Segédanyag az *Idősorok és többdimenziós statisztika gyakorlat* (matematika BSc, matematikai elemző szakirány) tantárgyhoz.
2019. december 2.

Diszkrét idejű Markov-folyamatok (Markov-láncok)

Legyen S megszámlálható halmaz, neve: állapottér; $X_0, X_1, X_2, \dots$ valószínűségi változó sorozat, $P(X_i \in S) = 1$ minden i-re. Az állapottér elemeire sokszor egyszerűbb az $1, 2, \dots$ számokként gondolni.

Definíció. **Markov-tulajdonság.** Minden $0 \le n \in \mathbb{Z}$ és $i_j \in S,\ j = 1, 2, \dots$ esetén $P(X_{n+1} = i_{n+1} | X_n = i_n, \dots, X_0 = i_0) = P(X_{n+1} = i_{n+1} | X_n = i_n)$.

Definíció. **Markov-folyamat/lánc.** A Markov-tulajdonsággal rendelkező diszkrét idejű sztochasztikus folyamatokat Markov-folyamatoknak vagy Markov-láncoknak hívjuk.

Definíció. **Homogén Markov-lánc.** A Markov-lánc homogén (vagy stacionárius), ha a $P(X_{n+1} = j | X_n = i)$ feltételes valószínűség nem függ n-től, azaz minden n-re ugyanaz.
Jel. a kezdeti eloszlást $\mathbf{q}^T = (q_1, q_2, \dots)^T$, ahol $q_i = P(X_0 = i),\ \ i \in S$.
Jel. homogén Markov-lánc esetén $p_{i,j} := P(X_{n+1} = j | X_n = i),\ \ i, j \in S$. Ezek neve: egylépéses átmenetvalószínűségek vagy röviden átmenetvalószínűségek.
Jel. $\mathbf{P} = (p_{i,j})_{i,j \in S}$, neve: **átmenetvalószínűség mátrix**

Jel. homogén Markov-lánc esetén $p_{i,j}^{(n)} := P(X_n = j | X_0 = i),\ \ i, j \in S$. Ezek neve: n lépéses átmenetvalószínűségek.
Jel. $\mathbf{P}^{(n)} = \left(p_{i,j}^{(n)}\right)_{i,j \in S}$, neve: n lépéses átmenetvalószínűség mátrix

Tétel. **Chapman-Kolmogorov egyenletek.** Ha $m < k < n$, akkor

$$P(X_n = i_n | X_m = i_m) = \sum_{i_k \in S} P(X_n = i_n | X_k = i_k) \cdot P(X_k = i_k | X_m = i_m).$$

Következmény. $\mathbf{P}^{(n)} = \mathbf{P}^n$
Állítás. A $\mathbf{P}$ mátrix minden sorösszege 1.

Állítás. $P(X_n = j) = \sum_{i \in S} q_i \cdot (\mathbf{P}^n)_{i,j} = \left(\mathbf{q}^T \mathbf{P}^n\right)_j,\ \ j \in S$

Definíció. i állapotból j állapot **elérhető**, ha $\exists n \ge 1$: $p_{i,j}^{(n)} > 0$. Jel.: $i \to j$
Definíció. i és j **érintkeznek**, ha vagy $i = j$, vagy ($i \to j$ és $j \to i$). Jel.: $i \leftrightarrow j$

Állítás. Az elérhetőség tranzitív, az érintkezés pedig ekvivalenciareláció.

Definíció. **Pontosan n lépésben visszatérés valószínűsége.**
$f_n(i,i) := P(X_n = i, X_{n-1} \ne i, \dots, X_1 \ne i | X_0 = i),\ \ i \in S$
Definíció. **Visszatérés valószínűsége.** $f(i,i) := \sum_{n=1}^{\infty} f_n(i,i),\ \ i \in S$

Állítás. $p^{(n)}(i,i) = f_n(i,i) + \sum_{k=1}^{n-1} f_k(i,i) p^{(n-k)}(i,i)$

Definíció. **Visszatérő állapot.** Az i állapot visszatérő, ha $\forall j \in S$-re $i \to j \ \Rightarrow j \to i$.
Megjegyzés. Ekvivalens elnevezések: visszatérő=lényeges=perzisztens állapot.

Definíció. **Átmeneti állapot.** Az i állapot átmeneti, ha nem visszatérő.
Megjegyzés. Ekvivalens elnevezések: átmeneti=lényegtelen=**tranziens** állapot

Tétel. Az i állapot visszatérő $\iff f(i,i) = 1 \iff \sum_{n=0}^{\infty} p^{(n)}(i,i) = \infty$

Következmény. Az i állapot átmeneti $\iff f(i,i) < 1 \iff \sum_{n=0}^{\infty} p^{(n)}(i,i) < \infty$

Definíció. **Elnyelő állapot.** Az i állapot elnyelő, ha $p_{i,i} = 1$.

Definíció. Az i állapot periódusa: $d(i) := \text{lnko}\{n \ge 0 : p^{(n)}(i,i) > 0\}$

Definíció. **Periodikus állapot.** Az i állapot periodikus, ha $d(i) > 1$.

Definíció. **Aperiodikus állapot.** Az i állapot aperiodikus, ha $d(i) = 1$.

Definíció. **Irreducibilis ML:** az állapotai érintkeznek egymással.

Megjegyzés. Az irreducibilis Markov-lánc gráfja összefüggő.

Definíció. **Ergodikus ML:** irreducibilis, minden állapota visszatérő és aperiodikus.

Tétel. Legyen $\mathbf{P}$ egy ergodikus Markov-lánc átmenetvalószínűség mátrixa.
Ekkor $\forall i, j \in S$-re $p^{(n)}(i,j) \underset{n \to \infty}{\longrightarrow} \frac{1}{\sum_{n=1}^{\infty} n f_n(j,j)}$.

Vegyük észre, hogy az előző tétel alapján amihez konvergál, már nem függ a kiinduló i állapottól. Jelölje $\pi_j := \frac{1}{\sum_{n=1}^{\infty} n f_n(j,j)}\ \ j = 1, 2, \dots$, ezzel $\lim_{n \to \infty} \mathbf{P}^n = \begin{pmatrix} \pi_1 & \pi_2 & \cdots \\ \vdots & \vdots & \dots \\ \pi_1 & \pi_2 & \cdots \end{pmatrix}$.
Jel. $\boldsymbol{\pi}^T = (\pi_1, \pi_2, \dots)^T$, elnevezése: **stacionárius** vagy egyensúlyi **eloszlás**. A stacionárius eloszlás mutatja meg, hogy "hosszú idő után" a milyen valószínűséggel leszünk a Markov-lánc egyes állapotaiban.
A gyakorlatban a stacionárius eloszlást az alábbi egyenletrendszer megoldásával szokás kiszámolni: $\boldsymbol{\pi}^T = \boldsymbol{\pi}^T \mathbf{P}$, ahol $\sum_i \pi_i = 1$. Ennek értelmében tehát a $\boldsymbol{\pi}$ vektor a $\mathbf{P}$ mátrix baloldali, 1-re normált sajátvektora.

