Curs 11:

Analiza seriilor de timp

Structura

- Motivaţie
- Pre-procesarea seriilor de timp
- Predicţie
- Identificare şabloane
- Grupare şi clasificare
- Detecţie anomalii

Motivaţie

Problema: Se cunosc date săptămânale privind indexul Dow Jones şi se doreşte identificarea acţiunilor pentru care creşterea de profit va fi cea mai mare în săptămâna care urmează

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

Exemple de companii cotate şi pt care sunt înregistrate informaţii:

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

Motivaţie

Problema: care acţiune va înregistra cea mai mare creştere în săptămâna care urmează?

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

Motivaţie

Problema: care acţiune va înregistra cea mai mare creştere în săptămâna care urmează?

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

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

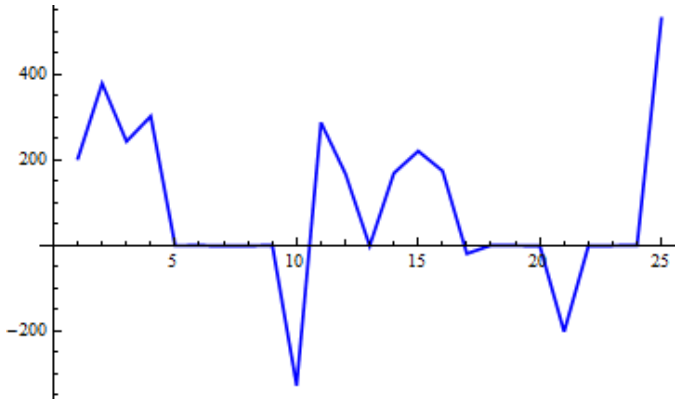
Motivaţie

Problema: care acţiune va înregistra cea mai mare creştere în săptămâna care urmează?

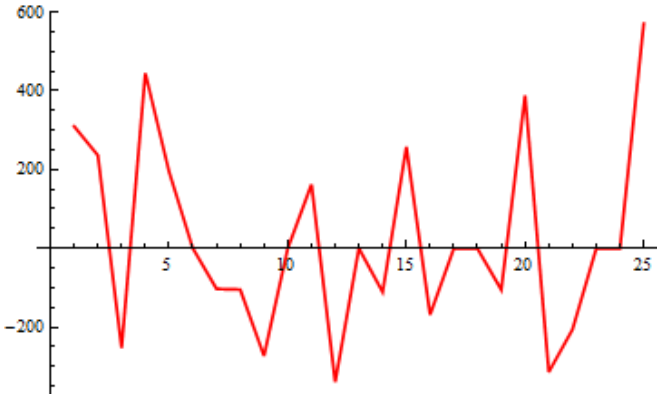
Exemplu [Dow Jones Index de la http://archive.ics.uci.edu/ml/datasets.html]

16 atribute

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



IBM

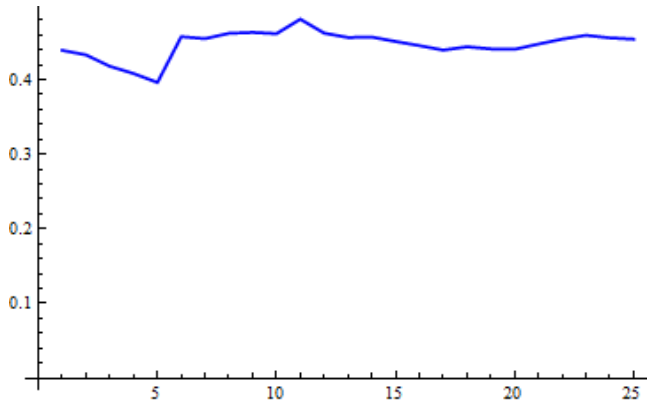


HP

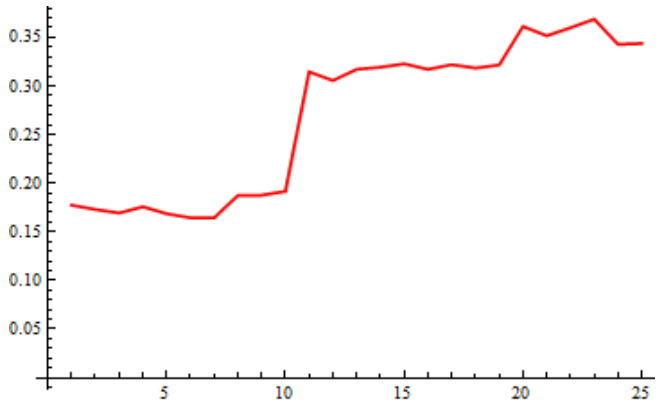
Motivaţie

Problema: care acţiune va înregistra cea mai mare creştere în săptămâna care urmează?

Exemplu [Dow Jones Index de la http://archive.ics.uci.edu/ml/datasets.html]
16 atribute
percent_return_next_dividend: the percentage of return on the next dividend



IBM



HP

Motivaţie

Pe lângă datele financiare există o mulţime 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 pacienţilor de la terapie intensivă
- Date de tip “web log” (clickstream data)
  - Secvenţe indicând vizite ale unor pagini web

Motivaţie

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

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

    Task: predicţie valori viitoare
- Date medicale
  - Electrocardiograma (ECG)
  - Electroencefalograma (EEG)
  - Date de monitorizare în timp reali a pacienţilor de la terapie intensivă

    Task: identificare comportament anormal

- Date de tip “web log” (clickstream data)
  - Secvenţe 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. Totuşi 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 secvenţe)

In raport cu atributele comportamentale

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

Pre-procesarea seriilor de timp

Valori absente

Problema:

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

Soluţie:

