Slučajni procesi

Uvod. Osnovne statističke definicije. Regresiona analiza

Ljubiša Stanković, Miloš Brajović

Univerzitet Crne Gore
Elektrotehnički fakultet

Prezentacija 1

Osnovne informacije

Slučajni procesi

- Fond časova: **3P+1V**
- Predmetni nastavnik:
  - Doc. dr Miloš Brajović (milosb@ucg.ac.me)
- Provjere znanja i raspodjela poena:
  - Dva kolokvijuma (svaki sa po maksimalno **20** poena)
  - Testovi na DL platformi (maksimalno **10** poena)
  - Završni ispit (maksimalno **50** poena)

Literatura

- LJ. Stanković, *Digital Signal Processing with Selected Topics*, CreateSpace Independent Publishing Platform, An Amazon.com Company
- LJ. Stanković, M. Brajović, *Teorija slučajnih procesa sa elementima statistike i vjerovatnoće*, skripta, 2023.
- A. Papoulis, and S. Pillai, *Probability, Random Variables, and Stochastic Processes*, McGraw Hill, Boston, Fourth edition, 2002.

Uvod. Osnovne statističke definicije: srednja vrijednost, medijan, varijansa i standardna devijacija

Ljubiša Stanković, Miloš Brajović

Univerzitet Crne Gore
Elektrotehnički fakultet

Prezentacija 1

Uvod

- Slučajne vrijednosti signala ne mogu biti definisane jednostavnim determinističkim matematičkim funkcijama, već se opisuju **stohastičkim alatima**.
- **Slučajni proces** je skup slučajnih promjenljivih koje su *indeksirane* u odnosu na nezavisnu promjenjivu, uobičajeno vrijeme.
- Skup vrijednosti nezavisno promjenljivih se naziva **indeksni set**. Slučajni proces u tom smislu možemo posmatrati i **kao skup slučajnih funkcija** odnosno **skup slučajnih signala**, gdje se vrijeme najčešće uzima kao nezavisno promjenljiva.
- Nezavisno promjenljiva slučajnih signala može uzimati diskretne vrijednosti, kada kažemo da se radi o **vremenski diskretnom slučajnom procesu**, odnosno o skupu diskretnih slučajnih signala, $\{x(n)\}$, gdje je n cio broj.
- Ako je indeksni set definisan intervalom sa mogućim kontinualnim vrijednostima nezavisno promjenljive onda se radi o **vremenski kontinualnom slučajnom procesu**, $\{x(t)\}$, gdje je $t \in \{\mathbb{T}\}$ unutar intervala $\mathbb{T}$.
- Primijetimo da, za neku određenu vrijednost nezavisne promjeljive, n ili t, vrijednosti slučajnog procesa predstavljaju ustvari **slučajne promjenljive**.
- Skup vrijednosti koje uzma slučajna promjenljiva naziva se i **prostor stanja** slučajnog procesa.
- Ako je i skup vrijednosti koje uzima slučajni signal diskretan, onda imamo slučajni proces koji odgovara formi **digitalnog slučajnog signala**.

Uvod

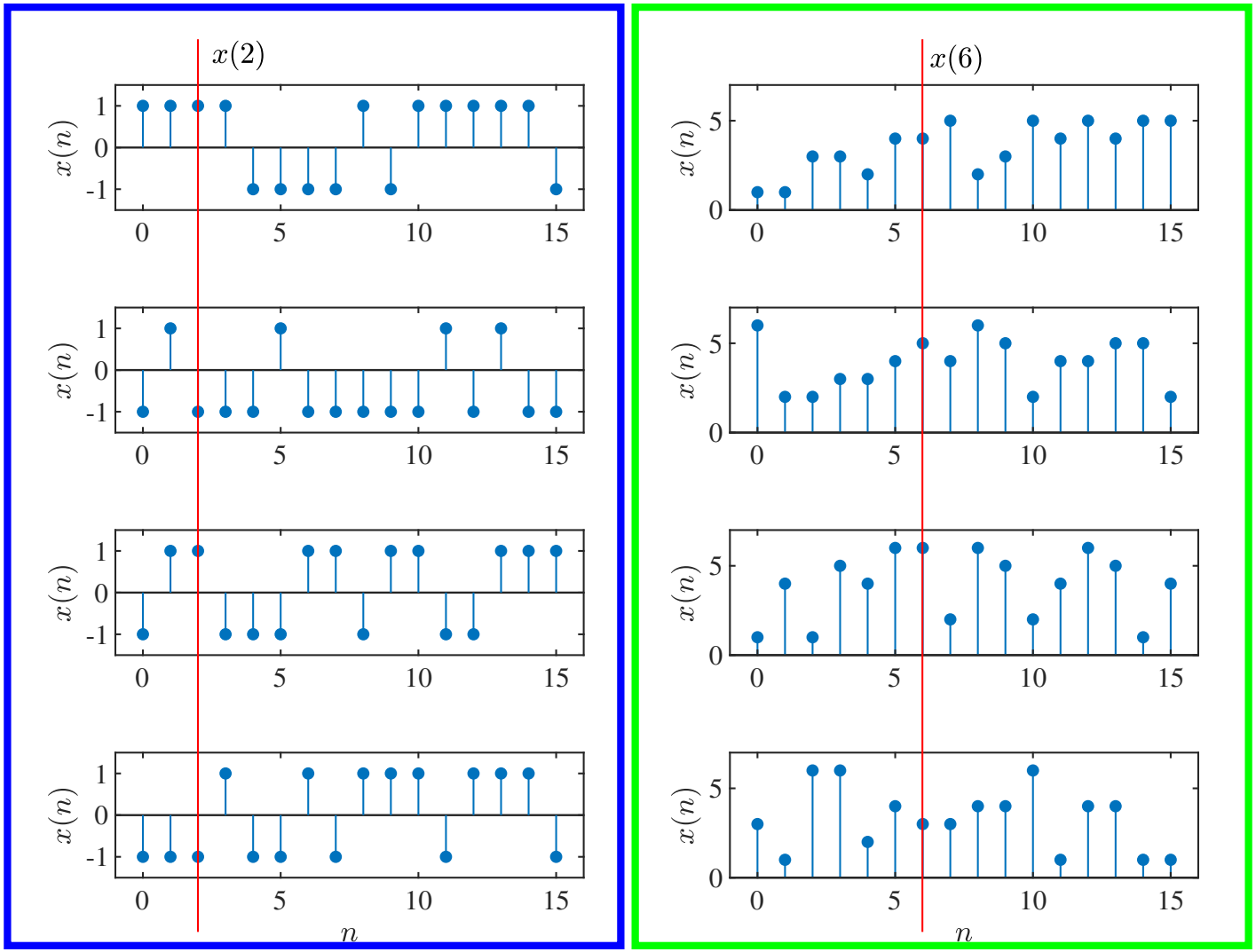
- Jedan mogući ishod slučajnog procesa naziva se **realizacija ili uzorak** slučajnog procesa, odnosno **uzorak slučajnog signala**, $x(n)$ ili $x(t)$.
- U praksi su posebno važni vremenski diskretni slučajni procesi, i oni će biti dominantno analizirani u sklopu ovog kursa.
- Za vremenski kontinualne slučajne procese, indeksni set je kontinualni interval, koji je sam tim neprebrojiv i zahtijeva složeniji matematički aparat, pa ćemo samo povremeno analizirati jednostavnije primjere takvih signala.

Ilustracija slučajnih signala

Prethodne pojmove ćemo ilustrovati sa nekoliko primjera slučajnih procesa (skupova slučajnih signala).

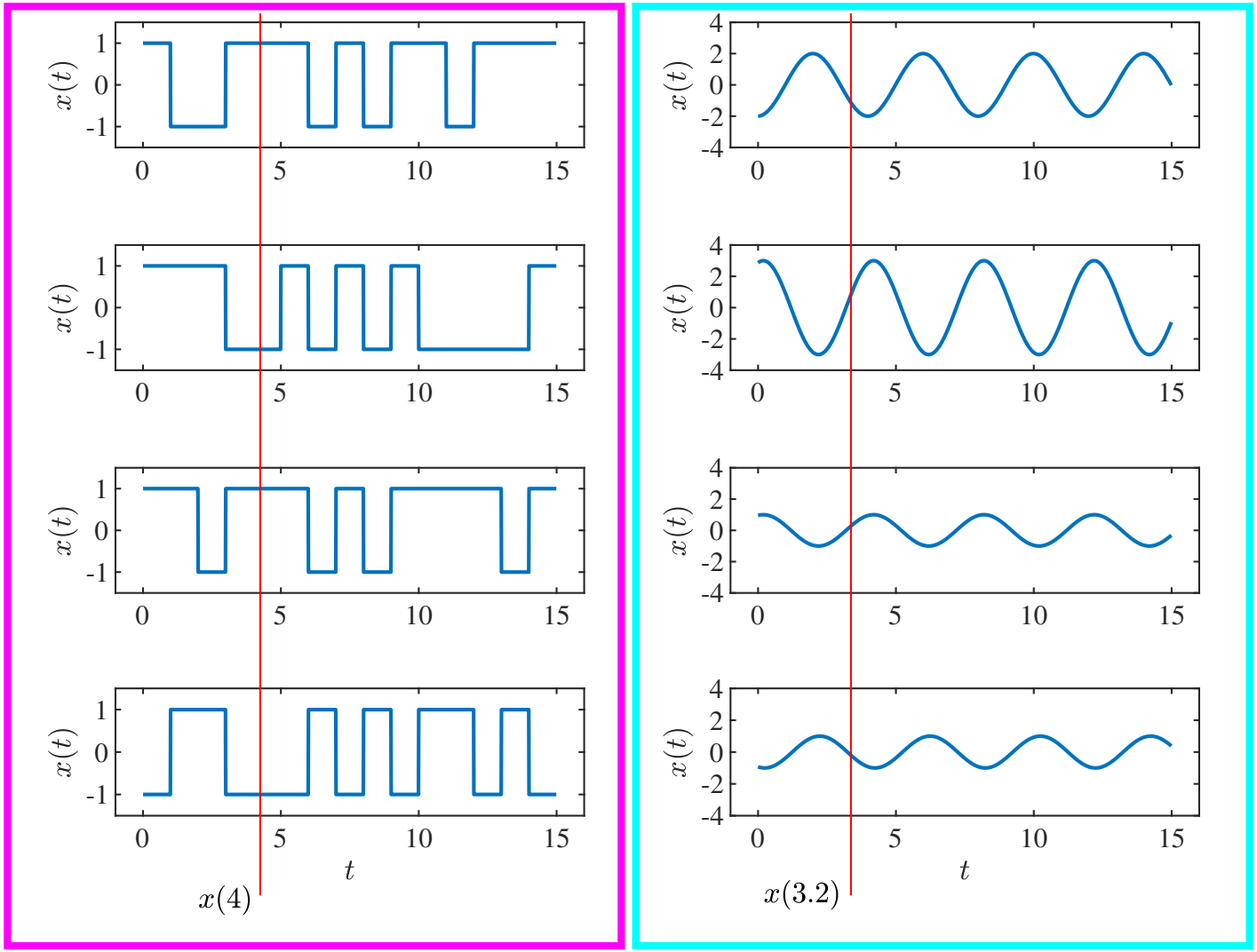
1. Kao primjer diskretnog slučajnog procesa posmatraćemo po četiri realizacije Bernulijevog procesa i četiri realizacije procesa gdje je vrijednost slučajnog signala jednaka broju koji se pojavljuje pri bacanju fer kocke.
2. Kao primer kontinualnog slučajnog procesa ćemo posmatrati sinusoidalni signal sa slučajnom početnom fazom, kao digitalni binarni signal sa kontinualnom diskretnom promjenljivom i dvije moguće vrijednosti amptude iz skupa od dvije vrijednosti.

Uvod



Slika: Po četiri realizacije slučajnih signala: Bernulijev slučajni signal (lijevo); broj koji se pojavljuje pri bacanju fer kocke (desno).

Uvod



Slika: Po četiri realizacije slučajnih signala: binarni signal (lijevo); sinusoide sa slučajnom amplitudom is skupa $\{1,2,3\}$ i slučajnom fazom (desno).