Markov-láncoknál egy lényeges kérdés, hogy átlagosan mennyi időbe (lépésbe) telik, míg az egyik állapotból egy másik állapotba eljutunk. Jelölje $m_{i,j}$: ha jelenleg az i állapotban vagyunk, akkor várhatóan ennyi lépésre van szükség, hogy a j állapotba kerüljünk. Általánosan ezeket az értékeket nem lehet közvetlenül egyszerűen kiszámolni, de a teljes várható érték tétel alapján felírható rájuk a következő egyenlet: $m_{i,j} = p_{i,j} + \sum_{k \ne j} p_{i,j}(1 + m_{k,j})$, amit $m_{i,j} = 1 + \sum_{k \ne j} p_{i,j} m_{k,j}$-ra lehet egyszerűsíteni.

Állítás. Ergodikus Markov-lánc esetén $m_{i,i} = \frac{1}{\pi_i}$

Elnyelő Markov-láncok:

- van s tranziens állapot: $t_1, \ldots, t_s$
- van m elnyelő állapot: $a_1, \ldots, a_m$

Particionáljuk ezek alapján az átmenetvalószínűség mátrixot: $\mathbf{P} = \begin{pmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & \mathbf{I}_m \end{pmatrix}$, ahol $\mathbf{Q} \in \mathbb{R}^{s \times s}$, $\mathbf{R} \in \mathbb{R}^{s \times m}$, $\mathbf{I}_m$ az m dimenziós egységmátrix.
Néhány lényeges mennyiség kiszámítása:

- ha t_i-ben vagyunk, akkor azon periódusok várható száma, amit t_j-ben töltünk, mielőtt egy elnyelő állapotba lépnénk: $\left((\mathbf{I}_s - \mathbf{Q})^{-1}\right)_{i,j}$
- ha t_i-ben vagyunk, akkor annak a valószínűsége, hogy a_j-be kerülünk: $\left((\mathbf{I}_s - \mathbf{Q})^{-1}\mathbf{R}\right)_{i,j}$

Elnyelő láncok esetén nem beszélhetünk olyan értelemben stacionaritásról, mint az ergodikus láncoknál, egyfajta egyensúly csak akkor érhető el, ha minden időszakban van(nak) új belépő(k) a rendszerbe. Tekintsük az n-edik időperiódust az $n-1$-edik és az n-edik időpont között eltelt időnek, $n = 1, 2, \ldots$
Vezessünk be jelöléseket:

- H_i: az egyes időperiódusok elején az i-edik állapotba belépő egyedek száma
- $N_i(n)$: az n-edik időperiódus elején az i-edik állapotban lévő egyedek száma
- $r_{i\bullet} = \sum_{j=1}^{m} r_{i,j}$, ami az i-edik állapotból egy elnyelő állapotba lépés valószínűsége
- $\widetilde{\mathbf{Q}} := \left(\begin{array}{c|c} \mathbf{Q} & \begin{matrix} r_{1\bullet} \\ \vdots \\ r_{s\bullet} \end{matrix} \end{array}\right)$

Kérdés, hogy léteznek-e a $\lim_{n\to\infty} N_i(n)$ határértékek. Ha léteznek, akkor jelöljük őket N_i-vel, amikből képezzük az $\mathbf{N} = (N_1, \ldots, N_s)^T$ egyensúlyi egyedszám vektort. Amennyiben létezik ilyen egyensúlyi helyzet, akkor minden időperiódusban az i-edik állapotba belépő egyedek számának (a lenti egyenletben a baloldal) meg kell egyeznie az onnan kilépő egyedek számával (jobboldal):

$$H_i + \sum_{k \neq i} N_k \cdot \widetilde{q}_{k,i} = N_i \cdot (1 - \widetilde{q}_{i,i}) \qquad i = 1, \ldots, s$$

Ajánlott irodalom: Wayne L. Winston: Operációkutatás, 17. fejezet

---

A Poisson-folyamat

Definíció. **Sztochasztikus folyamat:** $(X_t)_{t \in T}$, ahol T a paramétertér és minden t-re X_t valószínűségi változó.

Definíció. Az X_t sztochasztikus folyamat **független növekményű**, ha tetszőleges $t_1 < t_2 \leq t_3 < t_4$ esetén $X_{t_2} - X_{t_1}$ és $X_{t_4} - X_{t_3}$ növekmények függetlenek egymástól.

Definíció. Az X_t sztochasztikus folyamat **stacionárius növekményű**, ha tetszőleges $t_1 < t_2$ és h esetén $X_{t_2} - X_{t_1} \sim X_{t_2+h} - X_{t_1+h}$, azaz tetszőleges növekmények tetszőleges eltoltjai ugyanolyan eloszlásúak.

Megjegyzés. Az előző két definíció során feltesszük, a tetszőlegesen választott időpontok olyanok, hogy azok nem vezetnek ki a T paramétertartományból.

Definíció. **Poisson-folyamat.**
$X_t, t \geq 0$ Poisson-folyamat $\lambda > 0$ intenzitással, ha

- $X_0 = 0$,
- független növekményű,
- stacionárius növekményű,
- $P(X_t = 1) = \lambda t + o(t)$ és $P(X_t \geq 2) = o(t)$, ha $t \to 0$

Tétel. **A Poisson-folyamat tulajdonságai.**
Legyen X_t Poisson-folyamat λ intenzitással. Jelöje τ_i az i. esemény bekövetkezésének időpontját, $i = 1, 2, \ldots$ Ekkor
a.) $X_t \sim \text{Poi}(\lambda t)$
b.) az autokovariancia függvénye $\text{Cov}(X_t, X_s) = \lambda \cdot \min(t, s)$;
c.) $\tau_1, \tau_2 - \tau_1, \tau_3 - \tau_2, \ldots$ függetlenek és azonos, $\text{Exp}(\lambda)$ eloszlásúak
d.) ha $t > s$, akkor $X_s | X_t \sim \text{Bin}\left(X_t, \frac{s}{t}\right)$
e.) ha $t < s$, akkor $X_s | X_t \sim \; X_t + \text{Poi}(\lambda(s-t))$

Következmény. $\tau_n \sim \Gamma(n, \lambda), \;\; n = 1, 2, \ldots$

Állítás. **Poisson-folyamatok egyesítése.**
Legyenek $X_t^1, \ldots, X_t^m$ független Poisson-folyamatok $\lambda_1, \ldots, \lambda_m$ intenzitásokkal. Ekkor $X_t = X_t^1 + \ldots + X_t^m$ is Poisson-folyamat, $\lambda_1 + \ldots + \lambda_m$ intenzitással.

