Slučajni procesi
Vježbe 2

14. oktobar 2022.

Linearna regresija i *ridge* regresija

1. Numerički implementirati sljedeći zadatak. Razmatra se $N = 20$ slučajnih odbiraka signala $x(t_n)$

$$[x(t_1), x(t_2), \ldots, x(t_N)]^T = [6.10, 3.09, 3.23, 6.90, 3.53, 3.67, 3.64, 3.97, 3.85, 4.08, \\ 3.68, 4.05, 4.30, 4.27, 4.90, 4.56, 4.75, 1.80, 4.57, 2.9]$$

uzetih u odgovarajućim slučajnim trenucima t_n

$$[t_1, t_2, \ldots, t_N]^T = [0.022, 0.034, 0.200, 0.303, 0.307, 0.376, 0.429, 0.443, 0.519, 0.525, \\ 0.538, 0.598, 0.704, 0.715, 0.837, 0.841, 0.899, 0.910, 0.953, 0.954].$$

   (a) Odrediti parametre a i b linearnog regresionog modela za posmatrani skup podataka.

   (b) Nakon toga, primijeniti RANSAC pristup na sljedeći način:

   - Uzeti $S = 4$ (slučajno) odabranih podataka $n \in \mathbb{S}$.
   - Odrediti parametre linearnog regresionog modela za ovaj podskup podataka.
   - Ukoliko je broj tačaka unutar opsega definisanog rastojanjem $d = 0.25$ od linije modela manji od unaprijed zadate vrijednosti $T = 10$, odabrati novi (slučajni) podskup $\mathbb{S}$ i postupak ponavljati dok se ne dobije željeni rezultat.
   - Ukoliko broj tačaka unutar opsega od interesa nije ispod $T = 10$, iskoristiti sve dostupne tačke (podatke) koji su unutar posmatranog opsega da određivanje parametara linearnog regresionog modela.

Rješenje:

```
clear all, clc, close all

load sp_v2_z1.mat %ucitavamo podatke iz .mat fajla
% x i t su vektori-vrste u sacuvanom fajlu
% (a)
x=x.';
T=[ones(length(tn),1),tn.'];
ba=pinv(T)*x;
a=ba(2);
b=ba(1);


%vizuelizacija rezultata pod (a) i zadatih mjerenja:
figure(1)
plot(tn,x,'k.','markersize',12)
```

```
hold on
t=0:0.01:1;
xa=a*t+b'
plot(t,xa,'.')
legend('Podaci','Regresioni model')

% (b) RANSAC algoritam

d=0.25; %zadata udaljenost
Th=10; % prag
dn=0; % inicijalizacija broja elemenata skupa D
it=0; % brojac iteracija
S=4; %broj elemenata podskupa koji RANSAC uzima
ITmax=100;


while dn<Th && it < ITmax
    it=it+1;
    poz=randperm(length(x),S);
    xtn=x(poz);
    ttn=tn(poz);
    T=[ones(S,1),ttn'];
    ba=pinv(T)*xtn;
    a=ba(2);
    b=ba(1);

    Dn=find(abs(a*tn+b-x')/sqrt(1+a.^2)<d);
    dn=length(Dn);
end

xtn=x(Dn);
ttn=tn(Dn);
T=[ones(dn,1),ttn'];

figure(2)

plot(tn,x,'k.','markersize',12)
hold on

t=0:0.01:1;
xa=a*t+b';
plot(t,xa,'.')
legend('Podaci','Regresioni model dobijen RANSAC-om')
```

2. Snimljeno je $N = 10$ uzoraka slučajne promjenljive $x(n)$:

$$\mathbf{x} = [0.26,\ 0.31, 0.64,\ 0.99, 1.00,\ 0.92,\ 0.85,\ 0.73, 0.58, 0.15]^T,$$

sa različitim vrijednostima nezavisne slučajne promjenljive, t_n, čije su vrijednosti date u formi vektora:

$$\mathbf{t} = [-0.8, -0.83, -0.60, -0.10, -0.01,\ 0.28,\ 0.39, 0.52,\ 0.65,\ 0.92]^T.$$

Slučajna varijabla $x(n)$ je modelovana korišćenjem polinoma petog reda

$$x(n) = a_0 + a_1 t_n + a_2 t_n^2 + a_3 t_n^3 + a_4 t_n^4 + a_5 t_n^5$$

odnosno, u matričnoj formi:

$$\mathbf{x} = \mathbf{T}\mathbf{a},$$

gdje je $\mathbf{T}$ matrica sa kolonama $\mathbf{t}^m$, odnosno $\mathbf{T} = [\mathbf{t}^0,\ \mathbf{t}^1, \ldots, \mathbf{t}^5]$, a $\mathbf{t}^m$ označava vektor-kolonu sa elementima t_n^m, $n = 1, 2, \ldots, N$.

