brută

FindMotif (y[1..n],L,eps)

countMax=0

FOR i=1,n-L+1 DO

 candidate=y[i..i+L-1]

 count=0

 FOR j=1,n-L+1 DO

 D=dist(y[i..i+L-1],y[j..j+L-1])

 IF (i!=j) and (D<=eps) THEN count=count+1

 ENDFOR

 IF count[i]>countMax THEN best=i; countMax=count

ENDFOR

RETURN (y[best..best+L-1])

Excepții (anomalii)

Există două tipuri de excepții (anomalii) într-o serie de date:

Excepții (anomalii) punctuale:

- Deviație semnificativă de la valoarea prezisă
- Corespunde unei schimbări bruște în seria de date

Excepții (anomalii) în privința formei:

- O succesiune de valori poate reprezenta o anomalie chiar dacă valorile individuale nu sunt neobișnuite
- De exemplu, într-o electrocardiogramă o bătaie neregulată a inimii poate fi considerată o anomalie

Excepții (anomalii)

Detecția anomaliilor punctuale:

Step 1: se determină valoarea prezisă (pe baza modelului construit folosind valorile anterioare) $(z_m, z_{m+1}, \dots, z_n)$

Step 2: se construiește seria deviațiilor $(d_m, d_{m+1}, \dots, d_n)$ cu $d_i = z_i - y_i$

Step 3: se calculează deviațiile standardizate $(s_m, s_{m+1}, \dots, s_n)$

cu

$$s_i = (d_i - \text{avg}(d)) / \text{stdev}(d)$$

Dacă valoarea absolută a lui s_i este mai mare decât un prag (e.g. 3) atunci se consideră că este anomalie

Excepții (anomalii)

Detecția anomaliilor de formă:

Step 1: se extrag toate subseriile corespunzătoare unui ferestre de dimensiune W

Step 2: se calculează distanța dintre fiecare subserie și toate celelalte corespunzătoare unor ferestre disjuncte

Step 3: Subseriile care diferă semnificativ de celelalte sunt considerate excepții potențiale

Probleme:

- Alegerea lui W
- Alegerea pragului

Măsurarea (di)similarității

- Serii “aliniate” (valorile din cele două serii corespund acelorasi momente de timp)
 - Se poate utiliza orice măsura de (di)similaritate corespunzătoare unor date vectoriale
 - Seriiile nu sunt aliniate (de exemplu sunt înregistrări audio cu viteze diferite)
 - Se poate folosi un algoritm de matching între serii – similar algoritmilor de aliniere de secvențe biologice
 - Dynamic Time Warping (bazat pe tehnica programării dinamice)
- Idee:
- $D(i,0)=D(0,j)=\text{inf}$
 - $D(i,j)=\text{dif}(x[i],y[j])+\min\{D(i-1,j), D(i,j-1), D(i-1,j-1)\}$
 - $\text{Dist}(x[1..n],y[1..m]) = D(n,m)$