Tétel. **Poisson-folyamat ritkítása.**
Legyen X_t Poisson-folyamatok λ intenzitással. Minden egyes esemény bekövetkezésekor egymástól függetlenül feldobunk egy érmét, a fejdobás valószínűsége $p \in [0; 1]$. Legyen X_t^1 a t időpontig kapott fejek száma, X_t^2 pedig a t időpontig kapott írások száma.
Ekkor X_t^1 és X_t^2 egymástól független Poisson-folyamatok $\lambda \cdot p$ és $\lambda \cdot (1-p)$ intenzitásokkal.

Ajánlott irodalom: Márkus L. előadásfóliái a Poisson-folyamatról: `http://web.cs.elte.hu/probability/markus/ElemzoTS1/Egyszeru_poisson2.pdf`

Definíció. X **val.változó eloszlásfüggvénye:** $F_X(x) = P(X < x)$.

Állítás. Az eloszlásfüggvény tulajdonságai:

- $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$;
- balról folytonos;
- monoton növő.

Nevezetes diszkrét eloszlások:

Eloszlás neve	Jelölése	Eloszlása	EX	D^2X
Karakterisztikus (indikátorvált.)	$\mathrm{Ind}(p)$	$P(X=1)=p$ $P(X=0)=1-p$	p	$p(1-p)$
Geometriai (Pascal)	$\mathrm{Geo}(p)$	$P(X=k)=p(1-p)^{k-1}$ $k=1,2,\dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Hipergeometriai	Hipgeo(N,M,n)	$P(X=k)=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$ $k=0,1,\dots,n$	$n\frac{M}{N}$	$n\frac{M}{N}\left(1-\frac{M}{N}\right)\left(1-\frac{n-1}{N-1}\right)$
Binomiális	$\mathrm{Bin}(n,p)$	$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ $k=0,1,\dots,n$	np	$np(1-p)$
Negatív binomiális	$\mathrm{NegBin}(n,p)$	$P(X=k)=\binom{k-1}{n-1}p^n(1-p)^{k-n}$ $k=n,n+1,\dots$	$\frac{n}{p}$	$\frac{n(1-p)}{p^2}$
Poisson	$\mathrm{Poi}(\lambda)$	$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$ $k=0,1,\dots$	λ	λ

Nevezetes abszolút folytonos eloszlások:

Eloszlás neve	Jelölése	Eloszlásfüggvény	Sűrűségfüggvény	EX	D^2X
Egyenletes	$\mathrm{E}(a,b)$	$\begin{cases}0 & \text{ha } x\le a\\ \frac{x-a}{b-a} & \text{ha } a<x\le b\\ 1 & \text{ha } b<x\end{cases}$	$\begin{cases}\frac{1}{b-a} & \text{ha } a<x\le b\\ 0 & \text{különben}\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponenciális	$\mathrm{Exp}(\lambda)$	$\begin{cases}1-e^{-\lambda x} & \text{ha } x\ge 0\\ 0 & \text{különben}\end{cases}$	$\begin{cases}\lambda e^{-\lambda x} & \text{ha } x\ge 0\\ 0 & \text{különben}\end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Standard norm.	$\mathrm{N}(0,1^2)$	$\Phi(x)=\dots$	$\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\quad x\in\mathbb{R}$	0	1
Normális	$\mathrm{N}(m,\sigma^2)$	$\dots$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}\quad x\in\mathbb{R}$	m	σ^2

További nevezetes abszolút folytonos eloszlások:

Eloszlás neve	Jelölése	Eloszlásfüggvény	Sűrűségfüggvény	EX	D^2X
Cauchy	$Cauchy(a,b)$ $a\in\mathbb{R}, b>0$	$\frac{1}{\pi}\mathrm{arctg}\left(\frac{x-a}{b}\right)+\frac{1}{2}$	$\frac{1}{\pi b\left[1+\left(\frac{x-a}{b}\right)^2\right]}\quad x\in\mathbb{R}$	$\nexists$	$\nexists$
Pareto*	$Pareto(\alpha,\beta)$ $a,b>0$	$\begin{cases}1-\left(\frac{\beta}{x}\right)^{\alpha} & \text{ha } x\ge\beta\\ 0 & \text{ha } x<\beta\end{cases}$	$\begin{cases}\frac{\alpha}{\beta}\left(\frac{\beta}{x}\right)^{\alpha+1} & \text{ha } x\ge\beta\\ 0 & \text{ha } x<\beta\end{cases}$	$\frac{\alpha\beta}{\alpha-1}$	$\frac{\beta^2\alpha}{(\alpha-1)^2(\alpha-2)}$

* A Pareto-eloszlásnak akkor van véges várható értéke a képletnek megfelelően, ha $\alpha>1$, szórásnégyzete pedig akkor, ha $\alpha>2$.

Eloszlás neve	Jelölése	Eloszlás-függvény	Sűrűségfüggvény	EX	D^2X
Khí-négyzet	$\chi^2_k\quad k\in\mathbb{N}$	$\dots$	$\frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}e^{-x/2}\quad x\in\mathbb{R}$	k	$2k$
Gamma	$\Gamma(\alpha,\lambda)\quad \alpha,\lambda>0$	$\dots$	$\begin{cases}\frac{1}{\Gamma(\alpha)}\lambda^{\alpha}e^{-\lambda x}x^{\alpha-1} & \text{ha } x\ge 0\\ 0 & \text{ha } x<0\end{cases}$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
Béta	$Beta(\alpha,\beta)$ $\alpha,\beta>0$	$\dots$	$\begin{cases}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} & x\in[0;1]\\ 0 & \text{különben}\end{cases}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Lognormális	$\mathrm{LN}(m,\sigma^2)$ $m\in\mathbb{R},\sigma>0$	$\dots$	$\begin{cases}\frac{1}{x\sqrt{2\pi}\sigma}e^{-\frac{(\log x-m)^2}{2\sigma^2}} & \text{ha } x\le 0\\ 0 & \text{ha} x<0\end{cases}$	$e^{m+\sigma^2/2}$	$(e^{\sigma^2-1})e^{2m+\sigma^2}$

Többváltozós valószínűségszámítás

Definíció. **Valószínűségi vektorváltozó:** $\mathbf{X}\colon \Omega\to\mathbb{R}^d$ (Borel-)mérhető függvény, azaz amire $\{\omega:\mathbf{X}(\omega)\in B\}\in\mathcal{A}$ minden $B\subseteq\mathbb{R}^d$ nyílt (Borel-)halmazra.