(a) Estimirati parametre modela korišćenjem polinomijalne aproksimacije (rješenje minimizacije $J(\mathbf{a}) = ||\mathbf{x} - \mathbf{T}\mathbf{a}||_2^2$ u smislu najmanjih kvadrata),

$$\hat{\mathbf{a}} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{x}. \tag{1}$$

(b) Estimirati parametre modela sa *ridge* regresijom (rješenje minimizacije funkcije $J(\mathbf{a}) = ||\mathbf{x} - \mathbf{T}\mathbf{a}||_2^2 + \lambda||\mathbf{a}||_2^2$) u formi

$$\hat{\mathbf{a}} = (\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T\mathbf{x}, \tag{2}$$

sa $\lambda = 0.1$.

(c) Ponoviti proračune iz (a) i (b) sa povećanim aditivnim šumom u podacima:

$$\mathbf{x} = [0.35,\ 0.33,\ 0.57,\ 0.92,\ 0.94,\ 0.89,\ 0.87,\ 0.86,\ 0.44,\ 0.29]^T.$$

(d) Predvidjeti vrijednost $x(1.12)$ u svim razmatranim slučajevima. Koristiti rezultat iz (a) kao referentni.

(e) Odrediti sistematsku grešku (*bias*) i matricu kovarijanse estimatora *ridge* regresije kao funkciju od λ, kada je šum u podacima $\mathbf{s}$ bijeli šum, sa varijansom σ_ε^2. pri čemu je razmatrani signal $\mathbf{x} = \mathbf{s} + \varepsilon$ (*napredna tema*).

Rješenje

(a) Parametri modela, koji se dobijaju kao rješenje minimizacije funkcije $J(\mathbf{a}) = ||\mathbf{x} - \mathbf{T}\mathbf{a}||_2^2$, za podatke

$$\mathbf{t} = [-0.8, -0.83, -0.60, -0.10, -0.01,\ 0.28,\ 0.39, 0.52,\ 0.65,\ 0.92]^T$$

i

$$\mathbf{x} = [0.26,\ 0.31, 0.64,\ 0.99, 1.00,\ 0.92,\ 0.85,\ 0.73, 0.58, 0.15]^T.$$

dati su izrazom: (pogledati definiciju matrice $\mathbf{T}$ sa predavanja)

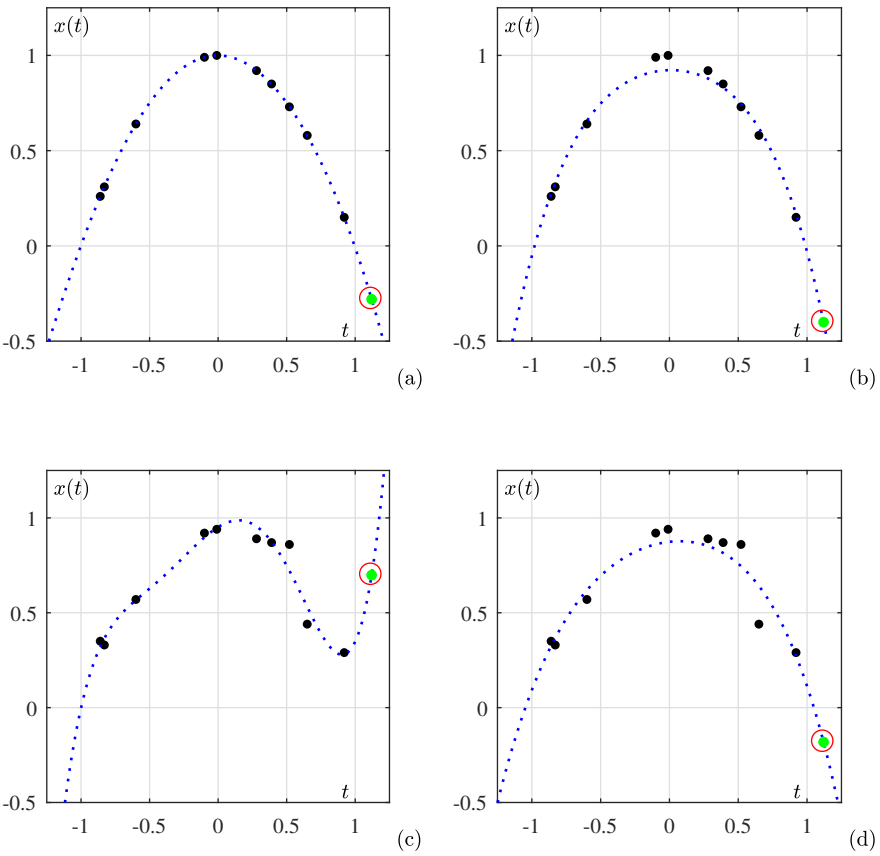
$$\hat{\mathbf{a}} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{x} = [0.9997, -0.0043, -0.9926,\ 0.0275, -0.0107, -0.0293]^T. \tag{3}$$

