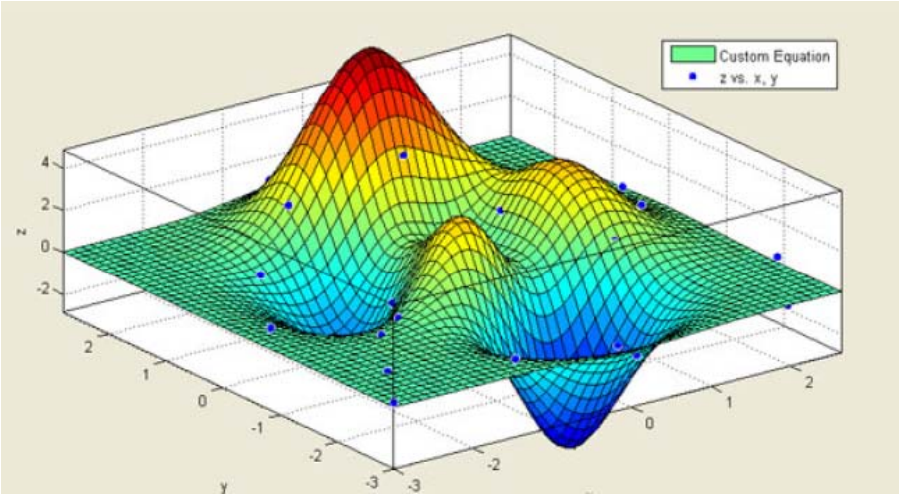
# CURVE FITTING TOOLBOX




**Nenad Bolf**
Sveučilište u Zagrebu
Fakultet kemijskog inženjerstva i tehnologije

bolf@fkit.hr http:// **LAM.fkit.hr**

# CURVE FITTING TOOLBOX™

**Podešavanje ("fitanje") krivulja i površina shodno realnim podacima primjenom tehnika regresije (regression), interpolacije (interpolation) i izglađivanja (smoothing).**

- Analiza podataka
- Preodbrada podataka
- Naknadna obrada podataka
- Usporedba potencijalnih modela
- Uklanjanje ekstremnih vrijednosti (**outliers**)
- Regresijska analiza s linearnim i nelinearnim modelima ili vlastitim jednadžbama
- Neparametarske tehnike modeliranja – **spline**, **interpolacija** i **izglađivanje**.

# CURVE FITTING TOOLBOX

## Osnovne karakteristike i funkcije

- ✓ Pregled i analiza podataka te vizualno i numeričko podešavanje
- ✓ Statistički pokazatelji za ocjenu kvalitete podešavanja
- ✓ Mogućnost ekstrapolacije, diferenciranja i integriranja
- ✓ Spremanje podataka u nekoliko formata (M-datoteke, binarno, radni prostor)

## Dva načina rada

- **Grafičko korisničko sučelje** (**Grafical User Interface – GUI**)
- *Matlab*-ov **naredbeni redak** (**MATLAB command line**)

Za većinu zadataka preporuča se upotreba **GUI**-a.

Prije rada u *Curve Fiiting Toolbox*-u varijable s podacima moraju postojati u radnom prostoru ( *workspace*-u ).

# ŠTO JE CURVE FITTING TOOLBOX?

## Pokretanje grafičkog sučelja:

Utipkati **cftool** u *Command promt*-u

## Čemu služi ovaj toolbox?

- **Predobrada** podataka
  - podjela na sekcije (**sectioning**) i izglađivanje podataka (**smoothing**)
- **Podešavanja** (**fitting**):
  - **parametarsko** podešavanje
    (polinomne, eksponencijalne i racionalne funkcije,
    sume Gaussovih funkcija, posebne jednadžbe itd. )
  - **neparametarsko** podešavanje

    ***spline*** ili **interpolacije** – standardne linearne i nelinearne metode najmanjih kvadrata, robusni postupci podešavanja

# PREDOBRADA PODATAKA – IZGLAĐIVANJE

## IZGLAĐIVANJE (smoothing)

- Ako su podaci opterećeni **šumom** primjenjuje se **algoritam izglađivanja** kako bi se olakšalo parametarsko podešavanje;
- Dvije **osnovne pretpostavke** za izglađivanje su:
  - veza između **odziva** i **prediktora** (nezavisne varijable) je glatka,
  - postupak izglađivanja daje rezultate koji daju bolju procjenu originalnih vrijednosti jer je šum smanjen;
- Izglađivanje se može smatrati lokalnim podešavanjem jer se nove vrijednosti generiraju za svaku originalnu vrijednost odziva;
- Prema tome, izglađivanje je slično neparametarskom podešavanju (npr. **smoothing spline** ili **cubic** *interpolation*);
- Ipak, ova vrsta podešavanja razlikuje se od parametarskog čiji rezultat su **globalni parametri**.

# METODE IZGLAĐIVANJA

**Filtiranje** (**averaging**) i **lokalna regresija**

- Kod svake metoda definira se određeni raspon (**span**) –područje susjednih točaka koje se uključuju u proračun nove točke;

- Ovaj raspon se pomiče duž podataka korak po korak za svaku novu vrijednost prediktora;

  - **Veliki raspon** povećava glatkoću, ali **smanjuje rezoluciju** izglađenih podataka;
  - **Mali raspon** smanjuje glatkost, ali **povećava rezoluciju** izglađenih podataka;

- Optimalni raspon ovisi o skupu podatka, metodi izglađivanja i obično zahtijeva ponešto eksperimentiranja.