Definíció. **$\mathbf{X}$ valószínűségi vektorváltozó eloszlásfüggvénye:**
$F_{\mathbf{X}}(\mathbf{x})=P(\mathbf{X}<\mathbf{x})=P(X_1<x_1,\dots,X_d<x_d)$.

Definíció. **$\mathbf{X}$ valószínűségi vektorváltozó abszolút folytonos** , ha létezik olyan $f_{\mathbf{X}}(x_1,\dots,x_d)$ függvény, amelyre
$F_{\mathbf{X}}(x_1,\dots,x_d)=\int_{-\infty}^{x_1}\dots\int_{-\infty}^{x_d}f_{\mathbf{X}}(t_1,\dots,t_d)\,dt_1\dots dt_d$.
Ilyenkor $f_{\mathbf{X}}(\mathbf{x})$-et **sűrűségfüggvény**nek hívjuk.

$d=2$ esetén vezessük be a következő jelöléseket és elnevezéseket:

- $F_{X,Y}(x,y)=P(X<x,Y<y)$ $\rightsquigarrow$ együttes eloszlásfüggvény
- $F_X(x)=P(X<x)$, $F_Y(y)=P(Y<y)$ $\rightsquigarrow$ peremeloszlásfüggvények
- $f_{X,Y}(x,y)$ $\rightsquigarrow$ együttes sűrűségfüggvény
- $f_X(x), f_Y(y)$ $\rightsquigarrow$ peremsűrűségfüggvények
  - $F_X(x)=\lim\limits_{y\to\infty}F_{X,Y}(x,y)$ és $F_Y(y)=\lim\limits_{x\to\infty}F_{X,Y}(x,y)$

Állítás.
- $F_{X,Y}(x,y)=\int_{-\infty}^{y}\int_{-\infty}^{x}f_{X,Y}(u,v)\,du\,dv$
- $f_{X,Y}(x,y)=\partial_y\partial_x F_{X,Y}(x,y)=\partial_x\partial_y F_{X,Y}(x,y)$
- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f_{X,Y}(x,y)\,dx\,dy=1$
- $f_X(x)=\int_{-\infty}^{\infty}f_{X,Y}(x,y)\,dy$ és $f_Y(y)=\int_{-\infty}^{\infty}f_{X,Y}(x,y)\,dx$

Állítás. Legyen (X,Y) abszolút folytonos, $A\subseteq\mathbb{R}$, $B\subseteq\mathbb{R}^2$ mérhető halmazok.
- $P(X\in A)=\int_{x\in A}f_X(x)\,dx$
- $P((X,Y)\in B)=\iint_{(x,y)\in B}f_{X,Y}(x,y)\,d(x,y)$

Állítás.
- X,Y függetlenek $\Leftrightarrow$ $F_{X,Y}(x,y)=F_X(x)\cdot F_Y(y)$
- X,Y függetlenek $\Leftrightarrow$ $f_{X,Y}(x,y)=f_X(x)\cdot f_Y(y)$
- X,Y függetlenek $\Leftrightarrow$ $P(X=x,Y=y)=P(X=x)\cdot P(Y=y)$
- X,Y függetlenek $\Rightarrow$ $E(XY)=EX\cdot EY$

Definíció. **X és Y kovarianciája:** $\mathrm{Cov}(X,Y)=E\left[(X-EX)(Y-EY)\right]$.
Köv.: $\mathrm{Cov}(X,Y)=E(XY)-EXEY$.
Elnevezés: ha $\mathrm{Cov}(X,Y)=0$, akkor azt mondjuk, hogy X és Y **korrelálatlanok**.

Állítás.
- X és Y függetlenek $\Longrightarrow$ X és Y korrelálatlanok
- X és Y korrelálatlanok $\not\Longrightarrow$ X és Y függetlenek !!!!!

Definíció. X **és** Y **lineáris korrelációja:** $Cor(X,Y)=\begin{cases}\frac{\mathrm{Cov}(X,Y)}{DX\,DY} & \text{ha } DX, DY>0\\ 0 & \text{ha } DX=0 \text{ v. } DY=0\end{cases}$
Ez a Pearson-féle lineáris korreláció két valószínűségi változó közti *lineáris* kapcsolat irányát és erősségét méri.

Definíció. **Kovarianciamátrix.** Legyen $\mathbf{X}$ valószínűségi vektorváltozó. Ekkor $\Sigma(\mathbf{X}) := E(\mathbf{X}\cdot\mathbf{X}^T) - E(\mathbf{X})E(\mathbf{X})^T$
A többdimenziós adatelemzés lényeges eszköze a korrelációs mátrix, aminek (i,j)-edik eleme az $R(X_i,X_j)$ lineáris korrelációs együttható. A korrelációs mátrix átlójában csupa 1-ek szerepelnek.

A többdimenziós normális és az egyenletes a gyakorlatban legtöbbször előforduló abszolút folytonos többdimenziós valószínűségi változók.
Ha $\mathbf{X}$ d **dimenziós** nem-elfajuló **normális eloszlás**t követ $\mathbf{m}$ várható érték vektorral és $\mathbf{\Sigma} > 0$ kovarianciamátrixszal (jel.: $\mathbf{X} \sim N_d(\mathbf{m},\mathbf{\Sigma})$), akkor sűrűségfüggvénye:
$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-\frac{d}{2}}|\det(\mathbf{\Sigma})|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})\right\}$.

$\mathbf{X}$ **egyenletes eloszlás**t követ a d dimenziós tér $B\subseteq\mathbb{R}^d$ mérhető részhalmazán (jel.: $\mathbf{X}\sim E(B)$), ha sűrűségfüggvénye:
$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases}\frac{1}{t(B)} & \text{ha } \mathbf{x}\in B\\ 0 & \text{egyébként}\end{cases}$ ahol a $t(\cdot)$ függvény a d dimenziós térfogatot jelöli

Tétel. **Valószínűségi vektorváltozó transzformáltjának sűrűségfüggvénye.** Legyen $\mathbf{X}=(X_1,\dots,X_n)$ abszolút folytonos valószínűségi vektorváltozó $f_{\mathbf{X}}$ sűrűségfüggvénnyel, $A\subseteq\mathbb{R}^n$ összefüggő és nyílt halmaz. Legyen $\mathbf{g}: A\to A$ függvény, amely invertálható és inverze folytonosan differenciálható. Legyen $\mathbf{Y}=\mathbf{g}(\mathbf{X})$, $\mathbf{J}=\partial_{\mathbf{y}}\mathbf{g}^{-1}(\mathbf{y})$ a Jacobi-mátrix. Ekkor
$f_{\mathbf{g}(\mathbf{X})}(\mathbf{y}) = |\det(\mathbf{J})|\cdot f_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{y}))$

Ajánlott irodalom: Márkus L. előadásfóliái a többdimenziós normális eloszlásról: `http://web.cs.elte.hu/probability/markus/ElemzoTS1/Tobbdim_normalis_elo.pdf`

Feltételes várható érték

Legyen a valószínűségi mező a szokásos $(\Omega,\mathcal{A},P)$ hármas, $\mathcal{F}\subset\mathcal{A}$ σ-algebra.