Estimirani model je prikazan na slici 1(a) tačkastom linijom. Budući da je šum mali (izazvan zaokruživanjem podataka na dvije decimale), model se dobro poklapa sa podacima.

(b) Kada se doda regularizaciona konstanta, rješenje minimizacije funkcije $J(\mathbf{a}) = ||\mathbf{x} - \mathbf{T}\mathbf{a}||_2^2 + \lambda||\mathbf{a}||_2^2$, za *ridge* regresiju se dobija u obliku:

$$\hat{\mathbf{a}} = (\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T\mathbf{x} = [0.9225,\ 0.0133, -0.5775,\ 0.0005, -0.3871, -0.0028]^T. \tag{4}$$

U ovom slučaju, može se uočiti mala sistematska greška u rezultatima, slika 1(b).



Slika 1: Primjer polinomijalne aproksimacije. (a) Podaci sa malim šumom (crne tačke) i polinomijalna aproksimacija sa rješenjem koje je optimalno u smislu najmanjih kvadrata. (b) Podaci sa malim šumom i polinomijalna aproksiamcija sa regresionim modelom uz korišćenje regularizacione konstante $\lambda = 0.1$, koja izaziva malu devijaciju u podacima. (c) Podaci sa umjerenim šumom i polinomijalna aproksimacija sa rješenjem koje je optimalno u smislu najmanjih kvadrata. Šum izaziva značajne devijacije i preprilagođen (*over-fitted*) model. (d) Podaci sa umjerenim šumom i polinomijalna aproksimacija sa regresionim modelom uz korišćenje regularizacione konstante $\lambda = 0.1$, gdje se energija u svim koeficijentima modela održava niskom. Predviđena vrijednost $x(1.12)$ je označena kružićem.

(c) Kada je u podacima prisutan aditivni šum većeg intenziteta:

$$\mathbf{x} = [0.35,\ 0.33,\ 0.57,\ 0.92,\ 0.94,\ 0.89,\ 0.87,\ 0.86,\ 0.44,\ 0.29]^T,$$

rješenja u smislu najmanjih kvadrata i rješenja *ridge* regresije su:

$$\hat{\mathbf{a}} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{x} = [0.9515,\ 0.4443, -1.1449, -1.6335,\ 0.3664, 1.3640]^T \qquad (5)$$

i

$$\hat{\mathbf{a}} = (\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T\mathbf{x} = [0.8740,\ 0.0795, -0.5277, -0.0622, -0.2434, -0.0068], \quad (6)$$

respektivno. Rezultati modela su dati na slici 1(c) i (d), respektivno. Koeficijenti višeg reda modela $\hat{\mathbf{a}}$ su veći u rješenju kada se regularizacija ne koristi.

(d) Predviđena vrijednost $x(1.12)$ je u svim razmatranim slučajevima označena kružićima. Može se vidjeti da umjereni šum izaziva značaje devijacije (*over-fitting*) od rezultata, slika 1(c), ukoliko se regularizacija ne koristi. U slučaju sa veoma malim šumom, regulatizacija malo pogoršava rezultate, unoseći sistematsku grešku (*bias*), što se može vidjeti na slici 1(b).