- Valoarea absentă este estimată folosind interpolare
- Caz simplu: interpolare liniară

Pre-procesarea seriilor de timp

Imputarea valorilor absente prin interpolare liniară

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

Presupunem ca lipseşte valoarea corespunzătoare momentului t cuprins între t_i şi 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 afectaţi de bruiaje, a.î. seria poate conţine valori generate în procesul de colectare a datelor şi care nu reflectă comportamentul real al atributului înregistrat

Modalităţi 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,….,y_n) este divizat în m subintervale conţinând 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

Observaţii:

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

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

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

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

Pre-procesarea seriilor de timp

Moving average smoothing

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

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

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

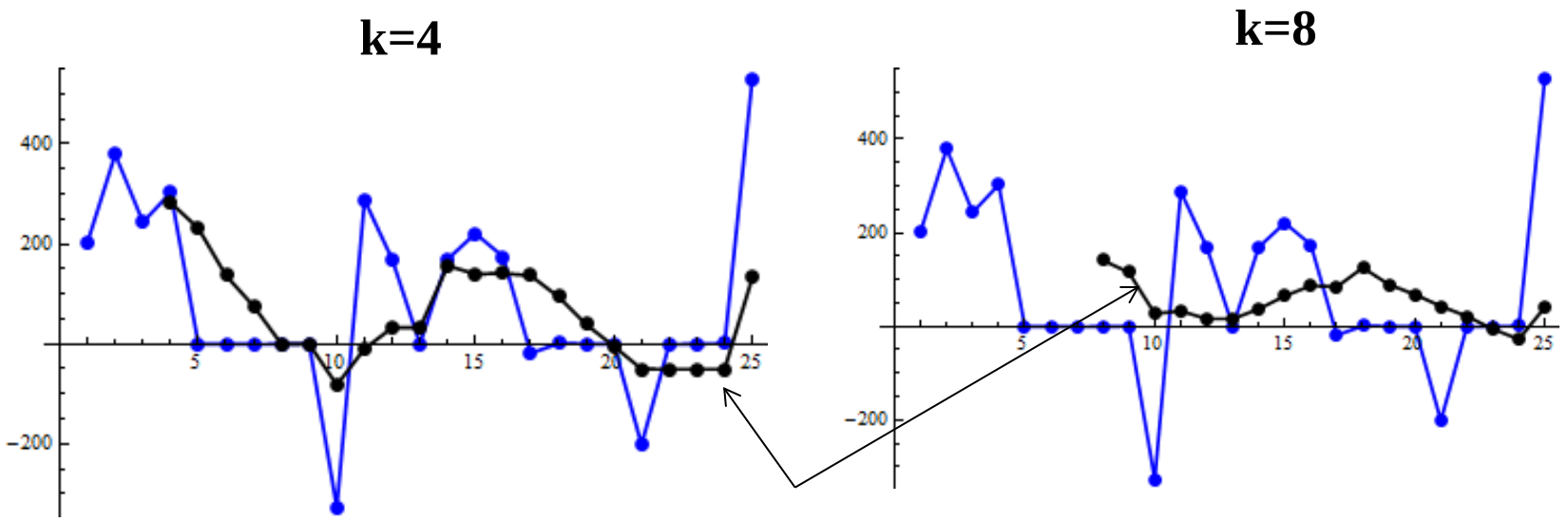
$$z_i = \frac{1}{k} \sum_{j=1}^{(m-1)k+1} y_{i+j-1}, i = \overline{1,m}$$

Obs:

- Numărul de elemente din serie este redus de la *n* la *n-k+1*
- Variaţiile pe termen scurt pot fi pierdute prin mediere
- Medierea 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)



Serii netezite (negru)

Pre-procesarea seriilor de timp

Netezire exponenţială

Idee: valoarea netezită se defineşte ca o combinaţie liniară a valorii curente şi a valorii netezite anterioare

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

$$z_0 = y_1$$

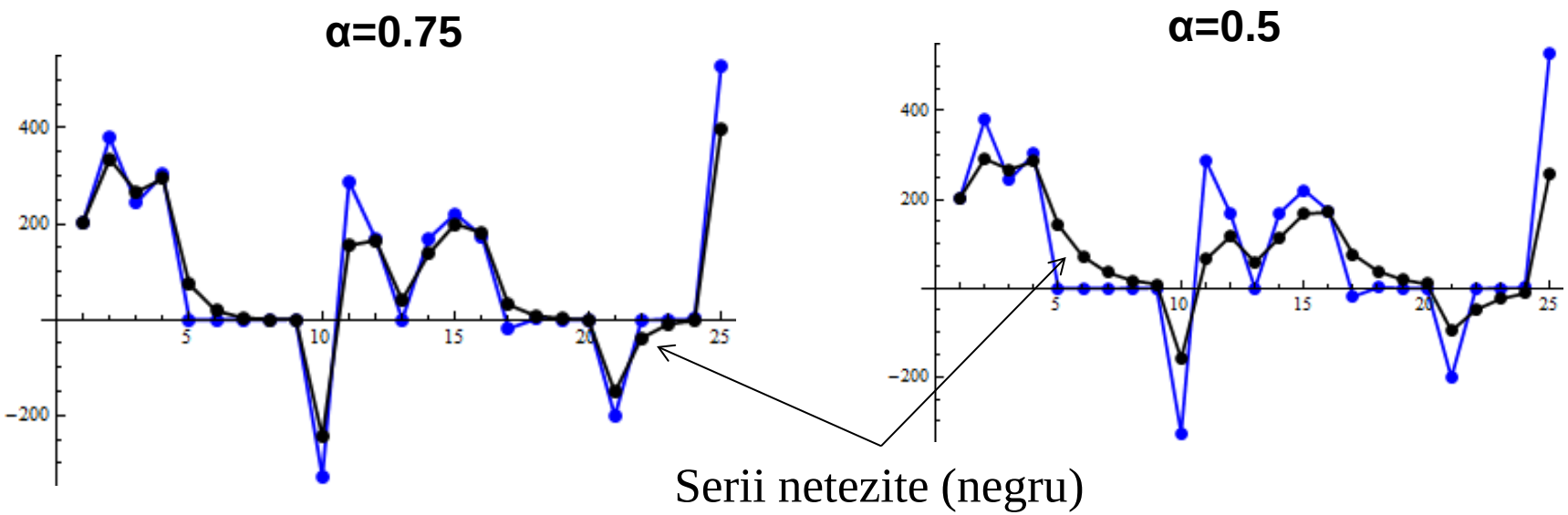
$$z_i = (1-\alpha)^i z_0 + \alpha \sum_{j=1}^{i} y_j (1-\alpha)^{i-j}, \quad i = \overline{1,m}$$