Definíció. $\mathcal{F}$**-mérhetőség.**
Az $X\colon\Omega\to\mathbb{R}$ valószínűségi változó $\mathcal{F}$-mérhető, ha minden $B\subseteq\mathbb{R}$ Borel-halmazra $X^{-1}(B)\in\mathcal{F}$.

Definíció. X **feltételes várható értéke** $\mathcal{F}$**-re nézve.**
Legyen X integrálható. Az $Y:=E(X|\mathcal{F})$ az a valószínűségi változó, amelyre egyrészt Y $\mathcal{F}$-mérhető, másrészt $\forall B\in\mathcal{F}$ halmazra $\int\limits_B X\,dP = \int\limits_B Y\,dP$.
Speciálisan, ha $\mathcal{F}=\sigma(Y)$, azaz $\mathcal{F}$-et az Y valószínűségi változó generálja, akkor $E(X|\mathcal{F})$ helyett $E(X|Y)$-t írunk.

Tehát $E(X|Y)$-ra úgy gondolunk, mint egy valószínűségi változóra, konkrétabban az Y valószínűségi változó egy mérhető $h(Y)$ függvényére; és ha Y egy adott értéket vesz fel, azaz ha $E(X|Y=y)$, akkor mint konkrét számra.

Abszolút folytonos eloszlások esetén a következő képlettel számítható:
$$\mathrm{E}(g(X)|Y) = \int\limits_{-\infty}^{\infty} g(x)f_{X|Y}(x|y)dx\Big|_{y=Y}$$
ahol $f_{X|Y}(x|y) = \begin{cases}\frac{f_{X,Y}(x,y)}{f_Y(y)} & \text{ha } f_Y(y)>0\\ 0 & \text{különben}\end{cases}$ a feltételes sűrűségfüggvény.

Definíció. σ**-algebrától való függetlenség.**
X valószínűségi változó független az $\mathcal{F}$ σ-algebrától, ha $\forall A\in\sigma(X)$ és $\forall B\in\mathcal{F}$ eseményekre $P(A\cap B)=P(A)P(B)$.

Állítás. Tulajdonságok. Legyen g $\mathcal{F}$-mérhető függvény.
- $E(X|\mathcal{F})$ 1 valószínűséggel egyértelműen létezik
- $E(E(X|\mathcal{F}))=EX$ $\rightsquigarrow$ teljes várható érték tétel (TVÉT)
- X $\mathcal{F}$-mérhető $\Rightarrow$ $E[g(X)|\mathcal{F}]=g(X)$
- X független $\mathcal{F}$-től $\Rightarrow$ $E(X|\mathcal{F})=EX$
- X $\mathcal{F}$-mérhető $\Rightarrow$ $E(XY|\mathcal{F})=XE(Y|\mathcal{F})$

Állítás. Ha X független $\mathcal{F}$-től, Y mérhető $\mathcal{F}$-re nézve, $g(X,Y)$ integrálható, akkor $E(g(X,Y)|\mathcal{F}) = E(g(X,y))\big|_{y=Y}$.

Állítás. **Teljes valószínűség tétele folytonos esetben.**
Legyen A tetszőleges esemény, Y abszolút folytonos valószínűségi változó. Ekkor $P(A)=\int\limits_{\infty}^{\infty} P(A|Y=y)f_Y(y)\,dy$.

Ajánlott irodalom: Márkus L. előadásfóliái a feltételes várható értékről: `http://web.cs.elte.hu/probability/markus/ElemzoTS1/FeltVarhErt.pdf`

Sztochasztikus folyamatok alapjai

Definíció. **Sztochasztikus folyamat:** $(X_t)_{t\in T}$, ahol T a paramétertér és minden t-re X_t valószínűségi változó.
A sztochasztikus folyamat **diszkrét paraméterű** (vagy diszkrét idejű), ha T számossága legfeljebb megszámlálhatóan végtelen, tipikusan $T=\mathbb{Z}$ vagy $T=\mathbb{Z}_+$. A sztochasztikus folyamat **folytonos paraméterű** (vagy folytonos idejű), ha T kontinuum számosságú, jellemzően $T=[0;1]$, $T=\mathbb{R}$ vagy $T=\mathbb{R}_+$. A félév során előforduló sztochasztikus folyamatok:
- Poisson-folyamat: folytonos paraméterű
- Markov-folyamat: diszkrét vagy folytonos idejű, a gyakorlaton csak a diszkrét idejűekkel foglalkozunk

- Wiener-folyamat: folytonos paraméterű
- idősormodellek (autoregresszív, mozgóátlag folyamatok): diszkrét idejű

Definíció. **Gauss-folyamat:** olyan sztochasztikus folyamat, melynek bármely véges számú peremeloszlása együttesen normális eloszlású, azaz minden $n \in \mathbf{Z}_+$, $t_1 \in T, \ldots, t_n \in T$ esetén $(X_{t_1}, \ldots, X_{t_n})$ együttesen normális eloszlású.

Idősor: Olyan sztochasztikus folyamat, amikor a T paramétertartományra '*idő*'-ként gondolunk.

Definíció. **Erős stacionaritás.** $(X_t)_{t\in T}$ erősen stacionárius, ha minden $n \in \mathbb{Z}_+$, $t_1 \in T, \ldots, t_n \in T$ és $h \in T$ esetén $(X_{t_1}, \ldots, X_{t_n})$ együttesen ugyanolyan eloszlású, mint a h-val való eltoltja, $(X_{t_1+h}, \ldots, X_{t_n+h})$.

Definíció. **Autokovariancia függvény:** $R(t,s) = \operatorname{cov}(X_t, X_s)$

Definíció. **Gyenge stacionaritás.** $(X_t)_{t\in T}$ gyengén stacionárius, ha EX_t nem függ t-től (azaz konstans), illetve az autokovariancia függvény $R(t,s)$ értéke csak a $t-s$ eltéréstől függ.
Gyengén stacionárius sztochasztikus folyamat autokovariancia függvénye tehát tulajdonképpen egyváltozós, ezt az egyváltozós függvényt is R-rel fogjuk jelölni: $R(t,s) = R(t-s)$. Tehát gyengén stacionárius sztochasztikus folyamat autokovariancia függvénye $R(h) = \operatorname{cov}(X_{t+h}, X_t)$ módon számolható.

Megjegyzés. A gyenge stacionaritásból nem következik az erős, de az erős stacionaritásból se a gyenge (nem biztos, hogy léteznek momentumai).