Osnovne statističke definicije

- Slučajne vrijednosti signala ne mogu biti definisane jednostavnim determinističkim matematičkim funkcijama, već se opisuju stohastičkim alatima.
- *Statistika* je nauka ili naučna praksa koja se bavi prikupljanjem, analizom, interpretacijom i prezentacijom numeričkih podataka, određivanjem parametera na osnovu cijelog skupa podataka ili reprezentativnog uzorka.
- Pojam *statistička vrijednost* podrazumijeva numeričku činjenicu dobijenu analizom razmatranog skupa podataka, koja se koristi za opis cijelog skupa podataka. **Statistika prvog reda** je obično polazna tačka u opisivanju slučajnih signala.
- **Srednja vrijednost** ili *prosječna vrijednost uzoraka* slučajnog signala je jedan od parametara statistike prvog reda. Ako imamo skup odbiraka signala,

$$\mathbb{X} = \{x(n) |\ \ n = 1, 2, \ldots, N\},$$

srednja vrijednost ovog skupa odbiraka se računa po formuli:

$$\hat{\mu}_x = \text{mean}\{x(n) |\ \ n = 1, 2, \ldots, N\} = \frac{1}{N}(x(1) + x(2) + \cdots + x(N)).$$

- U cilju jednostavnosti notacije, takođe će biti korišćeno i $\hat{\mu}_x = \text{mean}\{x(n)\}$, čime se označava srednja vrijednost skupa podataka $\{x(n)\}$, za sve indekse n, za koje je signal dostupan.

Srednja vrijednost – primjer

Primjer 1

Razmatra se slučajni signal $x(n)$, čija je jedna realizacija predstavljena donjom tabelom.

54	62	58	51	70	43	99	52	57	76
56	53	38	61	28	69	87	41	72	80
23	26	66	47	69	71	69	81	68	79
31	55	52	23	60	34	83	39	66	59
37	12	54	42	67	95	89	67	42	63
35	55	54	55	49	77	18	64	73	70
67	56	42	66	50	47	49	25	50	57
61	84	48	67	71	74	35	59	60	42
40	77	52	63	57	42	44	64	36	71
66	39	50	31	11	75	45	62	60	55

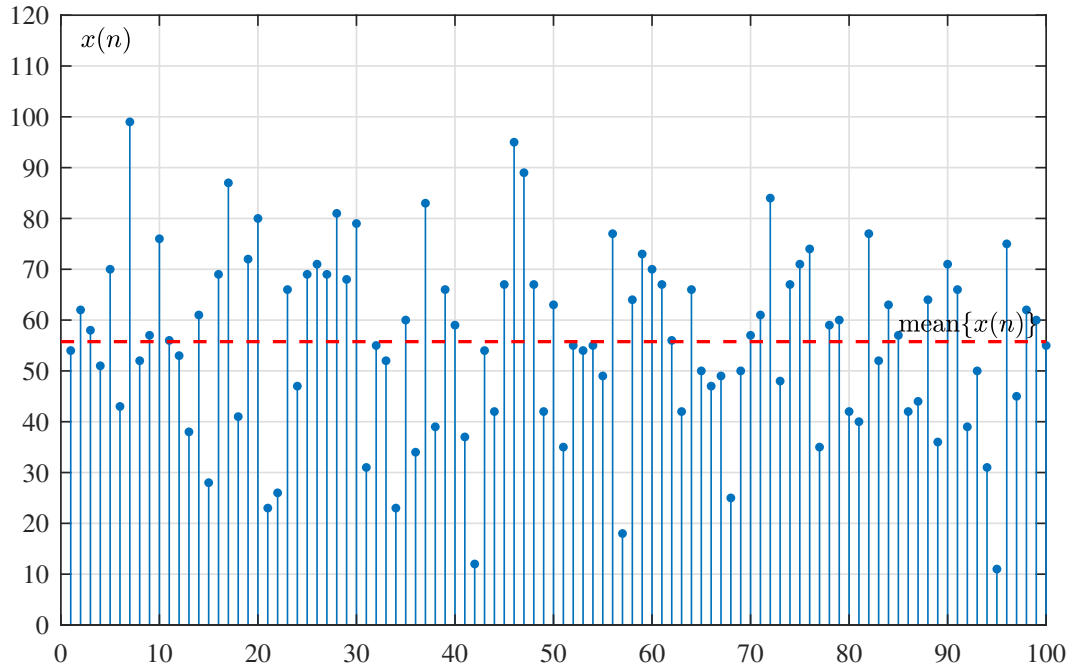
Odrediti srednju vrijednost ovog signala. Odrediti koliko se odbiraka signala nalazi u intervalima $[1,10]$, $[11,20]$,..., $[91,100]$. Grafički prikazati broj pojavljivanja vrijednosti signala $x(n)$ unutar ovih intervala, kao funkciju od opsega njihovih vrijednosti.

Primjer 1

- Realizacija razmatranog signala $x(n)$ je predstavljena na slici ispod. Srednja vrijednost odbiraka signala je

$$\hat{\mu}_x = \frac{1}{100}\sum_{n=1}^{100} x(n) = 55.76.$$

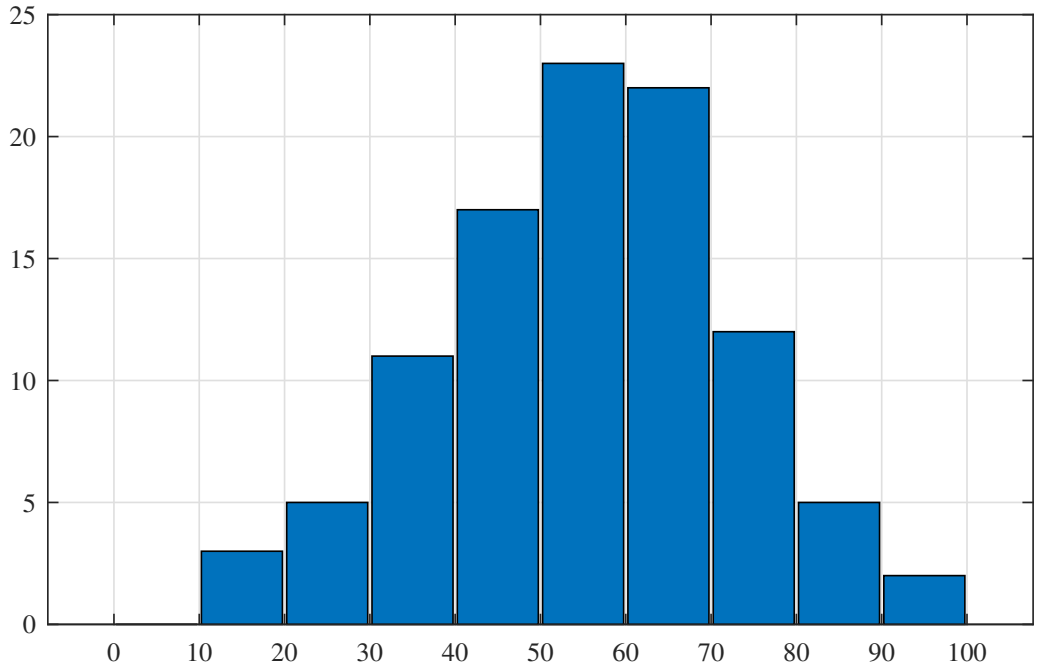
- Iz tabele ili njenog vizuelnog prikaza, može se zaključiti da, na primjer, ne postoji nijedna vrijednost signala unutar intervala $[1, 10]$.



Slika: Realizacija slučajnog signala $x(n)$

Primjer 1

- Unutar intervala $[11, 20]$ postoje dvije vrijednosti signala, ($x(42) = 12$ i $x(95) = 11$).
- Na sličan način se mogu prebrojati vrijednosti signala unutar preostalih intervala, čime se dobija rezultat prikazan na slici ispod.
- Ovakva vrsta reprezentacije slučajnog signala je poznata kao **histogram** signala $x(n)$, sa definisanim intervalima.



Slika: Histogram slučajnog signala $x(n)$, sa 10 intervala definisanih kao $[10i + 1, 10i + 10]$, $i = 0, 1, 2, \ldots, 9$.

Srednja vrijednost – primjer

Primjer 2

Pretpostaviti da se na osnovu signala $x(n)$ iz prethodnog primjera formira novi signal $y(n)$, u obliku:

$$y(n) = \text{int}\left\{\frac{x(n)+5}{10}\right\},$$

gdje $\text{int}\{\cdot\}$ označava najbliži cio broj. Navedeno znači da važi: $y(n) = 1$ za $1 \leq x(n) \leq 10$, $y(n) = 2$ za $11 \leq x(n) \leq 20$, $\ldots$, $y(n) = i$ za $10(i-1)+1 \leq x(n) \leq 10i$, sve do $i = 10$.

- Odrediti skup mogućih vrijednosti signala $y(n)$.
- Odrediti i grafički predstaviti broj pojavljivanja svih vrijednosti $y(n)$ u ovoj realizaciji signala. Odrediti srednju vrijednost novog signala $y(n)$ i diskutovati rezultat.

- Srednja vrijednost signala $y(n)$ je

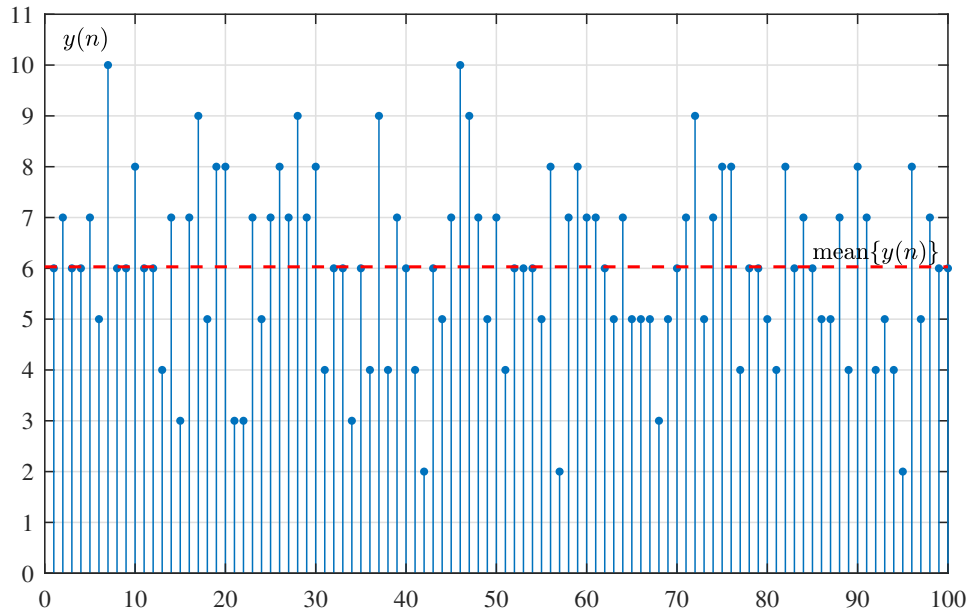
$$\hat{\mu}_y = \frac{1}{100}\sum_{n=1}^{100} y(n) = 6.13.$$

Primjer 2

- Signal $y(n)$ uzima vrijednosti iz skupa $\{2,3,4,5,6,7,8,9,10\}$ (slika ispod).
- Srednja vrijednost se može zapisati grupisanjem istih vrijednosti signala $y(n)$:

$$\hat{\mu}_y = \frac{1}{100}(1 \cdot n_1 + 2 \cdot n_2 + 3 \cdot n_3 + \cdots + 10 \cdot n_{10}) =$$
$$= 1 \cdot \frac{n_1}{N} + 2 \cdot \frac{n_2}{N} + 3 \cdot \frac{n_3}{N} + \cdots + 10 \cdot \frac{n_{10}}{N},$$

gdje $N = 100$ predstavlja ukupan broj dostupnih vrijednosti signala, dok n_i predstavlja broj koji pokazuje koliko se puta svaka vrijednost i pojavljuje u $y(n)$.



Slika: Slučajni signal $y(n)$

Primjer 2

- Ukoliko postoji dovoljan broj pojavljivanja svake vrijednosti (ishoda) i, tada se

$$P_y(i) = \frac{n_i}{N}$$

može smatrati **estimacijom vjerovatnoće** pojavljivanja vrijednosti i.