Obs:

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

Pre-procesarea seriilor de timp

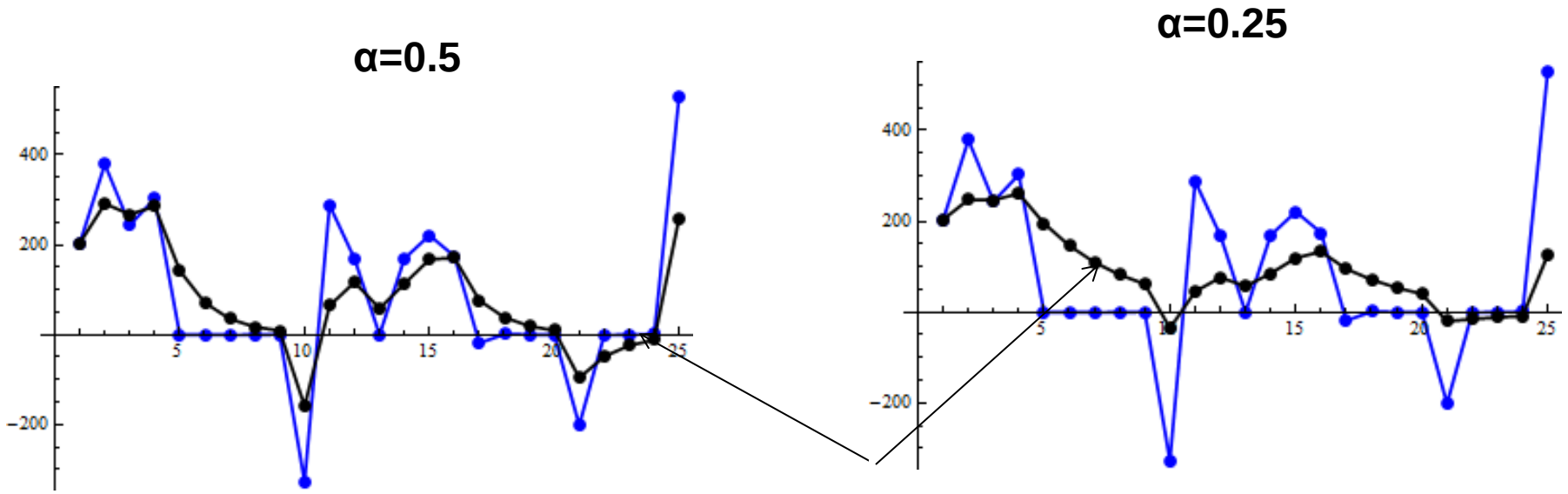
Exemplu (exponential smoothing)



Serii netezite (negru)

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

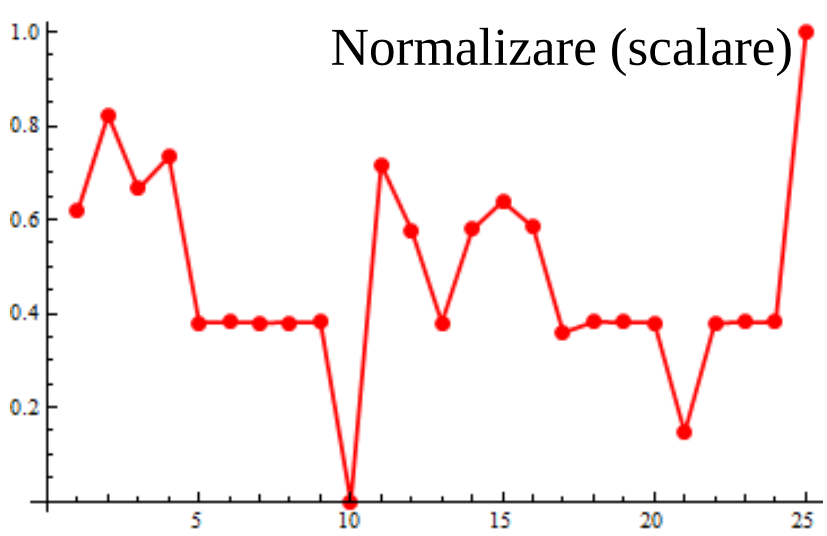
$$z_i = \frac{y_i - mean(y)}{stdev(y)}, \quad i = \overline{1, m}$$