(e) *Napredna tema:* Pretpostavimo da se razmatrani signal sastoji od pravih podataka, $\mathbf{s}$ i šuma ε, odnosno $\mathbf{x} = \mathbf{s} + \varepsilon$. Parametri pravog modela, $\mathbf{Ta} = \mathbf{s}$, predstavljaju rješenje problema minimizacije u smislu najmanjih kvadrata, $J(\mathbf{a}) = ||\mathbf{s} - \mathbf{Ta}||_2^2$, odnosno

$$\mathbf{a} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{s}.$$

Sistematska greška (*bias*) u estimiranom modelu ridge regresije dobija se iz

$$\hat{\mathbf{a}} = (\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T\mathbf{x} = (\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T(\mathbf{s} + \varepsilon),$$
$$\mathrm{E}\{\hat{\mathbf{a}}\} = (\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T\mathbf{s}, \ \text{ as}$$
$$\mathrm{bias}(\hat{\mathbf{a}}) = \mathrm{E}\{\hat{\mathbf{a}}\} - \mathbf{a} = \Big((\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1} - (\mathbf{T}^T\mathbf{T})^{-1}\Big)\mathbf{T}^T\mathbf{s}.$$

Za $\lambda = 0$, estimator je bez sistematske greške, $\mathrm{bias}(\hat{\mathbf{a}}) = 0$, dok za veliko λ sistematska greška raste ka $|\mathrm{bias}(\hat{\mathbf{a}})| = |(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{s}|$.

Matrica kovarijanse estimatora je, po definiciji,

$$\mathrm{Cov}(\hat{\mathbf{a}}) = \mathrm{E}\{(\hat{\mathbf{a}} - \mathrm{E}\{\hat{\mathbf{a}}\})(\hat{\mathbf{a}} - \mathrm{E}\{\hat{\mathbf{a}}\})^T\} = \mathrm{E}\{\Big((\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T\varepsilon\Big)\Big((\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T\varepsilon\Big)^T\}$$
$$= \sigma_\varepsilon^2\Big((\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T\Big)\Big((\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T\Big)^T = \sigma_\varepsilon^2(\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^T\mathbf{T}(\mathbf{T}^T\mathbf{T} + \lambda\mathbf{I})^{-1}.$$

since $\mathrm{E}\{\varepsilon\varepsilon^T\} = \sigma_\varepsilon^2\mathbf{I}$.

Matrica kovarijanse je za $\lambda = 0$ data sa $\mathrm{Cov}(\hat{\mathbf{a}}) = \sigma_\varepsilon^2(\mathbf{T}^T\mathbf{T})^{-1}$. Njeni elementi se smanjuju i teže ka 0 kada se λ povećava.

Očigledno, postoji optimalna vrijednost parametra λ, kada je tzv. kompromis između sistematske greške i varijanse (engl. *bias-to-variance trade off*), $||\mathrm{bias}(\hat{\mathbf{a}})||_2^2 + \mathrm{Trace}(\mathrm{Cov}(\hat{\mathbf{a}}))$, minimalan. $\mathrm{Trace}(\mathrm{Cov}(\hat{\mathbf{a}}))$ predstavlja sumu elemenata na dijagonali (varijansi) matrice kovarijanse.

Metod kros-validacije. U praksi, određivanje najbolje vrijednosti parametra λ predstavlja složen problem. Jedan pristup određivanju najbolje vrijednosti λ je poznat pod nazivom metod kros-validacije, gdje:

(a) Estimacija parametara modela se radi sa unaprijed odabranim skupom vrijednosti λ, gdje se jedna ili više tačaka $(t_n, x(n))$ isključuje iz razmatranja.

(b) Estimacija se ponavlja, ali uz isključivanje ostalih tačaka, jedne po jedne.

(c) Računa se ukupna kvadratna greška vrijednosti predviđenih modelom u odnosu na izostavljene vrijednosti, za svaku razmatranu vrijednost λ.

(d) Najbolja vrijednost λ je ona koja daje najmanju ukupnu kvadratnu grešku predikcije.