# METODE IZGLAĐIVANJA – Moving average filtering

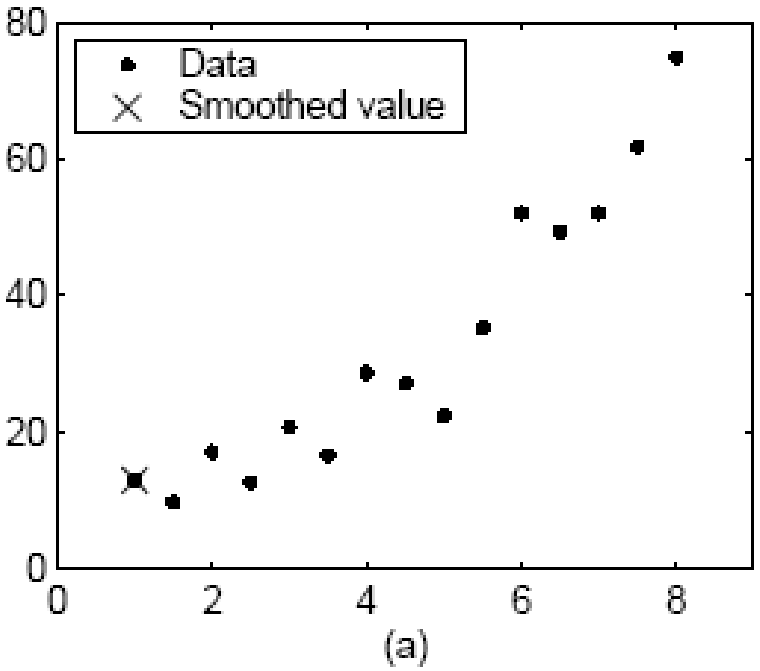
## Moving average filtering

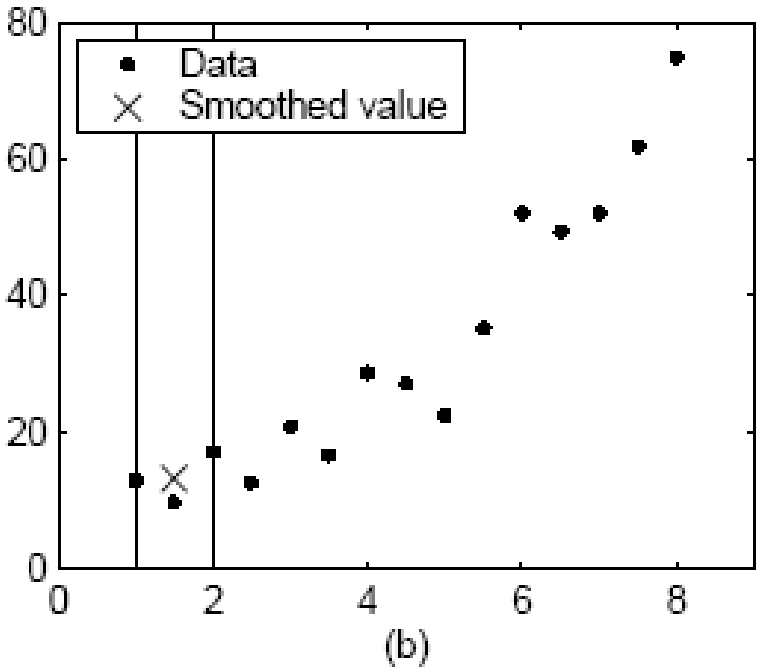
- Filtar koji **propušta niske frekvencije** i uzima sredinu od susjednih podataka;
- Izglađuje podatke tako da **svaki podatak zamjenjuje sa srednjom vrijednosti susjednih točaka** definiranih rasponom;
- Rezultat je dan s jednadžbom razlika:

$$y_s(i) = \frac{1}{2N+1}(y(i+N) + y(i+N-1) + \ldots + y(i-N))$$

| | |
|---|---|
| $y_s(i)$ | izglađena vrijednost *i*-tog podatka |
| $N$ | broj susjednih podataka s obje strane $y_s(i)$ |
| $2N+1$ | raspon |

# *MOVING AVERAGE* FILTRIRANJE



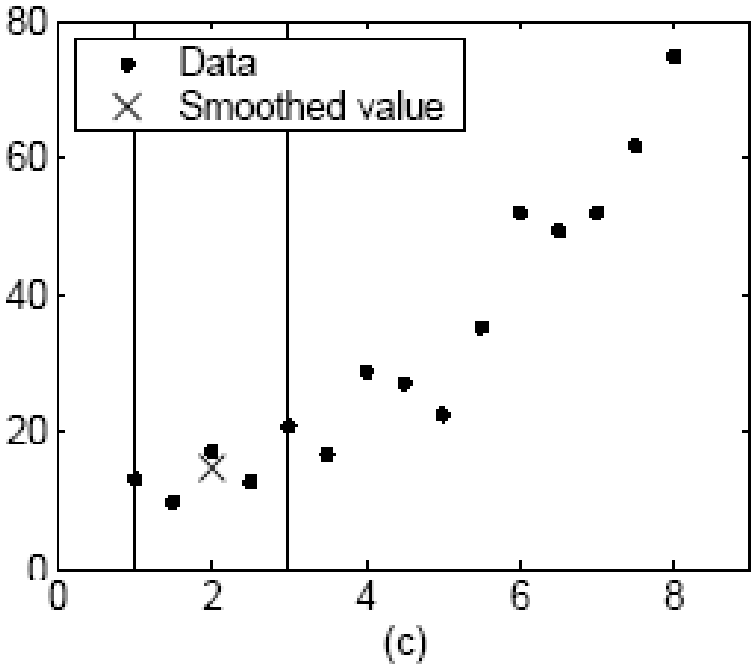



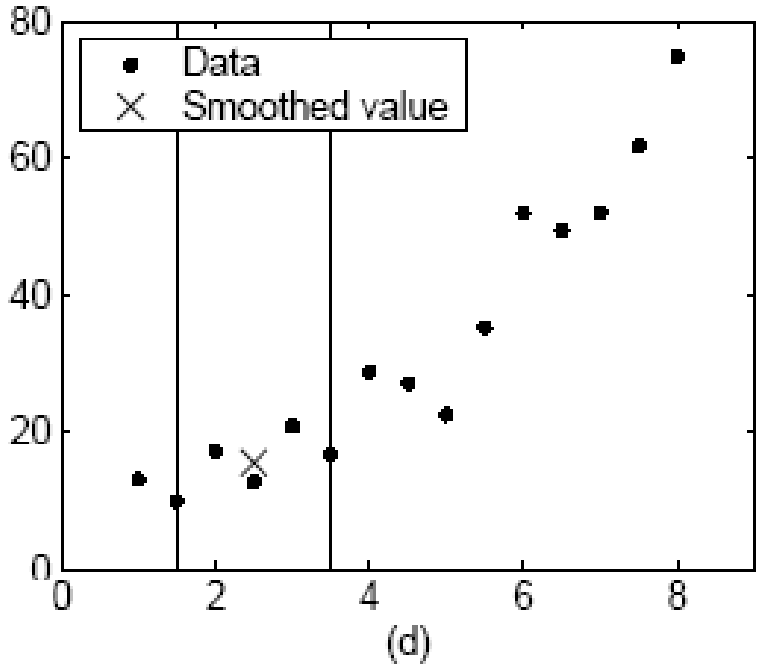

**(a)** Prva točka nije izglađena jer ne postoji raspon

**(b)** Druga točka je izglađena primjenom raspona od tri podataka

$$y_s(1) = y(1)$$
$$y_s(2) = (y(1)+y(2)+y(3))/3$$

# *MOVING AVERAGE* FILTRIRANJE




**(c)** i **(d)** za proračun izglađene vrijednosti primjenjuje se raspon od 5 točaka

$$y_s(3) = (y(1)+y(2)+y(3)+y(4)+y(5))/5$$
$$y_s(4) = (y(2)+y(3)+y(4)+y(5)+y(6))/5$$