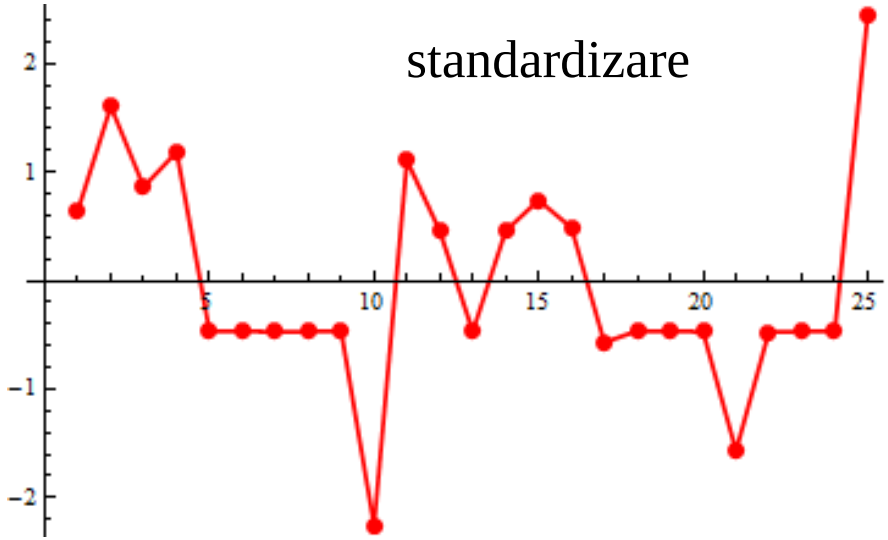
Obs:

- min(y) şi max(y) reprezintă valoarea minimă respectiv cea maximă din serie
- mean(y) şi stdev(y) sunt valoarea medie respectiv abaterea standard

Pre-procesarea seriilor de timp

Exemplu





Obs:

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

Predicţie

Scop:

- Estimarea preţului viitor al unei acţiuni, predicţia vremii, estimarea evoluţiei unor indicatori economici etc

Predicţie:

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

Cum poate fi abordată problema?

- Ca o problemă de regresie – se estimează explicit dependenţa dintre atributele comportamentale şi timp

- Utilizând modele care exprimă relaţia dintre valori curente şi valori anterioare ale seriei (modele autoregresive)

Predicţie

Obs: modelele de predicţie funcţionează bine pentru seriile staţionare

Intuitiv, o serie staţionară se caracterizează prin faptul că proprietăţile sale statistice (medie, varianţă, autocorelaţie) sunt constante în timp

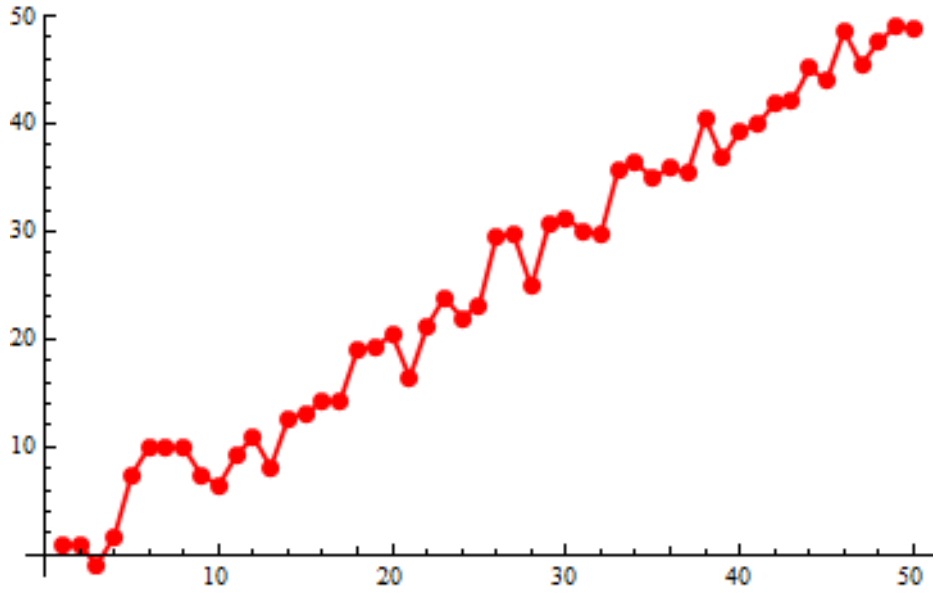
Staţionaritate strictă: distribuţia de probabilitate a valorilor din orice interval de timp [a,b] este identică cu distribuţia de probabilitate a valorilor din intervalul shiftat [a+h, b+h] (pentru un h>0 arbitrar)

Obs:

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

Predicţie

Exemplu: serie artificial construită: y_i=i+zgomot (zgomotul este generat folosind o distribuţie normală de medie 0 şi abatere standard 2)



Predicţie

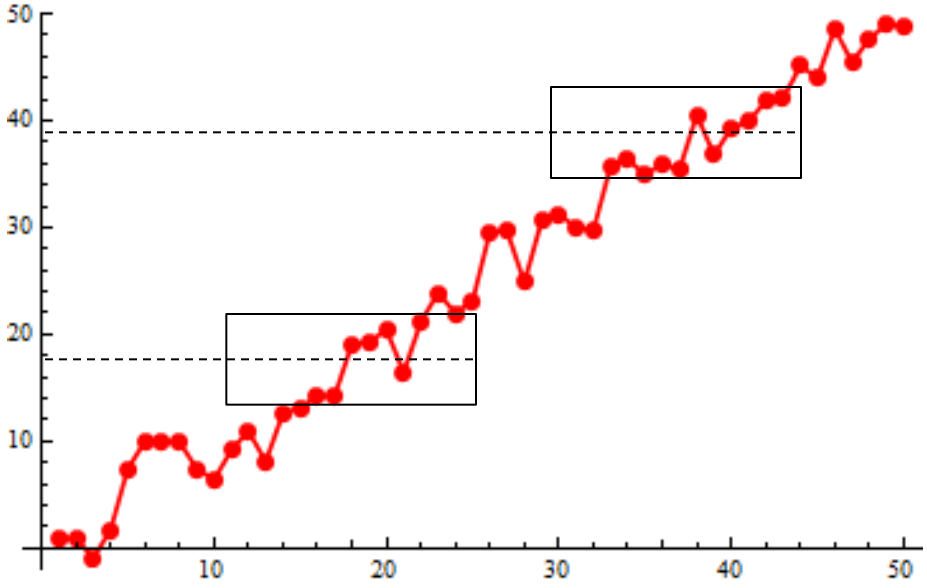
Exemplu: serie artificial construită: y_i=i+zgomot (zgomotul este generat folosind o distribuţie normală de medie 0 şi abatere standard 2)