- U tom smislu, može se uočiti da se izraz za srednju vrijednost može zapisati u obliku:

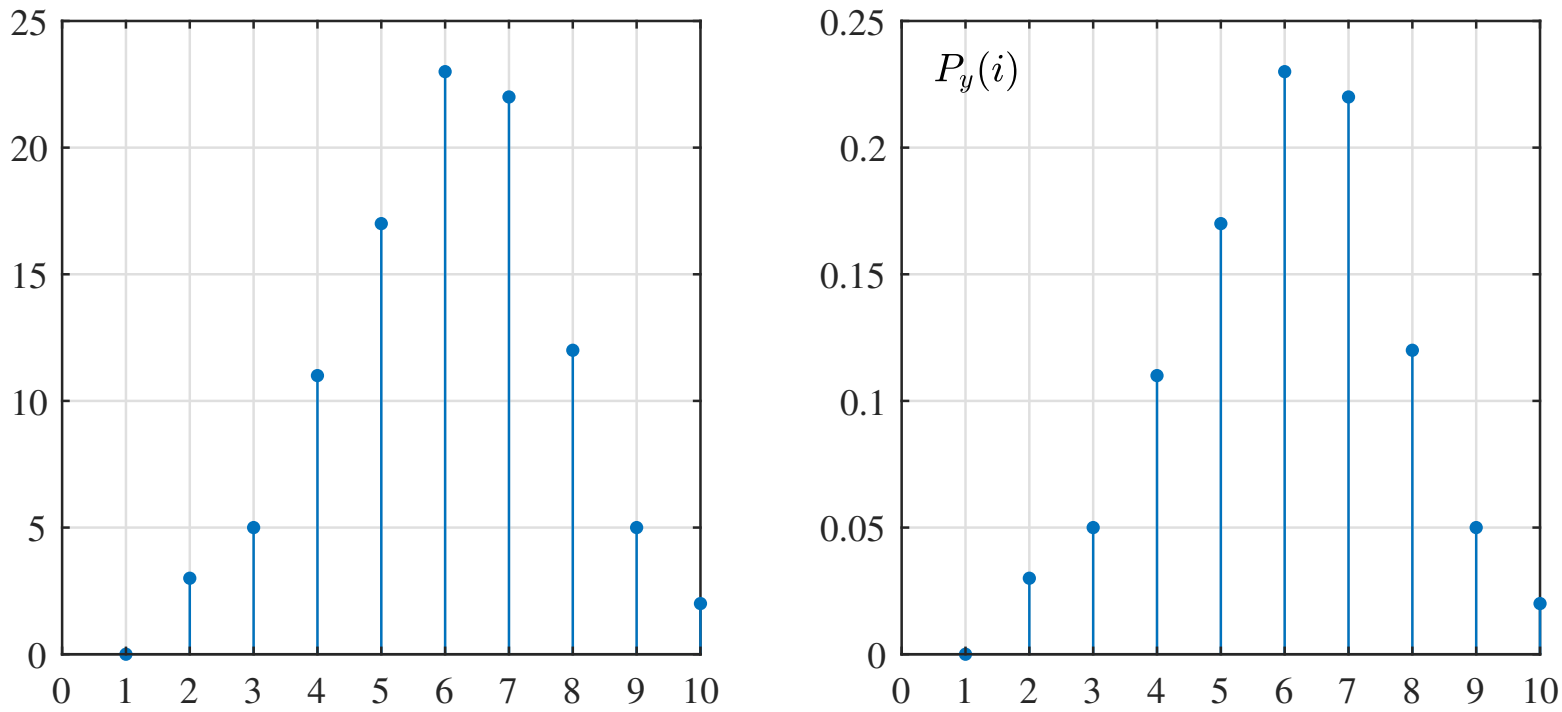
$$\hat{\mu}_y = 1 \cdot P_y(1) + 2 \cdot P_y(2) + 3 \cdot P_y(3) + \cdots + 10 \cdot P_y(10) = \sum_{i=1}^{10} y(i) P_y(i),$$

uz

$$\sum_{i=1}^{10} P_y(i) = \sum_{i=1}^{10} \frac{n_i}{N} = 1.$$

- Vrijednosti estimiranih vjerovatnoća $P_y(i)$ su prikazane na slici na narednom slajdu (desno).
- Za signal $y(n)$ se može, umjesto histograma, prikazati broj pojavljivanja svake vrijednosti koju $y(n)$ može uzeti, kao na slici na narednom slajdu (lijevo).

Primjer 2



Slika: Broj pojavljivanja svake moguće vrijednosti $y(n)$ (lijevo) i estimirane vjerovatnoće da slučajni signal $y(n)$ uzima vrijednosti $i = 1, 2, \ldots, 10$ (desno).

- Prethodni primjer ilustruje suštinsku vezu koja postoji između histograma (razmatrali smo signal sa cjelobrojnim vrijednostima) i estimiranih vjerovatnoća pojavljivanja odgovarajućih vrijednosti signala.

Srednja vrijednosti pojedinačnih odbiraka (zavisnost od n)

- U opštem slučaju, srednje vrijednosti **pojedinačnih odbiraka** signala se mogu međusobno razlikovati.
- Na primjer, ako vrijednosti signala predstavljaju najveću dnevnu temperaturu tokom godine, tada srednje vrijednosti veoma zavise od razmatranih odbiraka.
- U cilju računanja srednje vrijednosti temperature, neophodno je imati više realizacija ovih slučajnih signala (odnosno, mjerenja tokom M godina), koja ćemo označiti sa $\{x_i(n)\}$.
- Ovdje argument $n = 1, 2, 3, \ldots, N$ predstavlja redni broj dana u godini, dok je $i = 1, 2, \ldots, M$ indeks realizacije (indeks godine).
- Srednja vrijednost se tada računa na sljedeći način:

$$\hat{\mu}_x(n) = \frac{1}{M}(x_1(n) + x_2(n) + \cdots + x_M(n)) = \frac{1}{M}\sum_{i=1}^{M} x_i(n),$$

  za svako n.
- U ovom slučaju, postoji **skup** (odnosno signal) srednjih vrijednosti $\{\hat{\mu}_x(n)\}$, za $n = 1, 2, \ldots, 365$.

Srednja vrijednosti pojedinačnih odbiraka (zavisnost od n)

Primjer 3

Razmatra se signal $x(n)$ čije su realizacije date u tabeli ispod. Vrijednosti $x(n)$ su jednake mjesečnom prosjeku maksimalnih dnevnih temperatura u gradu mjerenih od $2001.$ do $2015.$ godine.

- Odrediti srednju vrijednost ovih temp. za svaki mjesec tokom razmatranog perioda godina.
- Koja je srednja vrijednost temperatura tokom svih mjeseci i godina?

Jan	Feb	Mar	Apr	Maj	Jun	Jul	Avg	Sep	Okt	Nov	Dec
10	4	18	17	22	29	30	28	27	17	17	5
6	7	11	23	22	32	35	33	22	26	22	8
10	11	10	16	21	26	32	31	23	19	17	4
3	11	13	19	22	26	34	29	26	22	12	9
7	10	13	21	27	29	30	34	24	20	16	11
7	11	17	17	27	25	37	34	33	22	14	14
7	12	13	19	23	32	34	38	21	21	12	10
12	5	9	20	21	37	34	34	27	22	20	7
7	12	13	23	27	33	29	31	25	21	6	11
8	12	10	17	27	33	38	32	23	20	15	9
8	10	13	24	23	33	33	31	27	21	16	8
4	6	15	18	25	26	27	33	23	23	13	11
3	6	16	17	27	28	30	32	29	24	12	10
11	12	14	18	22	29	34	34	23	21	20	11
6	13	8	22	22	29	30	34	23	18	15	8

Srednja vrijednosti pojedinačnih odbiraka (zavisnost od n)

- Razmatrani signal je za interval od $2001.$ do $2007.$ godine vizuelno predstavljen na narednim slajdovima.
- Srednja vrijednost temperature tokom posmatranog niza godina je za n-ti mjesec data izrazom:

$$\hat{\mu}_x(n) = \frac{1}{15}\sum_{i=01}^{15} x_{20i}(n),$$

gdje je iskorišćena simbolička notacija $20i$, koja označava niz 2001, 2002, $\ldots$, 2015, za $i = 01, 02, \ldots, 15$.

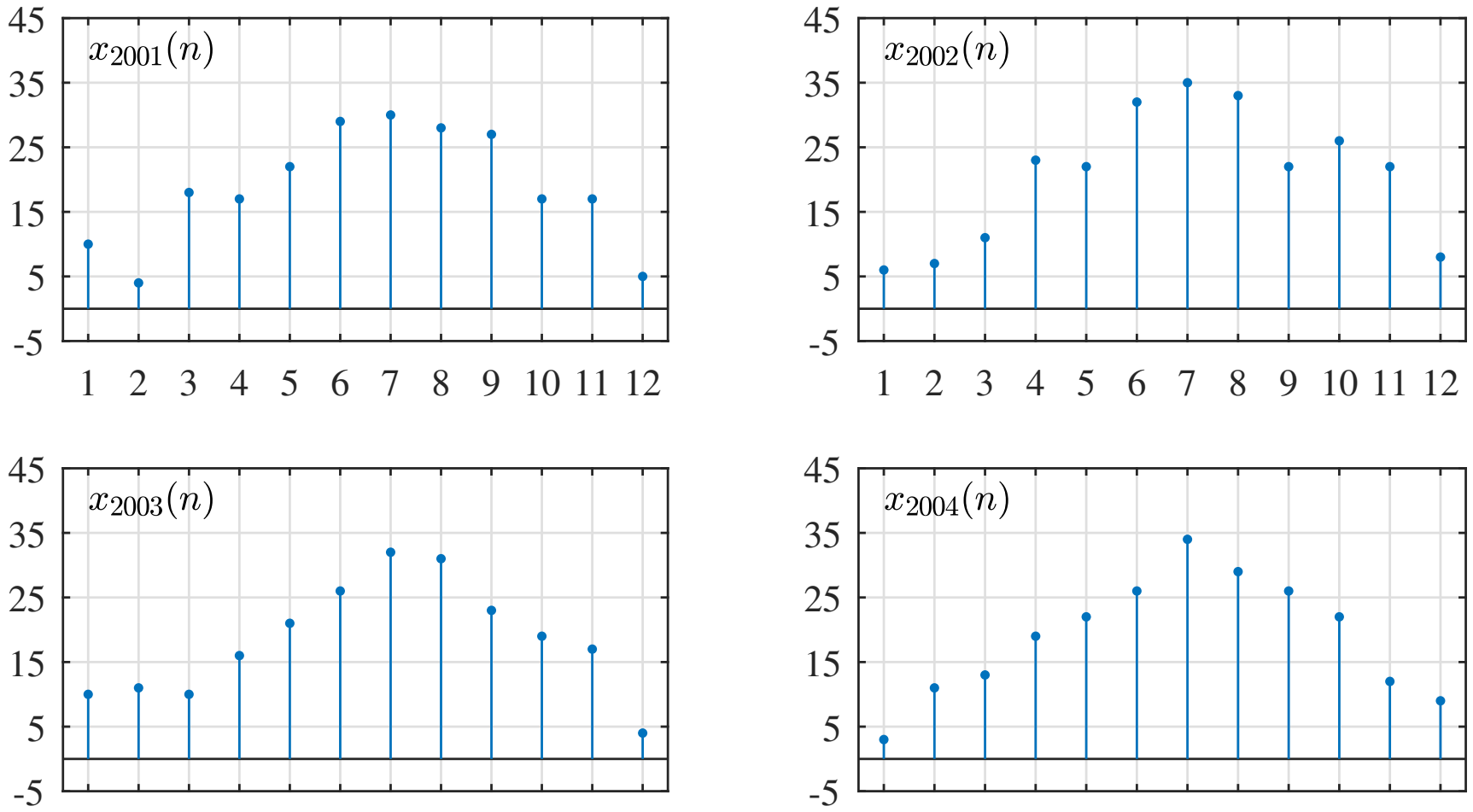
- Srednja vrijednost signala, $\hat{\mu}_x(n)$, prikazana je na posljednjem podgrafiku slike sa narednih slajdova.
- Srednja vrijednost računata za sve mjesece i godine je:

$$\hat{\mu}_x = \frac{1}{15 \cdot 12}\sum_{n=1}^{12}\sum_{i=01}^{15} x_{20i}(n) = 19.84.$$