Megjegyzés. A *stacionaritás* szó bizonyos szempontból időbeli állandóságot, stabilitást jelent. Szeretjük, ha egy idősor stacionárius, és igyekszünk adatainkat úgy transzformálni, hogy azok "közel", illetve "látszólag" stacionáriusak legyenek.

Állítás. $R(0) = D^2X_t$ minden t-re.

Definíció. **Autokorreláció függvény (ACF):** $r(h) = \operatorname{corr}(X_t, X_{t+h})$, $h \in T$.

Állítás. $r(h) = \frac{R(h)}{R(0)}$

Definíció. **Parciális autokorreláció függvény (PACF):**
$\rho(h) = \operatorname{corr}(X_t, X_{t+h} \mid X_{t+1}, X_{t+2}, \ldots, X_{t+h-1})$, $h \in \mathbb{Z}$.

Definíció. **Független értékű zaj folyamat:** $\varepsilon_t \sim i.i.d.(0, \sigma^2)$, ha
$E\varepsilon_t = 0$, $D^2\varepsilon_t = \sigma^2$, valamint ε_t és ε_s minden $t \neq s$ esetén független egymástól.

Definíció. **Fehér zaj folyamat (white noise):**
$\varepsilon_t \sim WN(0, \sigma^2)$, ha $E\varepsilon_t = 0$, $D^2\varepsilon_t = \sigma^2$ és $\operatorname{cov}(\varepsilon_t, \varepsilon_s) = 0$, ha $t \neq s$.

Megjegyzés. gyakran kényelmes feltenni a fehér zajról, hogy Gauss-folyamat, ilyenkor Gauss-féle fehér zajról beszélünk (GWN).

Definíció. **Lineáris folyamat:**
$X_t = \mu + \sum_{j=-\infty}^{\infty} \beta_j \varepsilon_{t-j}$, ahol $\mu \in \mathbb{R}$, $\varepsilon_t \sim WN(0, \sigma^2)$ és $\sum_{j=-\infty}^{\infty} |\beta_j| < \infty$.

Megjegyzés. A lineáris folyamat definíciójában lévő ε_t zaj folyamatot *innováció*nak vagy *meghajtó folyamat*nak is szokták nevezni. Ez az elnevezés más modellek esetén is használatos.

Állítás. Az X_t lineáris folyamat stacionárius.

Definíció. **MA(∞) folyamat:** Olyan lineáris folyamat, amely nem függ a zaj jövőbeli értékeitől, azaz $\beta_{-1} = \beta_{-2} = \ldots = 0$.

Megjegyzés. MA=moving average, magyarul mozgóátlag.

Megjegyzés. Ha adott egy $x_1, \ldots, x_n$ tapasztalati minta, akkor az adatokban lévő nemstacionárius komponens kimutatására alkalmas simítási eljárás lehet (nem ez az egyetlen) a *mozgóátlagolás*. Például ha az adataink negyedévesek, akkor érdemes 4 lépésben átlagokat számítani: $\widetilde{x}_t = \frac{x_t + x_{t+1} + x_{t+2} + x_{t+3}}{4}$, $t = 1, 2, \ldots, n-3$. A mozgóátlagolással az idősor rövidebb lesz, adatokat veszítünk.

Legyen $X_1, \ldots, X_n$ minta egy stacionárius folyamatból, melynek várható értéke μ, autokovariancia függvénye $R(h)$ és autokorreláció függvénye pedig $r(h)$. Tekintsük a következő becsléseket:

- μ becslése: $\overline{X}_n = \frac{X_1 + \ldots + X_n}{n}$;
- $R(h)$ becslése: $\widehat{R}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} \left(X_{t+|h|} - \overline{X}_n\right)\left(X_t - \overline{X}_n\right)$, ahol $|h| < n$ egész;
- $r(h)$ becslése: $\widehat{r}(h) = \frac{\widehat{R}(h)}{\widehat{R}(0)}$, ahol $h = -(n-1), \ldots, -1, 0, \ldots, n-1$.

Tétel. Amennyiben $X_1, \ldots, X_n$ egy fehér zaj folyamatból származó minta, akkor $\hat{r}(h)$ megközelítőleg normális eloszlású 0 várható értékkel és $\frac{1}{n}$ szórásnégyzettel.

Következmény. Az előző tétel alapján $\hat{r}(h)$ autokorrelációkra közös $1-\alpha$ megbízhatóságú konfidenciaintervallum készíthető: $0 \pm \frac{\Phi^{-1}(1-\frac{\alpha}{2})}{\sqrt{n}}$.

Ajánlott irodalom: Márkus L. előadásfóliái idősorelméletről: `http://web.cs.elte.hu/probability/markus/ElemzoTS1/Idosorok_1_Slides_2017_12_07.pdf`

Idősorok "illeszkedésvizsgálata"

A modellező fő célja az, hogy egy adott minta (tapasztalati idősor) esetén kiválassza azt a modellt, amelyik legjobban illeszkedik az adataira. Ez gyakran meglehetősen nehéz feladat. Az idősor modellek túlnyomó többsége valamilyen fehér zaj folyamatból (amit innovációnak is szokás hívni, és gyakran megfigyelési, mérési 'hiba'-ként vagy egyéb véletlen hatások összességeként tekintenek rá) indulnak ki. Az adott idősor modell illeszkedését úgy szokás megvizsgálni, hogy az illesztés – paraméterek becslése – után visszabecsüljük az innovációkat, majd megnézzük, hogy ezek vajon fehér zaj folyamatot követnek-e. Amennyiben a visszabecsült innovációk, más néven reziduálisok esetén *nem* vethető el, hogy fehér zaj folyamatból származnak, akkor az adott idősor modell alkalmazása ellen *nem* találtunk statisztikai bizonyítékot.
Jellemzően mik tudják elrontani a 'fehér zaj'-ságot:

- autokorreláció – a reziduálisok nem autokorrelálatlanok
- heteroszkedaszticitás – a reziduálisok szórása időben változik ($\longleftrightarrow$ homoszkedaszticitás: a szórás időben konstans)