Obs:

- Aceasta este o serie nestaţionară întrucât mediile valorilor corespunzând unor ferestre de timp diferite sunt diferite

Medie 2
(a doua fereastră)

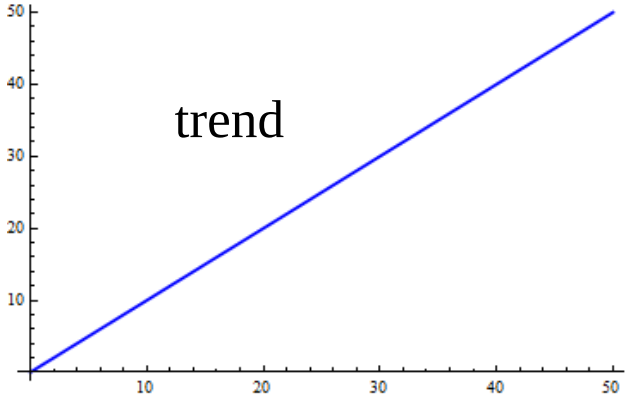
Medie 1
(prima fereastră)

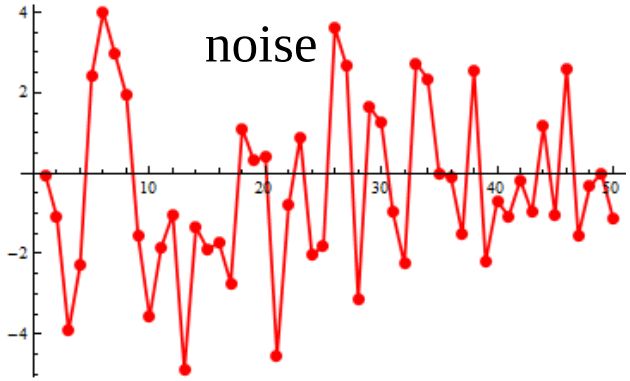


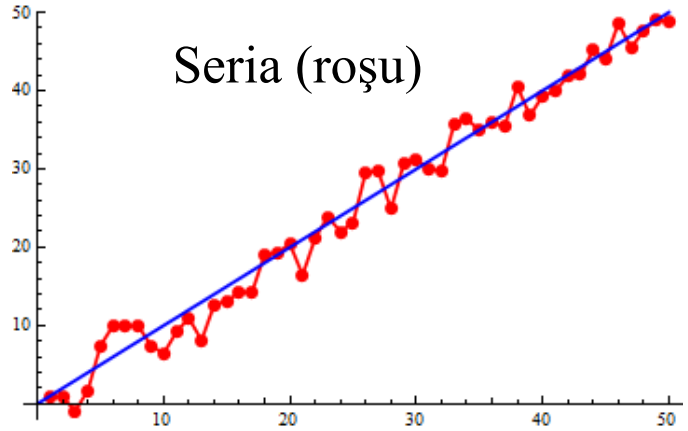
Predicţie

Exemplu: serie artificial construită: y_i=i+zgomot (zgomotul este generat folosind o distribuţie normală de medie 0 şi abatere standard 2)

Obs. Sunt 2 componente: tendinţa (trend) şi zgomot (noise)





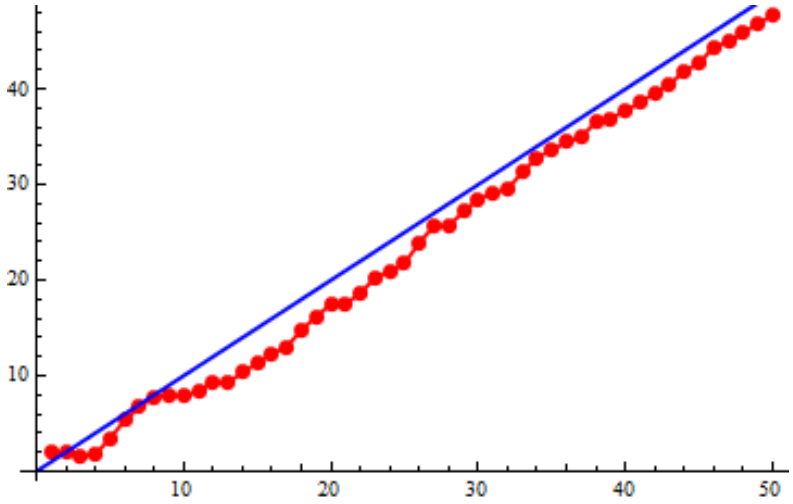


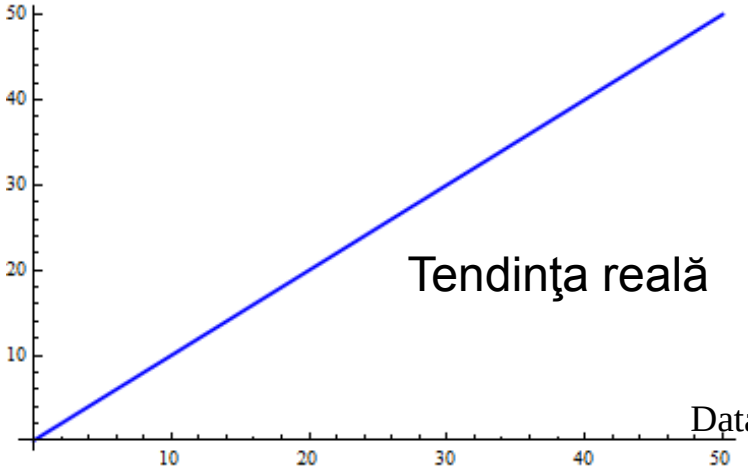
Cum pot fie extrase cele două componente din seria iniţială?