Srednja vrijednost za svaku razmatranu godinu je:

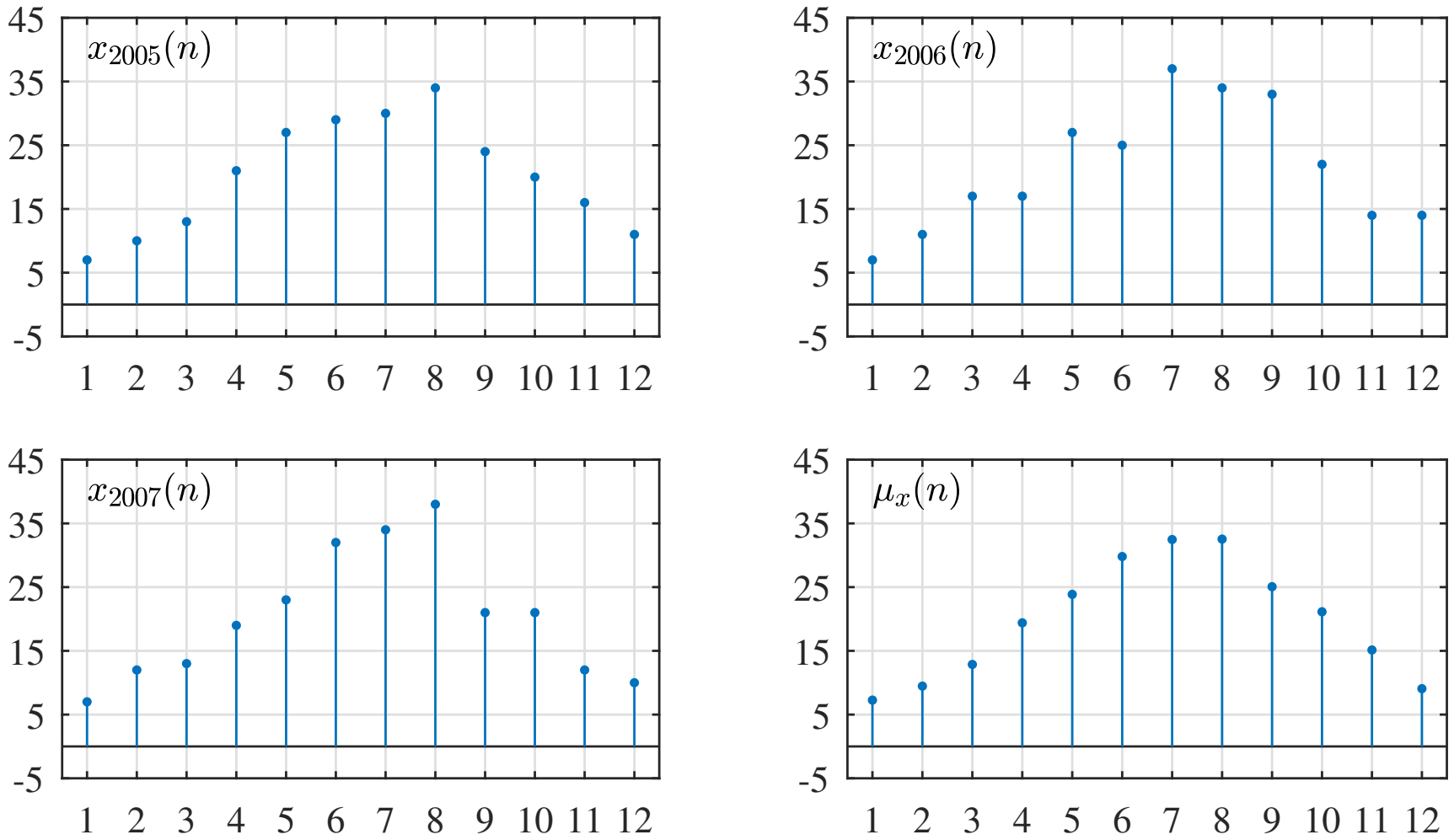
$$\hat{\mu}_x(20i) = \frac{1}{12}\sum_{n=1}^{12} x_{20i}(n).$$

Srednja vrijednosti pojedinačnih odbiraka (zavisnost od n)



Slika: Nekoliko realizacija slučajnog signala $x_{20i}(n)$, za $i = 01, 02, 03, 04$. Nastavak slike je dat na sljedećem slajdu

Srednja vrijednosti pojedinačnih odbiraka (zavisnost od n)



Slika: Nekoliko realizacija slučajnog signala $x_{20i}(n)$, za $i = 05, 06, 07$ i srednja vrijednost $\mu_x(n)$ za svaki odbirak (mjesec) tokom 15 dostupnih realizacija.

Srednja vrijednost kao prosječna vrijednost uzoraka

- Srednja vrijednost računata kao prosječna vrijednost odbiraka (uzoraka) se koristi vrlo često, zbog jednostavnosti procesa izračunavanja.
- Kasnije ćemo vidjeti da je prosjek odbiraka **optimalan estimator** *prave srednje vrijednosti signala*, kada su njegove realizacije oštećene vrlo čestim oblikom smetnje poznate pod nazivom Gausov šum (Gaus je inače uveo svoju poznatu distribuciju kao najbolju osnovu za estimaciju srednje vrijednosti uzoraka).
- Srednja vrijednost, računata kao prosjek odbiraka pomoću relacija sa prethodnih slajdova, predstavlja **rezultat sljedećeg minimizacionog problema**.
- Dat je skup slučajnih realizacija odbiraka $x(n)$, označen sa $\{x_i(n)\}$, gdje je $i = 1,2,\ldots,M$ indeks realizacije. Cilj je estimirati pravu srednju vrijednost signala $\mu(n)$ pomoću $\hat{\mu}(n)$, tako da je njena kvadratna razlika (devijacija) od dostupnih realizacija $x_i(n)$, $i = 1,2,\ldots,M$, minimalna. Ovaj zahtjev se može zapisati relacijom:

$$\hat{\mu}_x(n) = \min_{\alpha}\left((x_1(n)-\alpha)^2 + (x_2(n)-\alpha)^2 + \cdots + (x_M(n)-\alpha)^2\right)$$
$$= \min_{\alpha} ||\mathbf{x}(n)-\alpha||_2^2 = \min_{\alpha} f(\alpha),$$

gdje je $\mathbf{x}(n) = [x_1(n), x_2(n), \ldots, x_M(n)]^T$, dok $f(\alpha) = ||\mathbf{x}(n)-\alpha||_2^2$ predstavlja drugu (Euklidsku) normu vektora $\mathbf{x}(n)-\alpha$.

Srednja vrijednost kao prosječna vrijednost uzoraka

- Rezultat ove minimizacije se dobija na osnovu

$$\frac{d}{d\alpha}\left((x_1(n)-\alpha)^2+(x_2(n)-\alpha)^2+\cdots+(x_M(n)-\alpha)^2\right)=0$$

  i dat je izrazima koje smo koristili na prethodnim slajdovima.
- Estimacija usrednjavanjem odbiraka je *vrlo osjetljiva* na moguće pogrešno snimljene realizacije odbiraka $x(n)$, kao i na realizacije oštećene vrlo jakim smetnjama usljed određenih vanrednih okolnosti.
- Ovakve realizacije signala, koje se značajno razlikuju od pravih vrijednosti odbiraka poznate su kao ***outlier***-i, za razliku od realizacija koje su sa relativno malim greškama, koje su poznate pod nazivom ***inlier***-i.
- Računanje srednje vrijednosti odbiraka će dati potpuno pogrešan rezultat ukoliko se pojavi (desi) najmanje jedan *outlier* u razmatranom skupu realizacija $\{x_i(n)\}$.
- Najmanji mogući udio odbiraka (u odnosu na ukupan broj realizacija) koji treba zamijeniti sa *outlier*-ima da bi estimator postao neograničen (engl. *unbounded*) naziva se **tačkom prekida** (*breakdown point*) estimatora.
- Za usrednjavanje odbiraka primjenom ranije razmatranih formula, tačka prekida je najmanja moguća, odnosno, $1/M$ (gdje je M broj dostupnih realizacija), budući da samo jedan odbirak (*outlier*) usrednjavanje može učiniti neograničenim (*unbounded*).

Medijan

- Estimatori koji su robustni na moguće vrijednosti *outlier*-a u podacima se definišu i analiziraju u okviru robustne statistike. U nastavku će biti analiziran najednostavniji alat **robustne statistike – medijan odbiraka**.
- **Medijan odbiraka** je još jedan statistički alat za opisivanje slučajnih vrijednosti.
- Medijan skupa podataka je vrijednost koja se nalazi u sredini tog skupa, nakon sortiranja elemenata po njihovoj veličini. Ako sortirane vrijednosti od $x(n)$ označimo sa $s(n)$:

$$s(n) = \text{sort}\{x(n)\},\ n = 1, 2, \ldots, N$$

tada je medijan definisan kao

$$\text{median}\{x(n)|\ \ n = 1, 2, \ldots, N\} = s\left(\frac{N+1}{2}\right), \text{ za neparno } N.$$

- Ako je N parno, tada se medijan definiše kao srednja vrijednost dva odbirka koja su najbliža poziciji $(N+1)/2$, odnosno:

$$\text{median}\{x(n)|\ \ n = 1, 2, \ldots, N\} = \frac{s\left(\frac{N}{2}\right) + s\left(\frac{N}{2}+1\right)}{2}, \text{ za parno } N.$$

Medijan

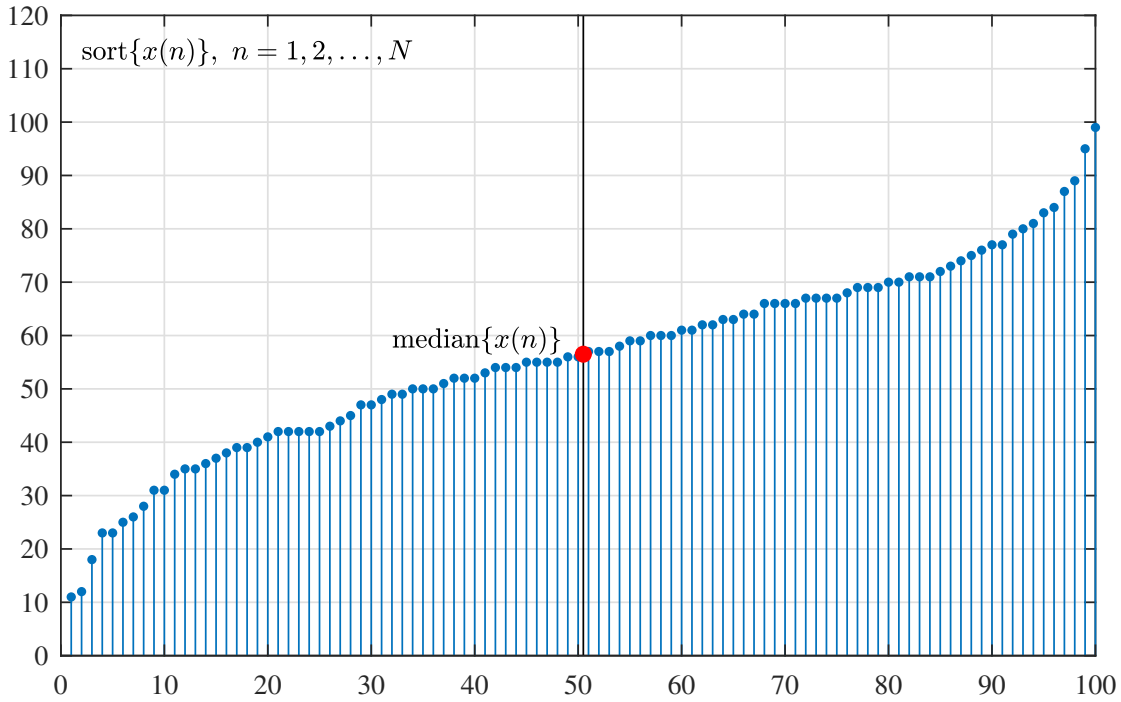
Primjer 4

Odrediti medijan sljedećih skupova:

1. $\mathbb{A} = \{-1,\ 1, -2,\ 4,\ 6, -9,\ 0\}$,
2. $\mathbb{B} = \{-1,\ 1, -1367,\ 4,\ 35, -9,\ 0\}$, i
3. skup čine odbirci signala $x(n)$ iz primjera 1.