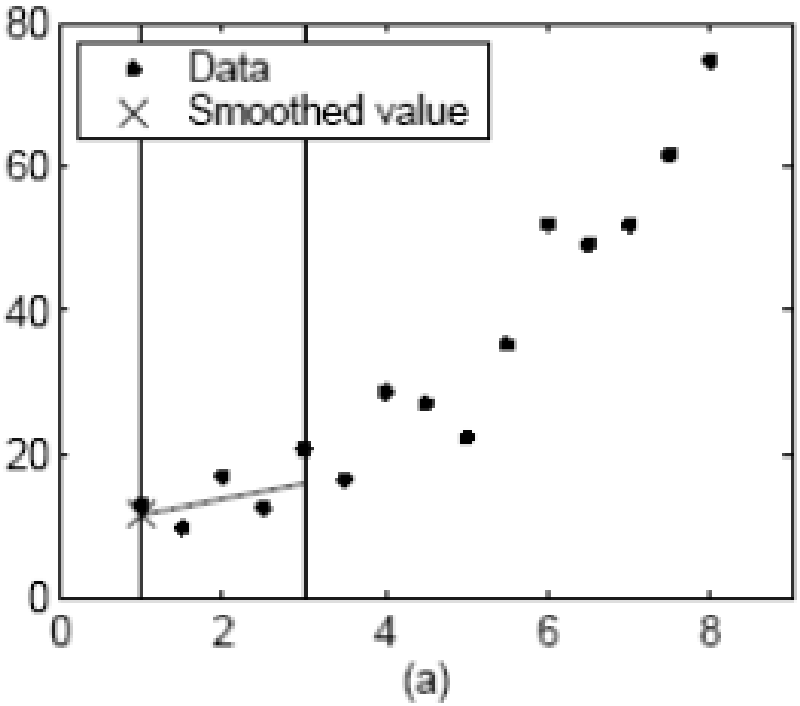
# LOWESS i LOES – izglađivanje primjenom lokalne regresije

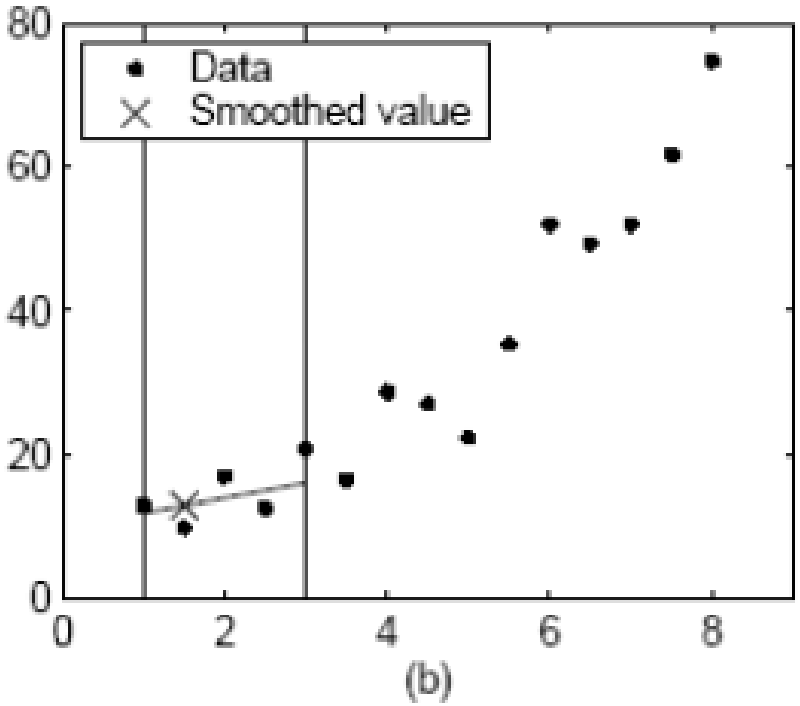
**Lowess**, **loess** – "***locally weighted scatter plot smooth***"

- Obje metode primjenjuju **lokalnu težinsku** linearnu regresiju;
- Smatra se "**lokalnom**" jer se vrijednost određuje na temelju **susjednih točaka** definiranih rasponom;
- Podaci imaju svoje težinske vrijednosti, a primjenjuje se i robusna težinska funkcija kako bi isključile ekstremne vrijednosti (***outlier***-i)
- **Lowess** – primjena linearnog polinoma 1. reda
  **Loess** – primjena kvadratnog polinoma 2. reda
- U **CFT**-u vrijede slijedeća pravila:

  - raspon može obuhvatiti neparan ili paran broj podataka
  - raspon se specificira kao postotak ukupnog broja podatka
    (npr. 0,1 znači 10% podataka)

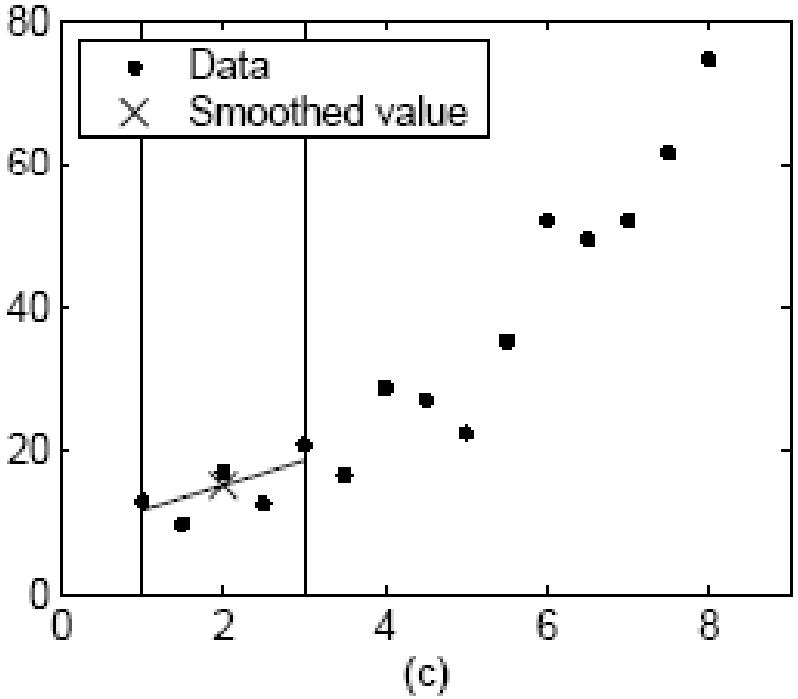
# LOWESS i LOES – izglađivanje lokalnom regresijom