Predicţie

Extragerea tendinţei = eliminarea zgomotului

Cum poate fi realizată? prin netezire



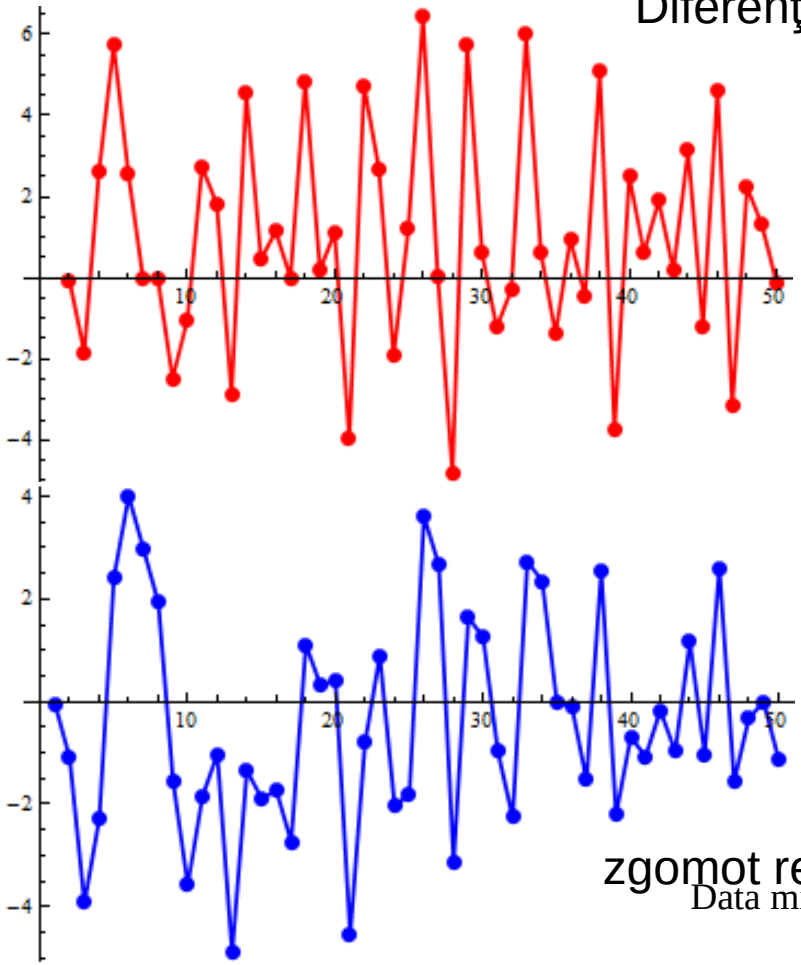


Predicţie

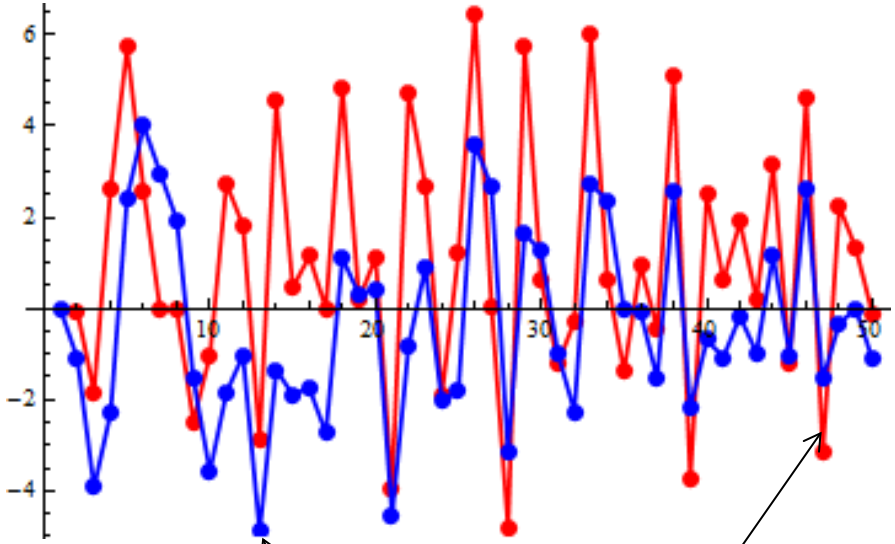
Extragerea zgomotului = eliminarea tendinţei

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

Diferenţa



zgomot real



Zgomot real

Zgomot estimat (diferenţa)