1. Nakon sortiranja vrijednosti iz skupa $\mathbb{A}$ dobija se $\mathbb{A} = \{-9, -2, -1,\ 0,\ 1,\ 4,\ 6\}$. Traženi medijan je stoga $\text{median}(\mathbb{A}) = 0$.
2. Slično kao u prethodnom slučaju, važi $\text{median}(\mathbb{B}) = 0$. Uočiti da se srednje vrijednosti posmatranih skupova podataka značajno razlikuju.
3. Sortirane vrijednosti signala $x(n)$, u oznaci $\text{sort}\{x(n)\}$, su prikazane na narednom slajdu. Budući da signal $x(n)$ ima $N = 100$ odbiraka, ne postoji jedan odbirak u sredini sortirane sekvence.
   Sredina je pozicionirana između sortiranih odbiraka na poziciji 50 i 51. Iz navedenog razloga, medijan se definiše kao srednja vrijednost odbiraka na pozicijama 50 i 51, u sortiranom skupu.

Primjer 4. Uticaj outlier-a na medijan



Slika: Sortirane vrijednosti i medijan signala $x(n)$.

- Na median neće uticati mogući mali broj *outlier*-a, odnosno, vrijednosti koje su značajno različite od vrijednosti u preostalom dijelu podataka.
- U najgorem slučaju, potrebno je njima zamijeniti $N/2$ vrijednosti, da bi bili sigurni da je odbirak sa vrijednošću u sredini sortiranog skupa među *outlier*-ima, odnosno, da medijan rezultata nije *inlier*. Stoga, tačka prekida (*breakdown point*) ovog estimatora je $(N/2)/N = 1/2$.

Izvođenje estimatora zasnovanog na medijanu

- Estimator zasnovan na usrednjavanju odbiraka je uveden minimizacijom kvadrata rastojanja (devijacije) od dostupnih realizacija $x_i(n)$.
- Budući da je kvadrirana vrijednost velikih grešaka vrlo velika, ovaj tip estimatora je vrlo osjetljiv na uticaj *outlier*-a.
- Uobičajen postupak za redukciju uticaja velikih grešaka jeste da se u minimizacionoj funkciji koju smo razmatrali za slučaj srednje vrijednosti umjesto kvadrata razlike koristi apsolutna vrijednost razlike, odnosno:

$$\min_{\alpha} \Big(|x_1(n) - \alpha| + |x_2(n) - \alpha| + \cdots + |x_M(n) - \alpha| \Big).$$

- Isto važi i u slučaju kada se razmatraju vrijednosti odbiraka signala $x(n)$, $n = 1, 2, \ldots, N$, odnosno:

$$\min_{\alpha} \Big(|x(1) - \alpha| + |x(2) - \alpha| + \cdots + |x(N) - \alpha| \Big).$$

- U nastavku će biti pokazano da je rezultat ove minimizacije zapravo medijan razmatranog skupa,

$$\underset{i=1,2,\ldots,M}{\text{median}} \{x_i(n)\} = \min_{\alpha} \Big(|x_1(n) - \alpha| + |x_2(n) - \alpha| + \cdots + |x_M(n) - \alpha| \Big),$$

gdje $\underset{i=1,2,\ldots,M}{\text{median}} \{x_i(n)\}$ označava $\text{median}\{x_i(n) |\ \ i = 1, 2, \ldots, M\}$.

Izvođenje estimatora zasnovanog na medijanu

- Razmatra se **funkcija cijene** (engl. ***cost function***):

$$f(\alpha) = |x_1(n) - \alpha| + |x_2(n) - \alpha| + \cdots + |x_M(n) - \alpha| = ||\mathbf{x}(n) - \alpha||_1,$$

  i bez gubljenja opštosti, smatra se da su odbirci u vektoru $\mathbf{x}(n) = [x_1(n), x_2(n), \ldots, x_M(n)]^T$ već sortirani, $x_1(n) \leq x_2(n) \leq \ldots, x_M(n)$, kao i da je M neparan broj.
- Minimum ove funkcije ne može se dobiti kao u izvođenju za srednju vrijednost, pošto ova funkcija nije diferencijabilna u tačkama $\alpha = x_1(n)$, $\alpha = x_2(n)$, $\ldots$, $\alpha = x_M(n)$. Međutim, funkcija $f(\alpha)$ jeste diferencijabilna za sve druge vrijednosti α, a takođe je i kontinualna za bilo koje α.
- Navedena svojstva ćemo iskoristiti za određivanje intervala parametra α za koje funkcija raste ili opada. Izvod funkcije $|x_i(n) - \alpha|$ je jednak:

$$\frac{d|x_i(n) - \alpha|}{d\alpha} = \begin{cases} -1, & \text{za } \alpha < x_i(n) \\ 1, & \text{za } \alpha > x_i(n). \end{cases}$$

- Ako se sada posmatra interval koji je lijevo od najmanje vrijednosti signala (pretpostavljeno je da je signal sortiran), $\alpha < x_1(n)$, lako se zaključuje da je izvod $f(\alpha)$ u tom intervalu jednak sumi izvoda $d|x_i(n) - \alpha|/d\alpha = -1$, za svaki član $i = 1, 2, \ldots, M$. Ova suma je $df(\alpha)/d\alpha = -M$.

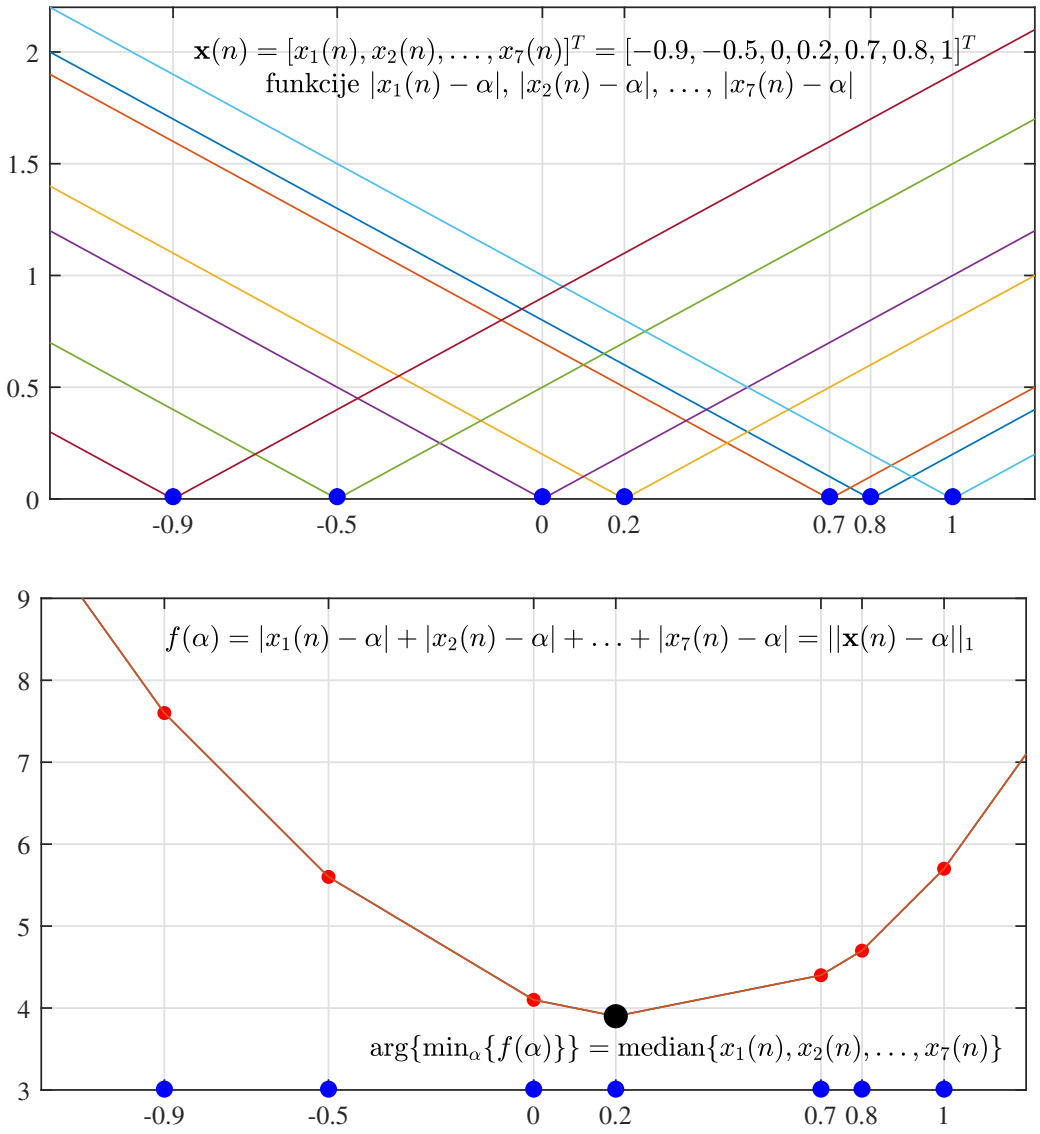
Izvođenje estimatora zasnovanog na medijanu

- Ako se sada pomjerimo desno duž ose α, na interval $x_1(n) < \alpha < x_2(n)$, tada se izvod od $|x_1(n) - \alpha|$ mijenja u 1, dok svi preostali članovi, njih $M-1$, imaju izvode jednake -1. Ovo znači da za posmatrani intervali važi $df(\alpha)/d\alpha = -M+2$.
- Ako se postupak ponovi za naredni interval, $x_2(n) < \alpha < x_3(n)$, i tako dalje, za sve preostale intervale, dobija se

$$\frac{df(\alpha)}{d\alpha} = \begin{cases} -M, & \text{za } \alpha < x_1(n) \\ -M+2, & \text{za } x_1(n) < \alpha < x_2(n) \\ \vdots & \\ -1, & \text{za } x_{(M-1)/2}(n) < \alpha < x_{(M+1)/2}(n) \\ 1, & \text{za } x_{(M+1)/2}(n) < \alpha < x_{(M+3)/2}(n) \\ \vdots & \\ M, & \text{za } \alpha > x_M(n). \end{cases}$$

- Ilustracija je predstavljena na narednom slajdu, za slučaj signala $\mathbf{x}(n) = [x_1(n), x_2(n), \ldots, x_7(n)]^T = [-0.9, -0.5,\ 0,\ 0.2,\ 0.7,\ 0.8,\ 1]^T$.
- Očigledno, funkcija cijene $f(\alpha)$ je opadajuća funkcija, $df(\alpha)/d\alpha < 0$, za $\alpha < x_{(M+1)/2}(n)$, odnosno, rastuća funkcija, $df(\alpha)/d\alpha > 0$, za $\alpha > x_{(M+1)/2}(n)$.

Medijan kao rješenje minimizacije ℓ_1-norme



Slika: Medijan kao rješenje minimizacije ℓ_1-norme.

Estimator zasnovan na medijanu. L-statistika.