Próbák arra vonatkozóan, hogy egy adott $X_1, \ldots, X_n$ minta független fehér zajból származik-e:

a.) A minta autokorreláció függvény *pontonként* egyenlő-e 0-val. Az előző fejezet legvégén lévő tételt és következményét fogjuk itt használni.
Ha a következményben $\alpha = 0,05$, akkor közelítőleg a $\left[-\frac{2}{\sqrt{n}}; \frac{2}{\sqrt{n}}\right]$ intervallumbecslés adódik. Amikor **R** segítségével egy idősor autokorreláció függvényét elkészítjük az `acf` függvény segítségével, akkor a két szaggatott kék vonal a $\pm\frac{2}{\sqrt{n}}$ értéknek felel meg. Az alábbi esetekben döntünk amellett, hogy a minta NEM fehér zajból származik:
   - az egyik autokorreláció nagyon kiugrik, "sokkal" a kék sávon kívül van. Mi az a "sok": a sáv hosszának legalább duplája az adott autokorreláció abszolút értéke.
   - ugyan nincs durván kiugró autokorreláció érték, de viszonylag sok érték a kék sávon kívül. Hüvelyujjszabály arra, mi minősül viszonylag soknak: ha m darab autokorrelációt rajzoltunk ki, akkor több, mint $0,05 \cdot m$ autokorreláció van kicsit a kék sávon kívül. Az R alap beállítása az $m = 40$, ezen ha legalább 3 érték van kicsit a kék sávon kívül, akkor elvetjük

   Ennél rendszerint többet is látunk, a belső "sáv"-on kívül eső autokorrelációk arra is utalnak, milyen jellegű nemstacionaritással szembesülünk, illetve milyen idősor modellel érdemes megpróbálni magyarázni az adatainkat.

b.) Portmanteau-tesztek. Ezek azt vizsgálják meg, hogy az első h darab (általában 20/30) autokorreláció *együttesen* egyenlő-e 0-val. A legfontosabb ezek közül *Ljung-Box próba*. Próbastatisztika: $Q_{LB} = n(n+2)\sum_{i=1}^{h}\frac{\hat{r}(i)}{n-i}$, ami H_0 esetén χ^2_h-hoz tart eloszlásban.

Heterokszedaszticitás próba idősorok esetén: *Mcleod Li teszt*. A próba nullhipotézise az, hogy az idősor homoszkedasztikus. A Ljung-Box próbákhoz hasonlóan a próbastatisztikát itt is az első néhány tapasztalati autokorrelációból lehet számítani.

ARMA folyamatok

Definíció. **Visszaléptetés operátor (backshift operator).**
$B : L_2(\Omega) \longrightarrow L_2(\Omega), \quad BX_t = X_{t-1}$

Állítás. A fenti B operátor
- iterálható: $B^k X_t = X_{t-k}, \ k \in \mathbb{Z}_+$;
- invertálható: $B^{-1}X_t = B_{t+1}$;
- unitér.

Definíció. **ARMA folyamat.** $X_t = \underbrace{\sum_{i=1}^{p}\alpha_i X_{t-i}}_{\text{AR}(p)\text{ rész}} + \underbrace{\sum_{j=0}^{q}\beta_j\varepsilon_{t-j}}_{\text{MA}(q)\text{ rész}}$, ahol p, q pozitív egész (nevük: *rend*); $\alpha_i, \beta_j \geq 0$ valós számok; $\varepsilon_t \sim WN(0, \sigma^2)$.

Megjegyzés. ARMA: autoregresszív mozgóátlag (autregressive moving average)

Megjegyzés. Az ARMA folyamat másik alakja: $\sum_{i=0}^{p}\tilde{\alpha}_i X_{t-i} = \sum_{j=0}^{q}\beta_j\varepsilon_{t-j}$, ahol $\tilde{\alpha}_0 = 1$ és $\tilde{\alpha}_i = -\alpha_i$ $(i = 1, \ldots, p)$.

Definíció. **Kauzalitás.** Az ARMA(p, q) folyamat kauzális (vagy oksági), ha léteznek olyan $\psi_i, \ i \in \mathbb{Z}$ konstansok, amikre $\sum_{i=1}^{\infty}|\psi_i| < \infty$ és $X_t = \sum_{i=1}^{\infty}\psi_i\varepsilon_{t-i}$ $\forall t$-re.

Megjegyzés. A kauzalitás a következőket jelenti:
- az ARMA "egyenletnek" létezik "megoldása";
- az ARMA folyamatnak létezik MA(∞) alakja.

Definíció. **Invertálhatóság.** Az ARMA(p, q) folyamat invertálható, ha léteznek olyan $\pi_i, \ i \in \mathbb{Z}$ konstansok, amikre $\sum_{i=1}^{\infty}|\pi_i| < \infty$ és $\varepsilon_t = \sum_{i=1}^{\infty}\pi_i X_{t-i}$ $\forall t$-re.

Megjegyzés. Az invertálhatóság helyett azt is szokás mondani, hogy a folyamatnak van AR(∞) alakja.

Definíció. **ARMA folyamat karakterisztikus polinomjai:** $P(x) = \sum_{i=0}^{q}\tilde{\alpha}_i x^{p-i}$,
$\widetilde{P}(x) = \sum_{i=0}^{q}\tilde{\alpha}_i x^i, \ \ Q(x) = \sum_{i=0}^{q}\beta_i x^i$.

Állítás. $P(x) = x^p \cdot \widetilde{P}\left(\frac{1}{x}\right)$

Állítás. $P(x)$ gyökei az egységkörön belül vannak $\Longleftrightarrow$ $\widetilde{P}(x)$ gyökei az egységkörön kívül vannak

Tétel. **Az ARMA folyamat "megoldása" és "inverze".**
- Ha $P(x)$ gyökei az egységkörön belül vannak, akkor a folyamatnak létezik stacionárius megoldása;
- Ha $Q(x)$ gyökei az egységkörön kívül vannak, akkor a folyamatnak létezik inverze.

Következmény. Az $X_t = \alpha X_{t-1} + \varepsilon_t$ AR(1) folyamat stacionárius $\quad\Longleftrightarrow\quad |\alpha| < 1$

Eddig jelöléseinkkel az ARMA folyamat a következő operátoros alakba írható: $\widetilde{P}(B)X_t = Q(B)\varepsilon_t$. Ezt már ki tudjuk fejezni az ARMA folyamatra, illetve a fehér zajra is:
- $X_t = \left[\widetilde{P}(B)\right]^{-1} Q(B)\varepsilon_t$
- $\varepsilon_t = [Q(B)]^{-1}\widetilde{P}(B)X_t$

Az persze nem triviális, hogy a fenti operátorinverzeket, majd az operátorszorzatokat hogyan állítsuk elő, ezen a ponton segítségül kell hívni a homogén lineáris differenciaegyenletek elméletét. Legyen célunk a fenti X_t előállítás explicit megadása, a másik hasonlóan megy. Jelölje $S(x) = \sum_{i=0}^{\infty} s_i x^i$ azt a polinomot, amire $S(B) = \left[\widetilde{P}(B)\right]^{-1} Q(B)$. Ebben az inverz úgy kapható meg, hogy keressük azt a $T(x) = \sum_{i=0}^{\infty} t_i x^i$ polinomot, amire $T(x)\widetilde{P}(x) = 1$.