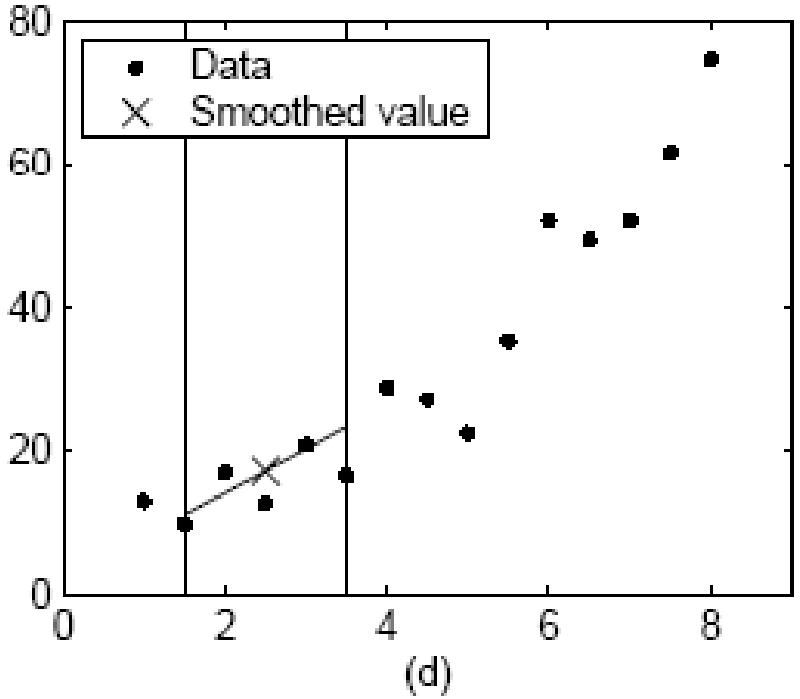



Raspon je stalan, a postupak izglađivanja provodi se od točke do točke.
Kod **(a)** i **(b)** upotrebljava se asimetrična težinska funkcija

# LOWESS i LOES – izglađivanje lokalnom regresijom

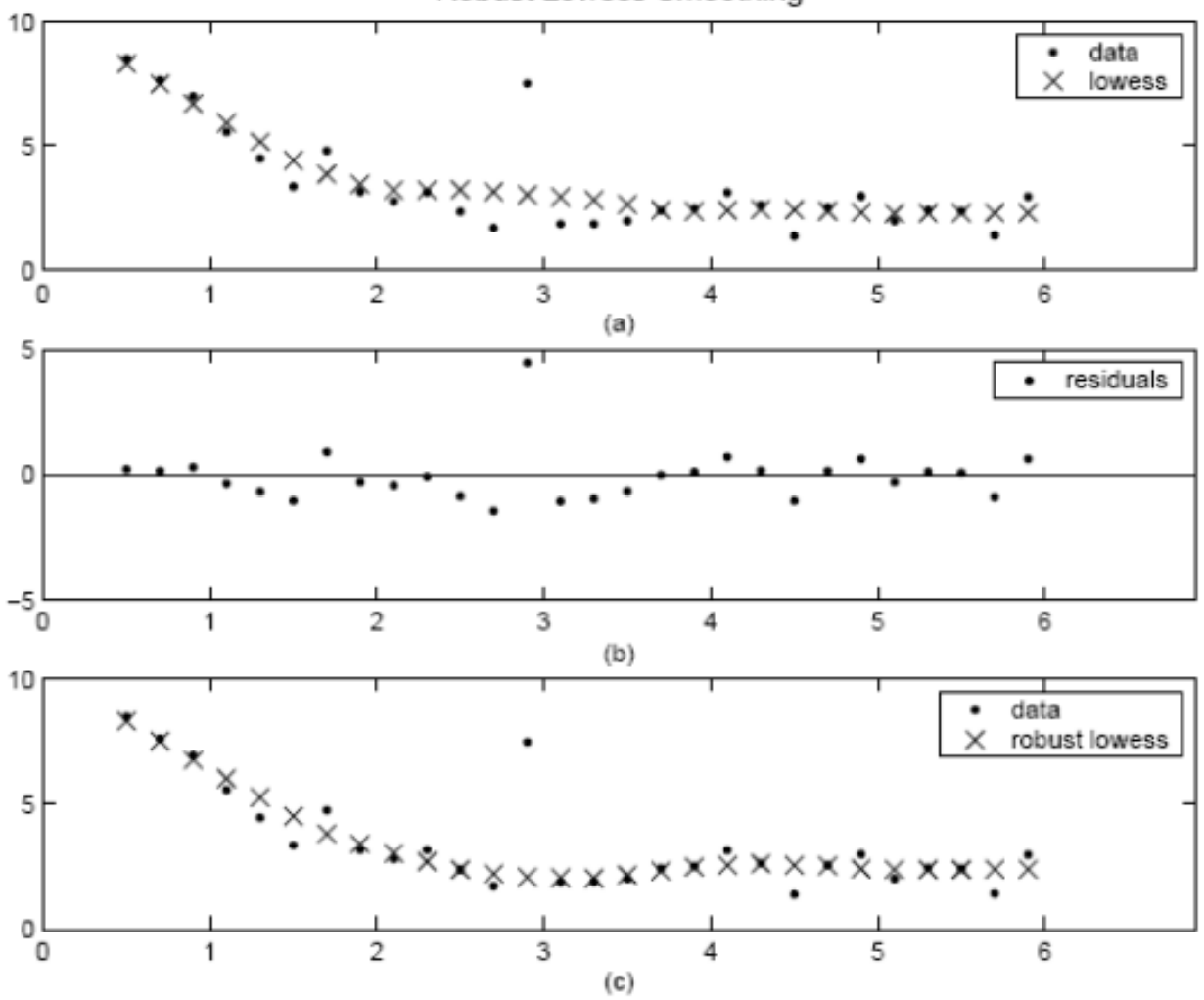



Ovisno o broju najbližih susjeda, težinske funkcije ne moraju biti simetrične oko točke oko koje se izglađuje.