- Pošto je funkcija $f(\alpha)$ kontinualna, prethodno razmatranje dokazuje da važi:
$$\underset{i=1,2,\ldots,M}{\text{median}}\{x_i(n)\} = \min_{\alpha}(f(\alpha)).$$
- Kada je M parno, u intervalu $x_{M/2}(n) < \alpha < x_{M/2+1}(n)$ važi da je $df(\alpha)/d\alpha = 0$.
- Navedeno znači da funkcija cijene opada za $\alpha < x_{M/2}(n)$, zatim da je konstantna u intervalu $x_{M/2}(n) < \alpha < x_{M/2+1}(n)$, a da zatim raste za $\alpha > x_{M/2+1}(n)$.
- U slučaju parnog M, srednja vrijednost od $x_{M/2}(n)$ i $x_{M/2+1}(n)$ se koristi kao medijan.
- U nekim slučajevima, *outlier*-i su mali. Tada će medijan odbaciti mnoge vrijednosti *inlier*-a koje bi mogle voditi dobroj estimaciji srednje vrijednosti.
- U ovim slučajevima, najbolji izbor bi bio da se koriste ne samo srednje vrijednosti sortiranog signala, već određeni broj vrijednosti oko medijana, čime se računa "odsječena" srednja vrijednost, koja je za neparno N, definisana izrazom:
$$\text{LSmean}\{x(n)|\ \ n = 1,2,\ldots,N\} = \frac{1}{2L+1}\sum_{i=-L}^{L} s\left(\frac{N+1}{2}+i\right).$$
- Sa $L = (N-1)/2$, koriste se sve vrijednosti signala, i prethodni izraz predstavlja standardnu srednju vrijednost signala, dok je za $L = 0$, vrijednost prethodnog izraza jednaka medijanu signala. Pristup je poznat kao **L-statistika**.

Varijansa i standardna devijacija

- Sljedeći važan parametar u statistici jeste devijacija (odstupanje) realizacija slučajnog odbirka od srednje vrijednosti.
- Najčešće korišćeni parametar za opisivanje ove statističke osobine jeste **standardna devijacija** ili njena kvadrirana vrijednost poznata kao **varijansa**.
- Za slučajni signal $x(n)$, čije su vrijednosti dostupne u M realizacija, **varijansa** se računa kao srednja kvadratna devijacija vrijednosti signala od odgovarajućih srednjih vrijednosti, $\mu_x(n)$:

$$\hat{\sigma}_x^2(n) = \frac{1}{M}\left(|x_1(n) - \mu_x(n)|^2 + \cdots + |x_M(n) - \mu_x(n)|^2\right).$$

- **Standardna devijacija** predstavlja kvadratni korijen od varijanse. Ona se može estimirati kao kvadratni korijen od srednje vrijednosti kvadrata centriranih podataka, odnosno:

$$\hat{\sigma}_x(n) = \sqrt{\frac{1}{M}\left(|x_1(n) - \mu_x(n)|^2 + \cdots + |x_M(n) - \mu_x(n)|^2\right)}.$$

- Ako je srednja vrijednost estimirana korišćenjem istog skupa podataka, $\hat{\mu}_x(n) = \frac{1}{M}\sum_{i=1}^{M} x_i(n)$, prethodno definisani estimator standardne devijacije ima tendenciju da daje manje vrijednosti standardne devijacije (u pitanju je vrijednost sa sistematskom greškom (engl. ***bias***).

Varijansa i standardna devijacija

- Iz prethodno navedenog razloga, u praksi se koristi prilagođena verzija, u literaturi poznata kao ***sample* standardna devijacija**, koja je data izrazom:

$$\hat{\sigma}_x(n) = \sqrt{\frac{1}{M-1}\left(|x_1(n) - \hat{\mu}_x(n)|^2 + \cdots + |x_M(n) - \hat{\mu}_x(n)|^2\right)}.$$

- Ova forma potvrđuje činjenicu da u slučaju kada je dostupan samo jedan odbirak (odnosno uzorak) $M = 1$, ne bi trebalo da postoji mogućnost da se estimira standardna devijacija.
- Ako standardnu devijaciju skupa podataka $\{x_i(n)\}$, gdje je $i = 1, 2, \ldots, M$, označimo sa $S(x_1(n), x_2(n)), \ldots, x_M(n)) = \sigma_x(n)$, tada ona zadovoljava **svojstvo skaliranja**:

$$S(ax_1(n) + b, ax_2(n) + b, \ldots, ax_M(n) + b) = |a| S(x_1(n), x_2(n)), \ldots, x_M(n)).$$

- Dokaz je jednostavan, korišćenjem definicije standardne devijacije i svojstva da je srednja vrijednost od $y_i(n) = ax_i(n) + b$ data sa $\hat{\mu}_y(n) = a\hat{\mu}_x(n) + b$.
- Standardna devijacija je osjetljiva na pojavu *outlier*-a, što se može zaključiti na osnovu njene definicije.

Varijansa i standardna devijacija

- Za estimaciju širenja (odnosno, odstupanja), može se koristiti i **apsolutna devijacija zasnovana na medijanu** (engl. *mean absolute deviation*, MAD), kao njena robustna forma.
- MAD se definiše kao

$$MAD_x(n) = \underset{j=1,2,\ldots,M}{\text{median}}\left\{\left|x_j(n) - \underset{i=1,2,\ldots,M}{\text{median}}\{x_i(n)\}\right|\right\},$$

po analogiji sa definicijom varijanse,

$$\sigma_x^2(n) = \underset{j=1,2,\ldots,M}{\text{mean}}\left\{\left|x_j(n) - \underset{i=1,2,\ldots,M}{\text{mean}}\{x_i(n)\}\right|^2\right\}.$$

- Važno je napomenuti da se, u svrhu dobijanja estimatora varijanse bez sistematske greške (odnosno, *bias*-a), umjesto srednje vrijednosti po indeksu j koristi suma koja se dijeli sa $(M-1)$.
- MAD vrijednost je u vezi sa standardnom devijacijom odbirka (uzorka),

$$MAD_x(n) = 0.6745\sigma_x(n),$$

za slučajnu varijablu sa Gausovom raspodjelom. Tačka prekida (***breakdown point***) za MAD je ista kao u slučaju medijana uzoraka.

Varijansa i standardna devijacija – primjer

Primjer 5

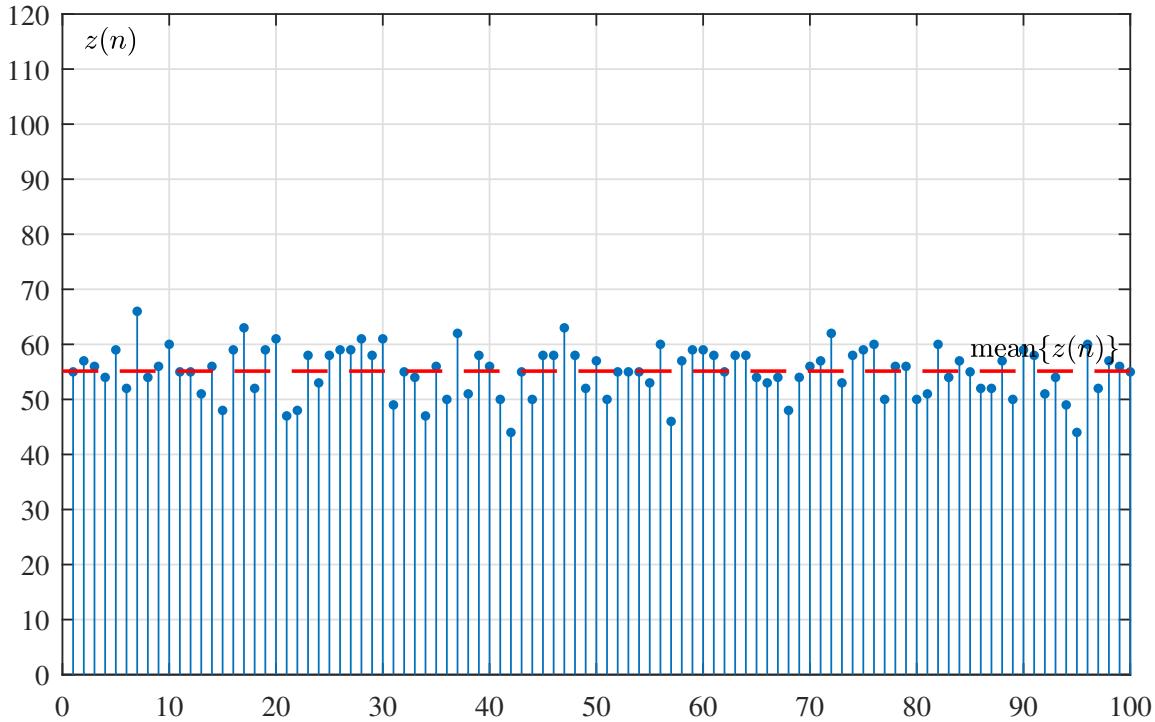
Za signal iz primjera 1 izračunati srednju vrijednost i varijansu. Uporediti je sa srednjom vrijednošću i varijansom signala $z(n)$ datom u tabeli ispod:

55	57	56	54	59	52	66	54	56	60
55	55	51	56	48	59	63	52	59	61
47	48	58	53	58	59	59	61	58	61
49	55	54	47	56	50	62	51	58	56
50	44	55	50	58	58	63	58	52	57
50	55	55	55	53	60	46	57	59	59
58	55	58	58	54	53	54	48	54	56
57	62	53	58	59	60	50	56	56	50
51	60	54	57	55	52	52	57	50	59
58	51	54	49	44	60	52	57	56	55

- Srednja vrijednost i varijansa signala $x(n)$ su $\hat{\mu}_x = 55.76$ i $\hat{\sigma}_x^2 = 314.3863$. Standardna devijacija je $\hat{\sigma}_x = 17.7309$. Ona predstavlja mjeru odstupanja vrijednosti signala od srednje vrijednosti.

Primjer 5

- Za signal $z(n)$, srednja vrijednost je $\hat{\mu}_z = 55.14$ (što je veoma blizu $\hat{\mu}_x$), dok je varijansa $\hat{\sigma}_z^2 = 18.7277$, a standardna devijacija $\hat{\sigma}_z = 4.3275$.
- Odstupanje signala $z(n)$ od njegove srednje vrijednosti je mnogo manje. Ako bi signali $x(n)$ i $z(n)$ bili mjerenja istog fizičkog procesa, tada bi pojedinačna mjerenja $z(n)$ bila mnogo *pouzdanija* od pojedinačnih mjerenja $x(n)$.
- Signal $z(n)$ je vizuelno predstavljen na slici ispod.



Slika: Slučajni signal $z(n)$ iz postavke zadatka.

Regresiona analiza. Linearna regresija

Ljubiša Stanković, Miloš Brajović

Univerzitet Crne Gore
Elektrotehnički fakultet

Prezentacija 1

Linearna regresija

- Regresiona analiza se bavi modelovanjem slučajnih varijabli i zastupljena je u brojnim naučnim oblastima, uključujući mašinsko učenje i predikciju podataka.
- Najčešći model je **linearna regresija**, gdje se pretpostavlja da rezultujuća slučajna varijabla odgovara linearnom modelu nezavisne (takođe slučajne) varijable.
- U kontekstu obrade signala, razmatraće se kontinualni slučajni signal, $x(t)$ odabran u slučajnim trenucima t_n. U linearnoj regresiji, model signala je linearna funkcija

$$x(t_n) = at_n + b + \varepsilon(t_n), \quad n = 1, 2, \ldots, N,$$

  gdje $\varepsilon(t_n)$ predstavlja slučajnu varijablu koja opisuje devijacije pojedinačnih realizacija, $x(t_n)$, od pretpostavljenog linearnog modela, $at_n + b$, sa konstantnim parametrima a i b.
- Vrijednosti $\varepsilon(t_n)$ su nepoznate.
- Cilj je estimirati parametre linearnog modela, a i b, na osnovu dostupnih podataka, a zatim ih koristiti za predikciju ili klasifikaciju novih podataka.
- Budući da su vrijednosti $x(t_n)$ i t_n dostupne, funkcija greške je data izrazom:

$$e(n) = x(t_n) - at_n - b.$$

Linearna regresija

- Funkcija cijene (engl. *cost function*), koja će se koristiti u procesu minimizacije, je:

$$J(a,b) = f(x(t_n) - at_n - b).$$

- Najčešći oblik funkcije cijene se definiše kao suma kvadrata vrijednosti funkcije greške, odnosno:

$$J(a,b) = \sum_{n=1}^{N} e^2(n) = \sum_{n=1}^{N} \Big(x(t_n) - at_n - b\Big)^2. \tag{1}$$