Predicţie

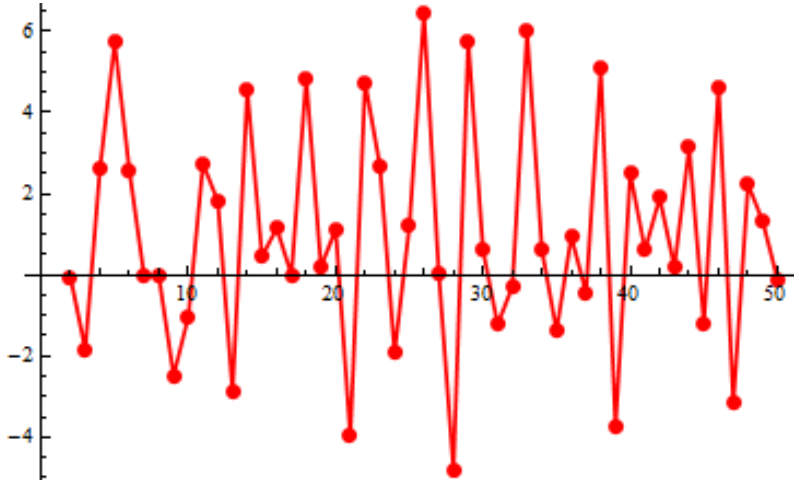
Extragerea zgomotului = eliminarea tendinţei

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

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

Obs:

- Seria obţinută prin diferenţiere este staţionară



Diferenţa

Predicţie

Extragerea zgomotului = eliminarea tendinţei

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

Alte variante:

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

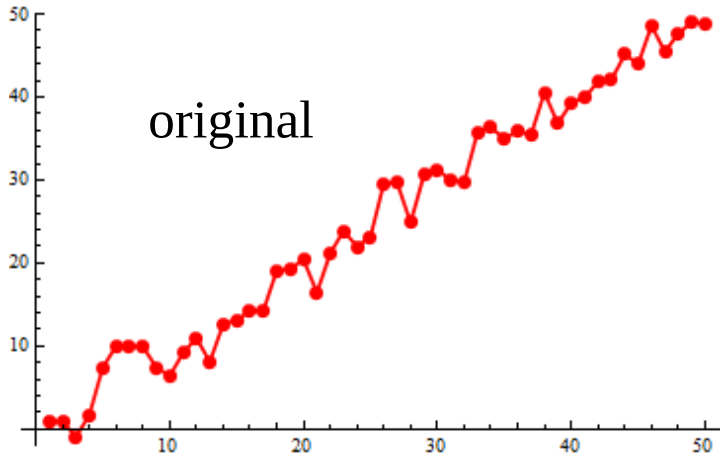
- La seriile cu creştere geometrică (de exemplu serii de preţuri în care factorul de inflaţie e constant) poate fi utilă logaritmarea înainte de calculul diferenţelor

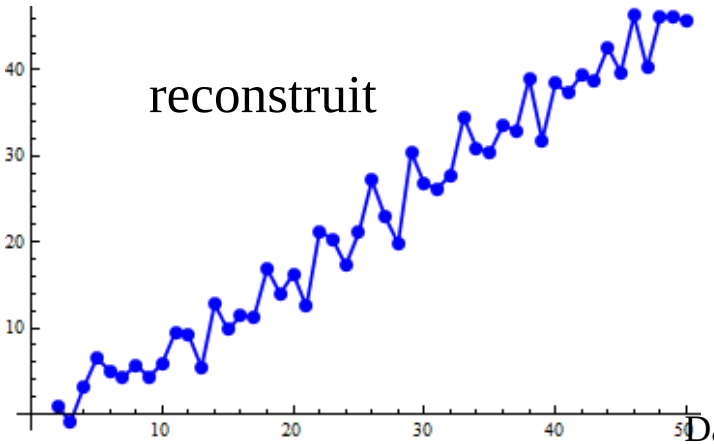
Întrebare:

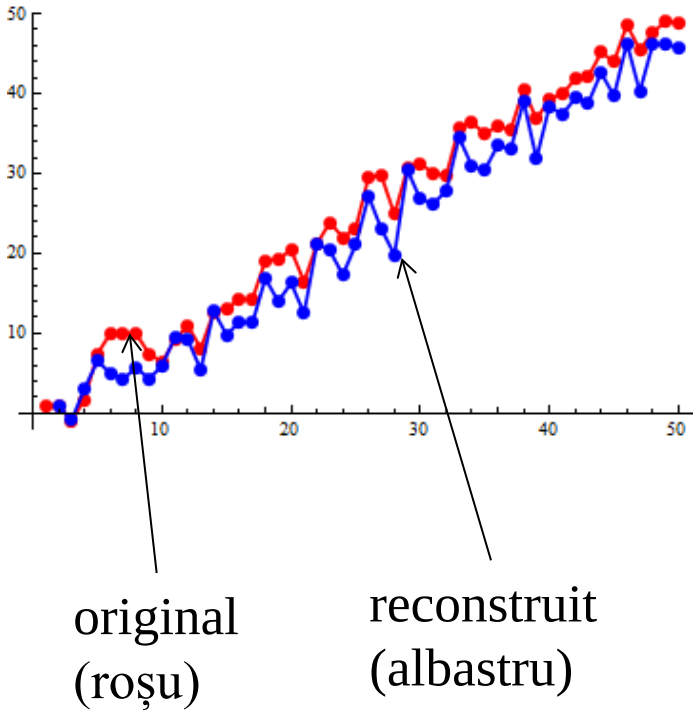
- Poate fi reconstruită seria iniţială pornind de la estimările tendinţei şi zgomotului?