Kod **(c)** i **(d)** upotrebljava se simetrična težinska funkcija.

# LOWESS i LOES – Robusno izglađivanje




(a) *Outlier* utječe na izglađenu vrijednost za nekoliko susjednih točaka
(b) Prikaz ostataka – veći su od 6 medijana apsolutnih odstupanja
(c) Izglađene vrijednosti oko *outlier*-a odražavaju većinu podataka

# *SAVITZKY-GOLAY* FILTAR

- Poopćeni "*moving average*" postupak pri čemu se određuje koeficijent filtra provedbom neteţinskog podešavanja linearnom metodom najmanjeg kvadrata primjenom polinoma određenog stupnja (naziva se još i digitalni polinomski filtar za izglađivanje – **digital smoothing polynomial filter** ili **least squares smoothing filter**)

- Višim redom polinoma može se postignuti visoka razina izglađivanje bez prigušenja podataka

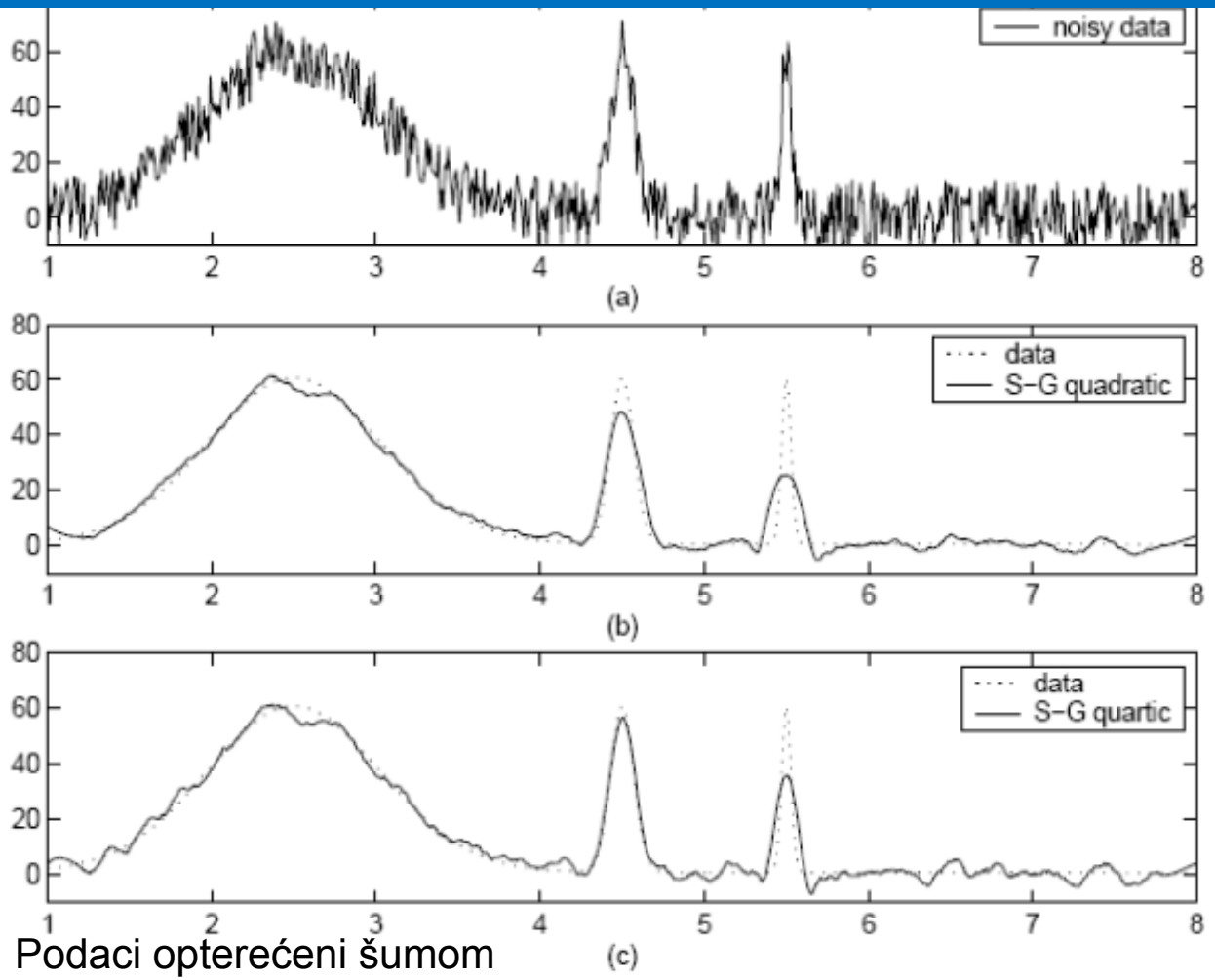
- Često se koristi pri obradi frekvencijskih ili spektroskopskih podataka (s pikovima)

- Kod frekvencijske analize djelotvoran je za očuvanje visokofrekventnih komponentni signala

- Kod spektroskopske analize dobar je za očuvanje vrhova pikova

- Za usporedbu, MA filtrira veliki dio visokofrekventnog sadržaja, a SG je manje uspješan od MA kod skidanja šuma