- Ova funkcija cijene je **optimalna** ukoliko su smetnje u mjerenjima, $e(n) = \varepsilon(t_n)$, takve da imaju **Gausovu raspodjelu**.
- Minimizacija ove funkcije (tzv. minimizacija u smislu **najmanjih kvadrata**, engl. *least squares* - LS) se obavlja izjednačavanjem njenih parcijalnih izvoda po nepoznatim parametrima a i b sa nulom:

$$\frac{\partial J(a,b)}{\partial a} = -2\sum_{n=1}^{N} t_n \Big(x(t_n) - at_n - b\Big) = 0$$

i

$$\frac{\partial J(a,b)}{\partial b} = -2\sum_{n=1}^{N} \Big(x(t_n) - at_n - b\Big) = 0.$$

Linearna regresija

- Prethodno razmatranje vodi do sljedećeg sistema jednačina:

$$\hat{a}\sum_{n=1}^{N} t_n^2 + \hat{b}\sum_{n=1}^{N} t_n = \sum_{n=1}^{N} t_n x(t_n) \tag{2}$$

$$\hat{a}\sum_{n=1}^{N} t_n + \hat{b}N = \sum_{n=1}^{N} x(t_n). \tag{3}$$

- Posmatrani sistem se može zapisati i u matričnoj formi kao $\mathbf{A}[\hat{a}\ \ \hat{b}]^T = \mathbf{B}$, daje estimirane vrijednosti parametara a i b odgovarajućeg modela linearne regresije, u oznaci $\hat{a}$ i $\hat{b}$.
- Estimirane vrijednosti su: $[\hat{a}\ \ \hat{b}]^T = \mathbf{A}^{-1}\mathbf{B}$.
- Nakon rješavanja sistema jednačina, dobijaju se odgovarajuće estimirane vrijednosti parametara:

$$\hat{a} = \frac{\hat{\mu}_{xt} - \hat{\mu}_x\hat{\mu}_t}{\hat{\mu}_{t^2} - \hat{\mu}_t^2} \quad \text{and} \quad \hat{b} = \hat{\mu}_x - \hat{a}\hat{\mu}_t,$$

gdje su: $\hat{\mu}_{xt} = \text{mean}\{x(t_n)t_n\}$, $\hat{\mu}_x = \text{mean}\{x(t_n)\}$, $\hat{\mu}_t = \text{mean}\{t_n\}$, i $\hat{\mu}_{t^2} = \text{mean}\{t_n^2\}$ odgovarajuće srednje vrijednosti veličina koje figurišu u (2) i (3).

Primjer 1

Slučajni signal $x(t)$, čije vremenske promjene imaju očekivanu linearnu karakteristiku, odabran je u trenucima t_n:

$$[t_1, t_2, \ldots, t_N]^T = [0.95,\ 0.23,\ 0.61,\ 0.49,\ 0.89, 0.76,\ 0.46,\ 0.02]^T.$$

Dobijene vrijednosti signala $x(t_n)$ su:

$$[x(t_1), x(t_2), \ldots, x(t_N)]^T = [4.81,\ 3.13,\ 4.25,\ 4.04,\ 4.55,\ 4.76,\ 4.16,\ 3.03]^T.$$

Odrediti linearni regresioni model korišćenjem pristupa najmanjih kvadrata. Izvršiti predikciju vrijednosti signala $x(t)$ u trenutku $t = 1.1$.

- Elementi matrica $\mathbf{A}$ i $\mathbf{B}$ u sistemu jednačina

$$\mathbf{A} \begin{bmatrix} \hat{a} & \hat{b} \end{bmatrix}^T = \mathbf{B}$$

koji se koristi za estimaciju parametara modela a i b su:

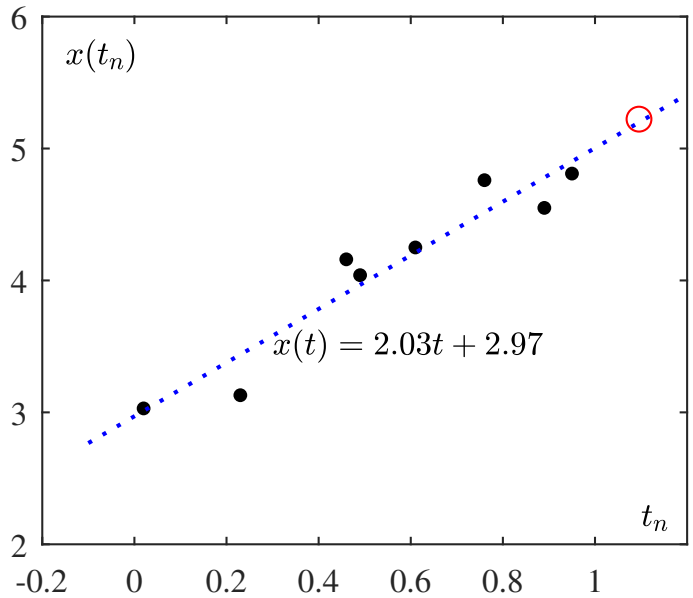
$$\mathbf{A} = \begin{bmatrix} 3.15 & 4.41 \\ 4.41 & 8.00 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 19.50 \\ 32.73 \end{bmatrix}.$$

Linearna regresija

- Estimirane vrijednosti su:

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} 3.15 & 4.41 \\ 4.41 & 8.00 \end{bmatrix}^{-1} \begin{bmatrix} 19.50 \\ 32.73 \end{bmatrix} = \begin{bmatrix} 2.03 \\ 2.97 \end{bmatrix}.$$

- Estimirani regresioni model je $x(t_n) = 2.03t_n + 2.97$. Za $t_n = 1.1$, predviđena vrijednost signala je $x(1.1) = 5.2$.



Slika: Podaci $x(t_n)$ mjereni u trenucima t_n za $n = 1, 2, 3, 4, 5, 6, 7, 8$ (tačke) i linearni model, $x(t_n) = 2.03t_n + 2.97$, dobijen na bazi metoda najmanjih kvadrata (tačkasta linija). Rezultat predikcije signala u trenutku $t_n = 1.1$ je označen kružićem.

Linearna regresija

- **Matrična forma** izraza (1) je

$$J(\mathbf{a}) = ||\mathbf{x} - \mathbf{T}\mathbf{a}||_2^2,$$

gdje su:

$$\mathbf{x} = [x(t_1), x(t_2), \ldots, x(t_N)]^T, \quad \mathbf{T} = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_N \end{bmatrix}, \quad \text{i} \quad \mathbf{a} = \begin{bmatrix} b \\ a \end{bmatrix}.$$

- Rješenje minimizacionog problema se dobija iz $\partial J(\mathbf{a})/\partial \mathbf{a}^T = \mathbf{0}$, odnosno, $-2\mathbf{T}^T(\mathbf{x} - \mathbf{T}\hat{\mathbf{a}}) = \mathbf{0}$. Rješenje je:

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

- Regresiona analiza se može generalizovati na slučajeve sa više od jedne nezavisne varijable.
- Ove forme regresije će biti razmatrane nakon prezentacije RANSAC pristupa.

Random Sample Consensus (RANSAC)

- Kada su poznati srednja vrijednost podataka i mjera raširenosti podataka oko srednje vrijednosti (uobičajeno standardna devijacija), tada je moguće definisati **kriterijum za identifikaciju *outlier*-a u podacima**. Funkcija koja se koristi za ovu namjenu je poznata kao **z-skor** i definiše se sljedećim izrazom:

$$z(n) = \frac{x(n) - \hat{\mu}_x(n)}{\hat{\sigma}_x(n)}.$$

- Uobičajeno je pretpostaviti prag sa vrijednošću $T = 2.5$ i deklarisati odbirke signala kod kojih z-skor zadovoljava $|z(n)| \leq T$ kao *inlier*-e, a ostale odbirke signala, za koje važi $|z(n)| > T$, kao *outlier*-e. Detaljnije objašnjenje praga $T = 2.5$ ćemo dati kasnije.
- Pošto vrijednosti prosjeka $\hat{\mu}_x(n)$ i standardne devijacije $\hat{\sigma}_x(n)$ uzoraka mogu biti značajno kompromitovane mogućim *outlier*-ima, preporučuje se korišćenje medijana i odgovarajućeg MAD-a u z-skoru.
- **Slučajni konsenzus uzoraka** (engl. *Random sample consensus* (**RANSAC**)) se koristi za linearnu regresiju u slučajevima kada se očekuju *outlier*-i u podacima.
- Neka se razmatra skup podataka (signal) $x(t_n)$, odabran u slučajnim trenucima t_n. Dalje se pretpostavlja da podaci odgovaraju linearnom modelu.

Random Sample Consensus (RANSAC) – algoritam

- Pošto se očekuje veliki broj *outlier*-a, linearni model može biti daleko od većine odbiraka. Stoga RANSAC pristup podrazumijeva sljedeće korake:
  1. Pretpostaviti mali podskup $\mathbb{S}$ sa S selektovanih indeksa trenutaka t_n, odnosno odbiraka $x(t_n)$, gdje je $n \in \mathbb{S}$.
  2. Odbirci sa indeksima iz skupa $\mathbb{S}$ se koriste za estimaciju parametara linearnog regresionog modela:

$$\hat{a}\sum_{n\in\mathbb{S}} t_n^2 + \hat{b}\sum_{n\in\mathbb{S}} t_n = \sum_{n\in\mathbb{S}} t_n x(t_n)$$
$$\hat{a}\sum_{n\in\mathbb{S}} t_n + \hat{b}N = \sum_{n\in\mathbb{S}} x(t_n).$$

  3. Nakon estimiranja parametara linearne regresije a i b iz

$$[\hat{a}\ \ \hat{b}]^T = \mathbf{A}_S^{-1}\mathbf{B}_S,$$

definiše se prava

$$x(t) = \hat{a}t + \hat{b}.$$

Zatim se računaju udaljenosti d_n svih tačaka $(t_n, x(t_n))$, $n = 1, 2, \ldots, N$ od dobijene prave, po formuli:

$$d_n = \frac{|\hat{a}t_n + \hat{b} - x(t_n)|}{\sqrt{1+\hat{a}^2}}.$$

Random Sample Consensus (RANSAC) – algoritam

4. Ako postoji dovoljan broj tačaka čija je udaljenost od linije modela manja od nekog unaprijed zadatog praga udaljenosti d, tada se sve ove tačke uključuju u novi skup podataka
$$\mathbb{D} = \{(t_n, x(t_n)) | \quad d_n \leq d\},$$
a zatim se vrši finalna estimacija parametara a i b (za mašinsko učenje, ili za predikciju), na osnovu svih podataka iz $\mathbb{D}$.
5. Ako ne postoji dovoljan broj tačaka sa rastojanjem manjim od d, bira se novi slučajni podskup podataka, sa indeksima $n \in \mathbb{S}$, i procedura se ponavlja od koraka 2.
6. Procedura se zaustavlja kada se dostigne željeni broj tačaka unutar $\mathbb{D}$, ili se dostigne maksimalni dozvoljeni broj pokušaja.

Napomena

Rastojanje između prave $\alpha x + \beta y + \gamma = 0$ i tačke (x_0, y_0) je jednako:

$$\frac{|\alpha x_0 + \beta y_0 + \gamma|}{\sqrt{\alpha^2 + \beta^2}}.$$

U našem slučaju, jednačina prave je $-\hat{a}t + x(t) - \hat{b} = 0$, dok je posmatrana tačka $(t_n, x(t_n))$. Gornja formula je korišćena u koraku 3 RANSAC algoritma.

Random Sample Consensus (RANSAC)

Primjer 2

Razmatra se $N = 20$ slučajnih odbiraka signala $x(t_n)$

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

Odbirci su uzeti u odgovarajućim slučajnim trenucima t_n.

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

- Odrediti parametre a i b linearnog regresionog modela za posmatrani skup podataka. Prokomentarisati rezultat.
- Nakon toga, primijeniti RANSAC pristup na sljedeći način: uzeti $S = 4$ (slučajno) odabranih podataka sa indeksima $n \in \mathbb{S} = \{8, 10, 18, 19\}$.
- Odrediti parametre linearnog regresionog modela za ovaj podskup podataka.
- Koliko je tačaka (koje odgovaraju vrijednostima podataka) unutar rastojanja $d = 0.25$ od prave koja reprezentuje dobijeni linearni model?