3. Napisati kod kojim se numerički implementiraju glavne tačke iz prethodnog zadatka.

Rješenje

```
clear all, clc, close all

load 'sp_v2_z2.mat'  % ucitavamo podatke iz fajla
T=[ones(size(t)),t,t.^2,t.^3,t.^4,t.^5];

figure(1)

subplot(2,2,1)
plot(t,x,'k.','markersize',10)
a=pinv(T)*x;
tt=-1.3:0.05:1.3;
xm=a(1)+a(2)*tt+a(3)*tt.^2+a(4)*tt.^3+a(5)*tt.^4+a(6)*tt.^5;
hold on
plot(tt,xm,'--')
title('Pol. aproks. bez regularizacije')
xlabel('t')
tpred=1.12;
xpred=a(1)+a(2)*tpred+a(3)*tpred.^2+a(4)*tpred.^3+a(5)*tpred.^4+a(6)*tpred.^5;
plot(tpred,xpred,'or','markersize',10)

subplot(2,2,2)
plot(t,x,'k.','markersize',10)

lambda=0.1;
a=inv(T'*T+lambda*eye(size(T'*T)))*T'*x;

tt=-1.3:0.05:1.3;
xm=a(1)+a(2)*tt+a(3)*tt.^2+a(4)*tt.^3+a(5)*tt.^4+a(6)*tt.^5;
hold on
plot(tt,xm,'--')
title('Pol. aproks. bez regularizacije')
xlabel('t')

tpred=1.12;
xpred=a(1)+a(2)*tpred+a(3)*tpred.^2+a(4)*tpred.^3+a(5)*tpred.^4+a(6)*tpred.^5;
plot(tpred,xpred,'or','markersize',10)


subplot(2,2,3)
plot(t,x2,'k.','markersize',10)
a=pinv(T)*x2;

tt=-1.3:0.05:1.3;
```

```
xm=a(1)+a(2)*tt+a(3)*tt.^2+a(4)*tt.^3+a(5)*tt.^4+a(6)*tt.^5;
hold on
plot(tt,xm,'--')
title('Pol. aproks. bez regularizacije')
xlabel('t')
tpred=1.12;
xpred=a(1)+a(2)*tpred+a(3)*tpred.^2+a(4)*tpred.^3+a(5)*tpred.^4+a(6)*tpred.^5;
plot(tpred,xpred,'or','markersize',10)


subplot(2,2,4)
plot(t,x2,'k.','markersize',10)

lambda=0.1;
a=inv(T'*T+lambda*eye(size(T'*T)))*T'*x2;

tt=-1.3:0.05:1.3;
xm=a(1)+a(2)*tt+a(3)*tt.^2+a(4)*tt.^3+a(5)*tt.^4+a(6)*tt.^5;
hold on
plot(tt,xm,'--')
title('Pol. aproks. bez regularizacije')
xlabel('t')
tpred=1.12;
xpred=a(1)+a(2)*tpred+a(3)*tpred.^2+a(4)*tpred.^3+a(5)*tpred.^4+a(6)*tpred.^5;
plot(tpred,xpred,'or','markersize',10)
```

4. Posmatra se signal iz zadatka 2, ali sa dva *outlier*-a:

$$\mathbf{x} = [0.35,\ 0.33,\ 0.57,\ -2,\ 0.94,\ 0.89,\ 4.2,\ 0.86,\ 0.44,\ 0.29]^T,$$

u trenucima t_4 i t_7, vektora $\mathbf{t}$:

$$\mathbf{t} = [-0.8, -0.83, -0.60, -0.10, -0.01,\ 0.28,\ 0.39, 0.52,\ 0.65,\ 0.92]^T.$$

Prikazati odgovarajući signal iz zadatka 2 (c), gdje nije bilo outlier-a, kao i rezultat linearne regresije sa regularizacijom, dobijen u zadatku 2 (c). Zatim estimirati parametre polinoma petog reda kojim se aproksimiraju podaci sa outlier-ima, primjenom:

(a) Polinomijalne regresije bez regularizacije (korišćenjem polinoma petog reda).

(b) Polinomijalne regresije sa regularizacijom, sa $\lambda = 0.1$.

(c) RANSAC algoritma kojim se procjenjuju parametri polinomijalne regresije, ako se koristi isti polinom petog reda. Veličina skupa $\mathbb{S}$ je $S = 4$. Koristiti prag $T = 7$, i granično rastojanje od modela $d = 0.2$. Rastojanje d_n između zadate tačke modela $(\hat{x}(t_n)$ i odgovarajućeg podatka $x(t_n)$ se definiše kao $d_n = |\hat{x}(t_n) - x(t_n)|$ (Euklidsko rastojanje, jer obije tačke računamo za isto t_n).

U svim prethodnim slučajevima izvršiti predikciju vrijednosti $x(1.12)$. Regularizaciju vršiti sa $\lambda = 0.1$.