# *SAVITZKY-GOLAY* FILTAR



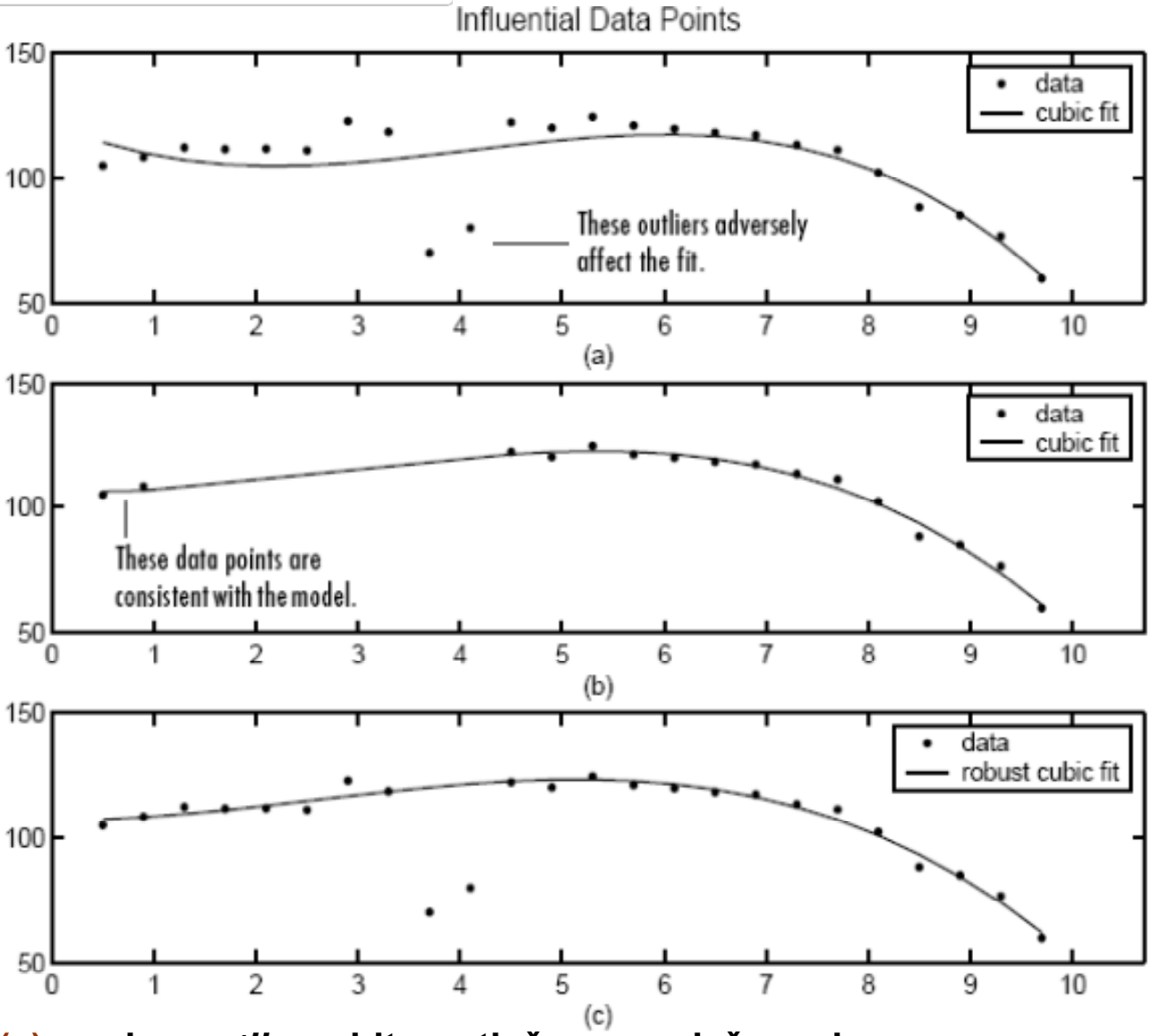

(a) Podaci opterećeni šumom

(b) Podaci bez šuma – izglađivanje kvadratnim polinom, ima problema s uskim pikovima

(c) Podaci bez šuma – izglađivanje kvartnim polinomom; općenito, što je veći red bolje se "hvataju" uski pikovi, ali slabije širi

# ISKLJUČIVANJE I PODJELA PODATAKA

- Ako postoji potreba, dio podataka može se **isključiti** iz skupa za podešavanje;
- Obično se isključuju podaci koji bi mogli **poremetiti** slijedeća podešavanja u nizu;
  (npr. podešavanje parametarskog modela s mjernim podacima kod kojih postoji **prekid zbog kvara senzora**)
- Unutar CFT podaci se isključuju na dva načina:
  - **Obilježavanje *outlier*-a** (**marking outliers**)
    *Outlieri* su **pojedine točke** koje su **nekonzistentne** sa statističkim karakteristikama ostalih podataka
    (npr. gruba mjerne pogreška, podatak koji znatno odstupa)
  - ***Podjela na sekcije*** ( **sectioning** )
    **Isključuje dio podataka**
    (npr. odvajanje većeg dijela podataka zbog sustavne pogreške)




**(a)** dva *outlier*-a bitno utječu na podešavanje
**(b)** dva podatka koja su konzistentna s modelom
**(c)** robusni postupak prihvatljiva je alternativa za obilježavanje *outlier*-a

# DODATNI KORACI PRI PREDOBRADI

- Dodatni koraci pri predobradi koji nisu dostupni u CFT su:

  - **transformiranje podataka**
  - **uklanjanje *INFs*, *NaNs* i *outlier*-a**

- U nekim slučajevima potrebno je podatke **transformirati**;
- Obično se primjenjuje **logaritmska** ln(y) i **eksponencijalna funkcija** kao što su $y^{1/2}$, $y^{-1}$;
- Na taj način mogu se **linearizirati** modeli, **pojednostaviti** modeli koji obuhvaćaju velike raspone (npr. *pH* umjesto $c(H^+)$) ili se **reducira broj koeficijenata** modela;
- Iako **CFT** ignorira ***Inf**s* i ***NaN**s* ponekad ih je potrebno prethodno ukloniti iz skupa podataka za što postoje odgovarajuće funkcije u MATLAB-u (**isinf**, **isnan**).