Primjer 2 (nastavak)

- 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 $\mathbb{S}$. U tom slučaju koristiti $\mathbb{S} = \{5, 11, 16, 19\}$ i ponoviti proceduru.
- 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.

- Za zadati skup podataka, estimirane vrijednosti parametara a i b linearnog regresionog modela se dobijaju iz

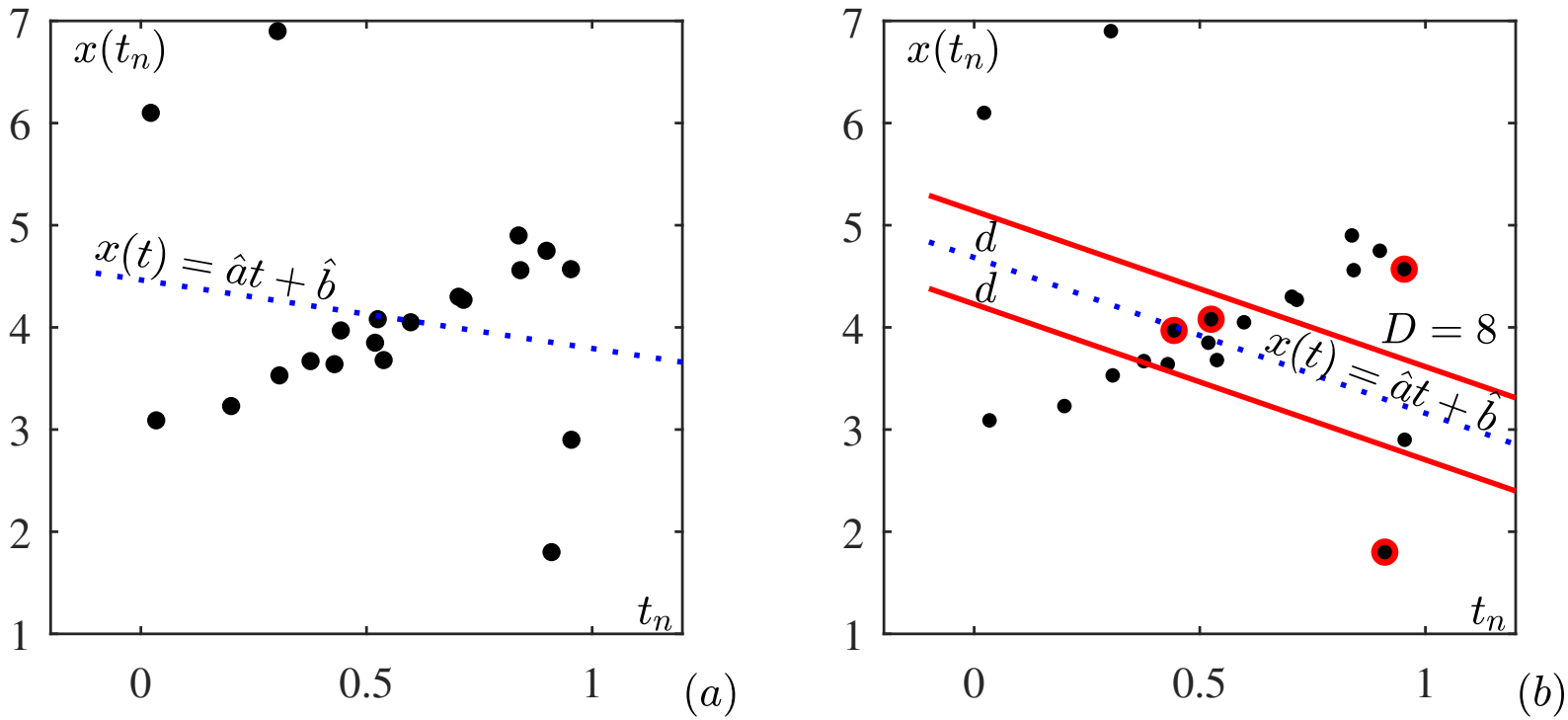
$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} 7.8107 & 11.1070 \\ 11.1070 & 20.0000 \end{bmatrix}^{-1} \begin{bmatrix} 44.3483 \\ 81.8400 \end{bmatrix} = \begin{bmatrix} -0.6707 \\ 4.4645 \end{bmatrix}$$

- Linearni model

$$x(t_n) = -0.6707 t_n + 4.4645.$$

ne odgovara podacima, zbog očiglednog prisustva *outlier*-a u trenucima t_1, t_4, t_{18}, i t_{20}, koji su korišćeni u izračunavanju, slika pod (a) na sljedećem slajdu.

Primjer 2



Slika: (a) Podaci $x(t_n)$ mjereni u trenucima t_n za $n = 1, 2, \ldots, 10$ (tačke) i linearni model dobijen minimizacijom u smislu najmanjih kvadrata (tačkasta linija). (b) Podaci (tačke) i linearni model dobijen metodom najmanjih kvadrata, korišćenjem slučajnog podskupa sastavljenog od 4 označena odbirka na pozicijama $\mathbb{S} = \{8, 10, 18, 19\}$ (tačkasta linija).

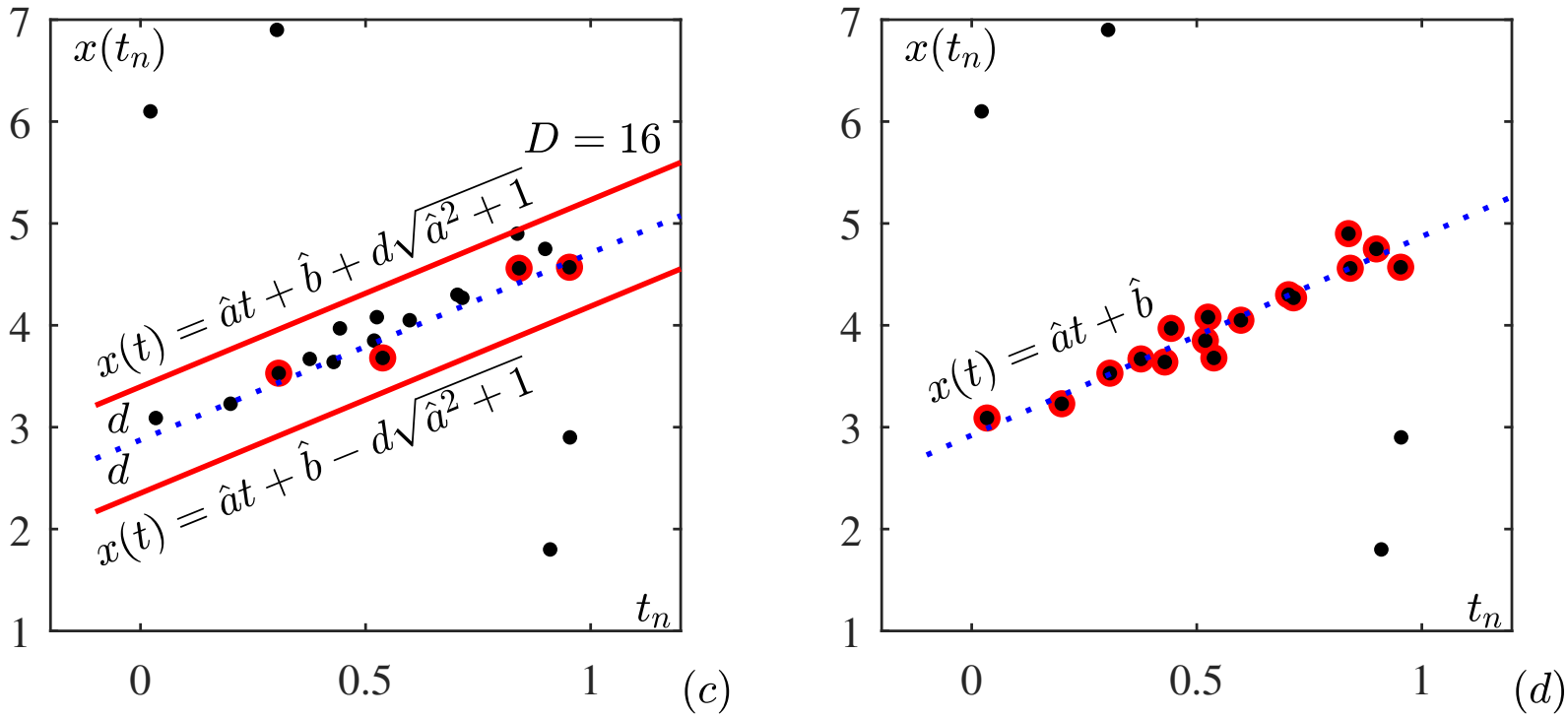
Primjer 2

- RANSAC pristup počinje slučajnom selekcijom podskupa podataka od $S = 4$ odbirka. U konkretnom primjeru, ove vrijednosti su određene indeksima $\mathbb{S} = \{8, 10, 18, 19\}$.
- Pošto se jedan od *outlier*-a, iz trenutka t_{18}, koristi u izračunavanju, estimacija parametara na osnovu

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} 2.2082 & 2.8310 \\ 2.8310 & 4.0000 \end{bmatrix}^{-1} \begin{bmatrix} 9.8939 \\ 14.4200 \end{bmatrix} = \begin{bmatrix} -1.5245 \\ 4.6840 \end{bmatrix}.$$

  čiji je odgovarajući linearni model $x(t_n) = -1.5245t_n + 4.6840$, ne odgovara podacima.
- Zaključak da prethodno dobijeni model ne odgovara podacima potvrđuje činjenica da je samo $D = 8$ tačaka koje odgovaraju podacima u unutrašnjosti regiona određenog linijama na udaljenosti $d = 0.25$ od prave kojom je reprezentovan model, kao što se može vidjeti na slici (b).
- Pošto je broj tačaka koje odgovaraju podacima unutar $\mathbb{D}$ manji od predefinisanog praga $T = 10$, sprovodi se nova estimacija parametara na osnovu novog slučajnog skupa, $\mathbb{S} = \{5, 11, 16, 19\}$.

Primjer 2



Slika: (c) Podaci (tačke) i linearni model dobijen minimizacijom u smislu najmanjih kvadrata, korišćenjem drugog podskupa od 4 označena odbirka na pozicijama $\mathbb{S} = \{5, 11, 16, 19\}$ (tačkasta linija). (d) Podaci (tačke) i linearni model dobijen minimizacijom u smislu najmanjih kvadrata, korišćenjem svih označenih odbiraka iz skupa $\mathbb{D}$ (tačkasta linija).

Primjer 2

- Na osnovu navedenih podataka, novoestimirani parametri se dobijaju iz:

$$\begin{bmatrix}\hat{a}\\ \hat{b}\end{bmatrix} = \begin{bmatrix}1.9992 & 2.6390\\ 2.6390 & 4.0000\end{bmatrix}^{-1}\begin{bmatrix}11.2537\\ 16.3400\end{bmatrix} = \begin{bmatrix}1.8342\\ 2.8749\end{bmatrix}$$

  sa odgovarajućim linearnim modelom $x(t_n) = 1.8342t_n + 2.8749$ koji odgovara podacima, pošto je sada $D = 16$ tačaka u skupu $\mathbb{D}$.
- Pošto je ovaj broj iznad predefinisanog praga $T = 10$, algoritam se zaustavlja i model linearne regresije se reestimira korišćenjem svih $D = 16$ uzoraka iz skupa $\mathbb{D}$, dajući:

$$\begin{bmatrix}\hat{a}\\ \hat{b}\end{bmatrix} = \begin{bmatrix}5.9802 & 8.9180\\ 8.9180 & 16.0000\end{bmatrix}^{-1}\begin{bmatrix}37.7188\\ 64.1400\end{bmatrix} = \begin{bmatrix}1.9502\\ 2.9218\end{bmatrix}$$

  kao i krajnju verziju estimiranog linearnog regresionog modela

$$x(t_n) = 1.9502t_n + 2.9218.$$

- Vjerovatnoća da je skup od $S = M$ podataka takav da u njemu nije pristutan nijedan od I *outlier*-a u nizu od N odbiraka može biti eksplicitno izračunata.
- Ova vjerovatnoća može da se iskoristi za estimaciju očekivanog broja iteracija RANSAC algoritma.