Tendinţa şi zgomot

Reconstruire: suma dintre estimarea tendinţei şi estimarea zgomotului







Predicţie

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

- estimarea unei noi valori cf modelului de tendinţă (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 distribuţie normală şi medie nulă) atunci valorile parametrilor (media şi abaterea standard) pot fi uşor estimată

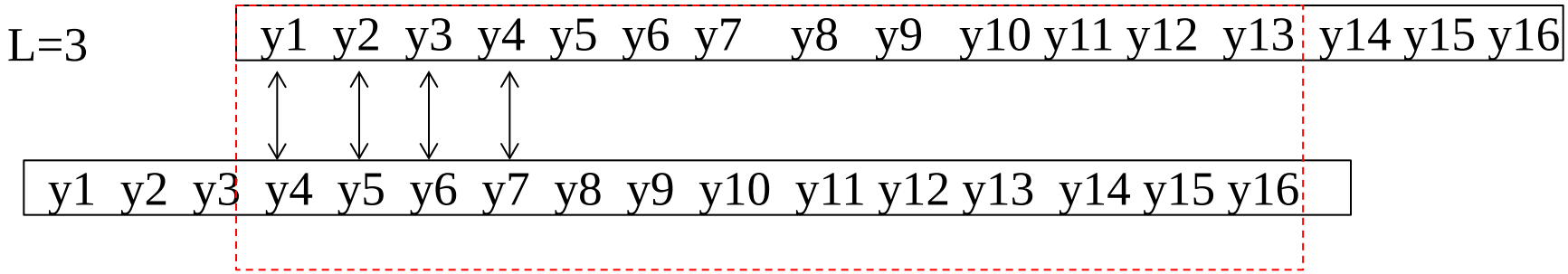
Altă abordare: se utilizează autocorelaţia = corelaţia dintre valorile corespunzătoare unor momente de timp învecinate

Modele autoregresive

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

Autocorelaţie pt o serie staţionară, (y_1,y_2,….,y_n) = corelaţia dintre valori separate prin întârzierea L

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



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 şi 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-corelaţiei este suficient de mică
- a_1, a_2,…., a_p şi c sunt parametri ai modelului şi se estimează folosind date de antrenare şi 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)

Motivaţie:

- Modelele autoregresive simple nu pot explica toate variaţiile (în mod particular schimbările bruşte, de tipul şocurilor)

Idee:

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

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

Obs:

- Presupunând că seria este staţionară şi zgomotul are medie 0 valoarea lui c este de fapt media valorilor din serie
- Parametrii b_1, b_2,…., b_q se estimează din date (problemă de fitare neliniară)

Modele autoregresive

Modele autoregresive combinate: ARMA(p,q)

Motivaţie:

- Se combină capacitatea de predicţie a modelelor autoregresive şi 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 şi q: ar trebui alese cele mai mici valori care asigură o bună aproximare a datelor – nu este uşor de identificat

Descoperirea şabloanelor

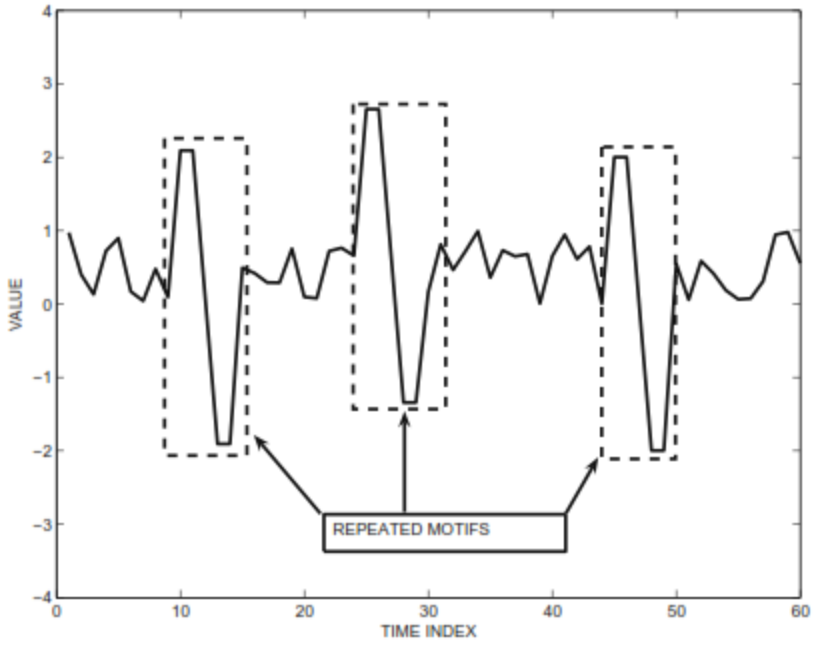
Şablon (motiv) = structură ce apare frecvent în serie

Procesul de descoperire

Intrare:

- Cel puţin o serie
- Lungimea L a şablonului
- Măsură de similaritate/ disimilaritate
- Prag pentru similaritate/ disimilaritate

Ieşire: subsecvenţă de lungime L ce apare frecvent în serie



C. Aggarwal, Data Mining – the Textbook, 2015

Descoperirea şabloanelor

Şablon (motiv) = structură ce apare frecvent în serie

Exemplu: Algoritm de tip forţă 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])
```

Excepţii (anomalii)

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

Excepţii (anomalii) punctuale:

- Deviaţie semnificativă de la valoarea prezisă
- Corespunde unei schimbări bruşte în seria de date

Excepţii (anomalii) în privinţa formei:

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

Excepţii (anomalii)

Detecţia anomaliilor punctuale:

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

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

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

cu

$s_i = (d_i - avg(d))/stdev(d)$

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

Excepţii (anomalii)

Detecţia anomaliilor de formă:

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

Step 2: se calculează distanţa dintre fiecare subserie şi toate celelalte corespunzătoare unor ferestre disjuncte

Step 3: Subseriile care diferă semnificativ de celelalte sunt considerate excepţii potenţiale

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

- Seriile 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 secvente biologice
  - Dynamic Time Warping (bazat pe tehnica programării dinamice)
  - Idee:
  - D(i,0)=D(0,j)=inf
  - D(i,j)=dif(x[i],y[j])+min{D(i-1,j), D(i,j-1), D(i-1,j-1)}
  - Dist(x[1..n],y[1..m]) = D(i,j)