# PODEŠAVANJE PODATAKA

## FITTING

- Postupkom podešavanja krivulja se usklađuje tako da slijedi postojeće eksperimentalne podatke;
- Krivulje se dobivaju tehnikama **regresije**, ***spline***-om ili **interpolacijom**;
- Podaci koji se obrađuju obično su izmjereni senzorima, dobiveni simulacijom, preuzeti iz baza podataka itd.;
- Cilj podešavanja je:

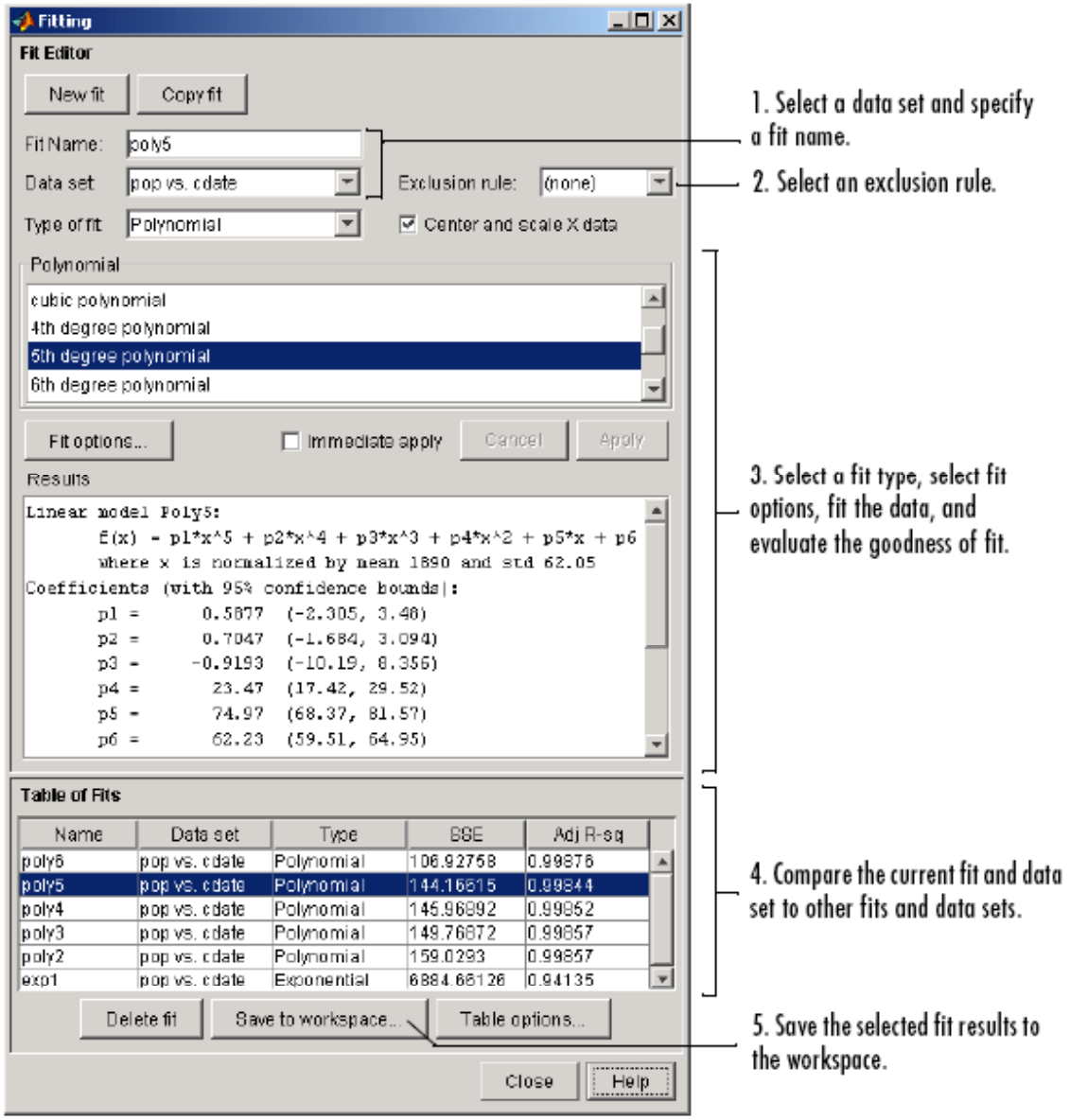
  - **Dobiti uvid u podatke** koje će omogućiti da se poboljša tehnika prikupljana podataka u budućim eksperimentima,
  - **prihvatiti** ili **odbiti** teorijski **model**,
  - iznaći **fizikalno značenje** za koeficijente i
  - izvući **zaključak o** izvornoj **populaciji**.

# KORACI PRI PODEŠAVANJU

- Provodi se u grafičkom korisničkom okruženju

- **Koraci**:

1. Odabir podataka i imena
2. Odabir pravila isključivanja
3. Izbor metode podešavanja i validacije
4. Usporedba rezultata podešavanja
5. Spremanje rezultata



# PARAMETARSKO PODEŠAVANJE

- Procjena vrijednosti koeficijenata (parametara) modela;
- Pretpostavlja se da su podaci uzeti statistički i da sadrže dvije komponente:

**podatak** = **deterministička komponenta** + **slučajna komponenta**

**podatak** = **fit** + **pogreška** (slučajna)

- ***fit*** predstavlja model koji je funkcija nezavisne varijable (prediktora);
- **Pogreška** je **slučajna varijacija** koja slijedi raspodjelu vjerojatnosti (obično *Gaussovu*);
- Podaci mogu sadržavati i sustavno odstupanje, ali se ono teško kvantificira;
- Podešeni koeficijenti mogu imati fizikalno značenje (npr. kod radioaktivnog raspada $T_{1/2}$ (vrijeme poluraspada), obično eksponencijalni odziv s obzirom na vrijeme:

$$y = y_0 e^{-\lambda r} \qquad \text{podatak} = y_0 e^{-\lambda t} + pogreška$$

# PROVJERA VALJANOSTI (VALIDIRANJE)

## VALIDATION

- Određivanje najboljeg od svih primjenjenih podešavanja;
- Da bi odredili najbolje podešavanje, potrebno je pregledati rezultate podešavanja **grafički** i **numerički**
- Početno se obično pregledavaju grafički prikaz rezultata i ostatka (**residual**);
- Numerička provjera na dva načina:

  - ***Goodnes of fit statistics***
    (kako dobro krivulja slijedi podatke)
  - ***Coefficient confidence intervals***
    (interval pouzdanosti pri određivanju koeficijenta)

# PROVJERA VALJANOSTI (VALIDIRANJE)

## Residual

- Razlika između odziva y i modelom predviđenog odziva

$$r = y - \hat{y}$$

- Rezidui aproksimiraju slučajne pogreške;
- Ako se rezidui na grafičkom prikazu vladaju slučajno to govori da model dobro opisuje podatke;
- Ako se javlja sustavno odstupanje model nije dobar.