Rješenje

```
clear all, clc, close all
load 'sp_v2_z2.mat'
%x2 -- signal sa outlier-ima
T=[ones(size(t)),t,t.^2,t.^3,t.^4,t.^5];

figure(1)
subplot(2,2,1)
plot(t,x,'k.','markersize',10)
lambda=0.1;
a=inv(T'*T+lambda*eye(size(T'*T)))*T'*x;
tt=-1.3:0.05:1.3;
xm=a(1)+a(2)*tt+a(3)*tt.^2+a(4)*tt.^3+a(5)*tt.^4+a(6)*tt.^5;
hold on
plot(tt,xm,'--')
title('Signal bez outlier-a (lin. regresija)')
xlabel('t')

tpred=1.12;
xpred=a(1)+a(2)*tpred+a(3)*tpred.^2+a(4)*tpred.^3+a(5)*tpred.^4+a(6)*tpred.^5;
plot(tpred,xpred,'or','markersize',10)
ylim([-5,5]) % zbog boljeg poredjenja y osu svodimo na opseg od -5 do 5

subplot(2,2,2)
plot(t,x3,'k.','markersize',10)
a=pinv(T)*x3;

tt=-1.3:0.05:1.3;
xm=a(1)+a(2)*tt+a(3)*tt.^2+a(4)*tt.^3+a(5)*tt.^4+a(6)*tt.^5;
hold on
plot(tt,xm,'--')
title('Signal sa outlier-ima (bez regul.)')
xlabel('t')
tpred=1.12;
xpred=a(1)+a(2)*tpred+a(3)*tpred.^2+a(4)*tpred.^3+a(5)*tpred.^4+a(6)*tpred.^5;
plot(tpred,xpred,'or','markersize',10)

subplot(2,2,3)
plot(t,x3,'k.','markersize',10)
lambda=0.1;
a=inv(T'*T+lambda*eye(size(T'*T)))*T'*x3;

tt=-1.3:0.05:1.3;
xm=a(1)+a(2)*tt+a(3)*tt.^2+a(4)*tt.^3+a(5)*tt.^4+a(6)*tt.^5;
hold on
plot(tt,xm,'--')
title('Signal sa outlier-ima (sa regul.)')
xlabel('t')
tpred=1.12;
xpred=a(1)+a(2)*tpred+a(3)*tpred.^2+a(4)*tpred.^3+a(5)*tpred.^4+a(6)*tpred.^5;
plot(tpred,xpred,'or','markersize',10)

% RANSAC za polinomijalnu regresiju
```

```
x=x3;
tn=t;
S=4;
prag=7;
d=0.2;
it=0;
ITmax=100;
N=length(x);
D=0; % broj elemenata skupa D
lambda=0.1;
while  D<prag && it<ITmax
    it=it+1;
    poz=randperm(N,S);
    xp=x(poz);
    tp=tn(poz);
    Tp=[ones(size(tp)),tp,tp.^2,tp.^3,tp.^4,tp.^5];
    Aest=inv(Tp'*Tp+lambda*eye(size(Tp'*Tp)))*Tp'*xp
    model=Aest(1)+Aest(2)*tn+Aest(3).*tn.^2+Aest(4).*tn.^3...
    +Aest(5).*tn.^4+Aest(6).*tn.^5;
    DD=abs(x-model);
    DDp=find(DD<d);
    D=length(DDp);
end

xp=x(DDp);
tp=tn(DDp);
Tp=[ones(size(tp)),tp,tp.^2,tp.^3,tp.^4,tp.^5];
Aest=inv(Tp'*Tp+lambda*eye(size(Tp'*Tp)))*Tp'*xp;

subplot(2,2,4)
plot(tn,x,'k.','markersize',12)

tt=-1.3:0.05:1.3;
xm=Aest(1)+Aest(2)*tt+Aest(3).*tt.^2+Aest(4).*tt.^3+Aest(5).*tt.^4+Aest(6).*tt.^5;

hold on
plot(tt,xm,'--')
title('Signal sa outlier-ima (RANSAC)')
ylim([-5,5])

% predikcija
tpred=1.12;
xpred=Aest(1)+Aest(2)*tpred+Aest(3)*tpred.^2+...
    Aest(4)*tpred.^3+Aest(5)*tpred.^4+Aest(6)*tpred.^5;
plot(tpred,xpred,'or','markersize',10)
```