***Goodness of fit statistics*** obuhvaća:

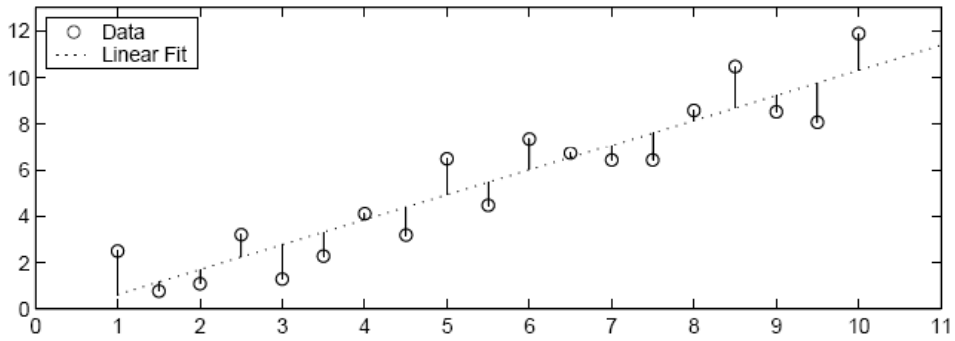
**Suma kvadrata pogreške**
(**SSE** – **the sum of squares due to error**)

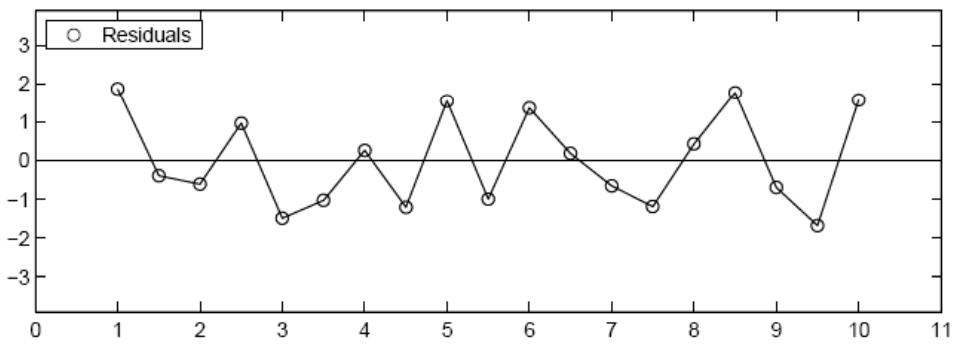
**R-kvadrat** (**R-square**)

**Podešeni R-kvadrat** (**adjusted R-square**)

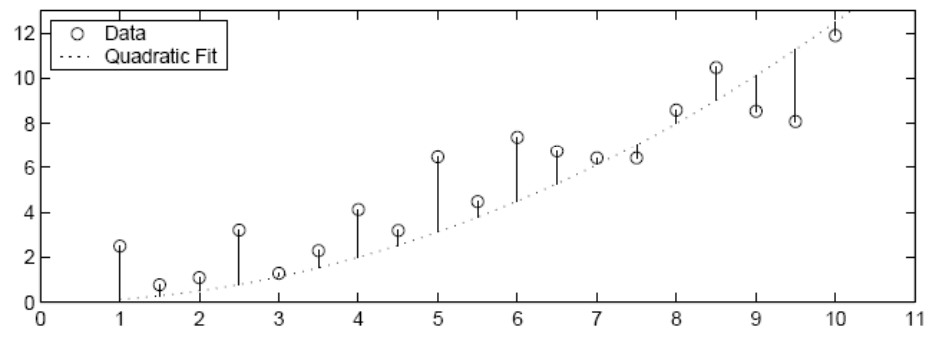
**Korijen kvadrata srednje vrijednosti**
(**RMSE** – **root mean squared error**)

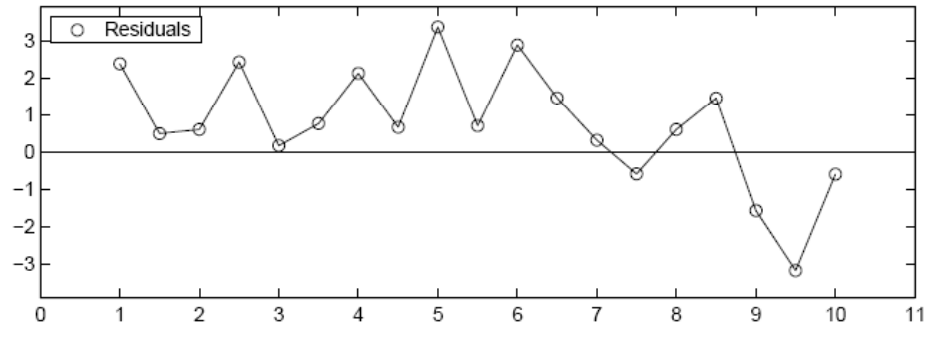
# PROVJERA VALJANOSTI (VALIDIRANJE)




Rezidui su slučajno raspodijeljeni oko nule što znači da model dobro opisuje podatke




Rezidui su pozitivni za većinu podataka što znači da model ne opisuje dobro podatke

# NEPARAMETARSKO PODEŠAVANJE

- Primjenjuje se kada nije potrebno određivati parametre ili ih nije potrebno interpretirati;
- Kroz podatke se provlači glatka krivulja;
- U **CFT**-u postoje dvije metode:

  - ***Interpolants***
    procjenjuje vrijednosti koje leže između podataka

  - ***Smoothing spline***
    provlači glatku krivulju koja prolazi kroz podatke

- Razina "izglađivanja" može se podesiti **promjenom parametara** koji korigiraju krivulju od ravne linije (aproksimacija metodom najmanjih kvadrata) do cubic *spline* interpolanta.

# IZGLAĐUJUĆI SPLINE (SMOOTHING SPLINE)

- Ako su podaci **opterećeni šumom** dobro je primijeniti **izglađujući spline**;
- Karakteriziraju ga parametar *p* i težine $w_i$
- Minimizira se:

$$p\sum w_i(y_i - s(x_i))^2 + (1-p)\int (\frac{d^2 s}{dx^2})^2 dx$$

$$0 < p < 1$$

$p = 0$ → *least square straight line fit*

$p = 1$ → *cubic spline interpolant*

- Ako težine nisu definirane pretpostavlja se da su jednake 1;
- Ako se ne definira *smoothing* parametar, automatski se odabire u "*interesting range*"-u $1/(1+h^3/6)$ pri čemu je *h* razmak između točaka
- Pošto *smoothing spline* ima parametar može ga se smatrati i parametarskim, no on je i *piecewise* polinomski kao npr. *cubic spline* ili *shape-preserving interpolants*, no u **CFT**-u se smatra neparametarskim.