Az ARMA modellek kiválasztásánál segítségünkre lehetnek azok tulajdonságai:

Modell	$R(h)$ autokorreláció fv.	$r(h)$ parciális autokorreláció fv.
AR(p)	tart 0-hoz, ha $\|h\| \to \infty$	nem 0, ha $\|h\| \leq p$; egyébként 0
MA(q)	nem 0, ha $\|h\| \leq q$; egyébként 0	tart 0-hoz, ha $\|h\| \to \infty$
ARMA(p,q)	tart 0-hoz, ha $\|h\| \to \infty$, az első q érték után kezdődik a konvergencia	tart 0-hoz, ha $\|h\| \to \infty$, az első p érték után kezdődik a konvergencia

Modellválasztás: a "legjobb" ARMA(p,q) folyamat kiválasztása. Ebben segítségünkre vannak az alábbi információs kritériumok:

- **Akaike-féle információs kritérium**: $AIC = 2k - 2\log\widehat{L}$, ahol
  - k: a becsülendő paraméterek száma
  - $\widehat{L}$ a likelihood-függvény értéke akkor, ha az ML-becslést használjuk (normális eloszlású hibáknál ez megegyezik a legkisebb négyzetes becsléssel)

  Minél kisebb, annál jobb.
- **Bayes-féle információs kritérium**: $BIC = \log n \cdot k - 2\log\widehat{L} \rightsquigarrow$ minél kisebb, annál jobb

Nagyon fontos: attól még, hogy kiválasztottuk az ARMA folyamatok közül a legjobbat, nem biztos, hogy az valóban jól is írja le idősorunk dinamikáját. Ehhez még azt is meg kell nézni, hogy a hibatagok valóban fehér zajból származnak-e. Amennyiben azt kapjuk, hogy az ARMA folyamat nem illeszkedik megfelelően, számos további teendőnk lehet:

- Van-e trend/szezonalitás az idősorban (ezek kiszűrésével illik kezdeni az elemzést, lásd a következő fejezetet)
- Próbálkozás másik idősormodellel (ilyet nem csinálunk ebben a tárgyban)

Nemstacionárius idősorok modellezése

Definíció. **Lag-1 differencia operátor.** $\nabla = 1 - B$, ahol B a visszaléptetés operátor.

Definíció. **Lag-d differencia operátor.** $\nabla_d = 1 - B^d$.

Ezáltal $\nabla X_t = X_t - X_{t-1}$ és $\nabla_d X_t = X_t - X_{t-d}$. A differencia operátort tetszőleges pozitív hatványra lehet emelni, ekkor $\nabla^m X_t = \nabla^{m-1}(X_t - X_{t-1})$, ami tovább iterálható.

Definíció. **ARIMA modell.** Az X_t folyamat ARIMA(p,d,q) folyamatot követ, amennyiben $Y_t = (1-B)^d X_t$ folyamat ARMA(p,q) folyamatból származik.

Megjegyzés. Az ARIMA-ban az I betű az 'Integrated' angol szó rövidítése (integrált).

Megjegyzés. ARIMA modelleknél az integráltságot kifejező d paraméter értékének megállapításában az ún. *egységgyök tesztek* segítenek. Az egységgyök elnevezés onnan ered, hogy a modell $P(x)$ karakterisztikus polinomjának van egységgyöke (az egyik gyökének 1 az abszolútértéke), ami azzal jár, hogy a folyamat nem stacionárius.

Klasszikus idősor-dekompozíciós modell: $X_t = m_t + s_t + Y_t$, ahol

- m_t: trend komponens – valamilyen szabályosan változó függvény, ami gyakran lineáris vagy négyzetes.
- s_t: szezonális komponens – a rendszeresen ismétlődő, azonos periodicitású és szabályos amplitúdójú, rendszerint rövid távú ingadozásokat tartalmazza. Ha $d \in \mathbb{Z}$ jelöli a szezonalitást leíró periódusok hosszát, akkor $s_t = s_{t+d}$ minden t-re. Feltesszük, hogy $\sum_{i=1}^{d} s_i = 0$. Közgazdasági alkalmazásokban a d értéke negyedéves adatoknál jellemzően 4, havi adatoknál pedig 12.
- Y_t: véletlen zaj tag, egy olyan stacionárius komponens, amit már valamilyen ismert idősor-modellel modellezhetünk. Feltesszük, hogy $EY_t = 0$, különben a konstans beolvasztható lenne a trend tagba.

A nemstacionárius idősoroknál a trend hatás kimutatására, illetve eltüntetésére két megközelítést lehet követni:

1. Trend becslése
   - *Paraméteres*: Alkalmas függvényt illesztünk, ami rendszerint egy polinom szokott lenni, azaz $m_t = \sum_{i=0}^{p} c_k t^k$ alakú, a c_k együtthatókat pedig legkisebb négyzetek módszerével lehet becsülni. Ezen a ponton kihasználhatjuk, hogy egy ilyen alakú regressziós modell a lineáris modell speciális esete.
     A paraméteres megközelítés hátránya, hogy feltesszük, a választott függvény a jövőben is jól fogja leírni az idősor dinamikáját, márpedig egy válság vagy akár egy váratlan pozitív esemény hatására erre nincs semmi biztosíték.
   - *Nemparaméteres*: Ilyenekre nem lesz idő.
2. Trend eliminálása differenciálással – annyiszor alkalmazzuk a ∇ differencia operátort a folyamatra, amíg el nem tűnik a trend. Például lineáris trend esetén már egyszeres differenciálás is eltünteti a trend komponenst.

Most áttekintjük, hogy mennyivel van több dolgunk, amennyiben a szezonális hatásokkal is kezdeni szeretnénk valamit. A nemstacionárius idősoroknál a trend ÉS a szezonális hatás kimutatására, illetve eltüntetésére két megközelítést lehet követni:

1. Trend és szezonalitás becslése
   Legyen $x_1, x_2, \ldots, x_n$ a tapasztalati mintánk. Először az előzőkben leírt valamilyen paraméteres vagy nemparaméteres módszerrel kiszűrjük az $\hat{m}_t$ trend hatást, majd az $x_t - \hat{m}_t$ eltérésekből átlagolással megbecsüljük az egyedi szezonhatásokat. Végül ezek átlagával korrigálunk, hogy az összegük 0 legyen. Tehát a két

lépés a szezonhatások számszerűsítésére:

I. Korrigálatlan egyedi szezonindexek becslése (Létrehozunk d darab halmazt, amelyekbe minden d-edik eltérést teszünk be, majd az egyes halmazokban lévő számok átlagát számítjuk. Például az első halmazba tartozik az 1.,(d+1).,(2d+1). stb. eltérések, a másodikba a 2.,(d+2).,(2d+2). stb. eltérések):

$$\tilde{s}_k = \frac{\sum\limits_{i:\ 1\le k+id\le n} (x_{k+id} - \hat{m}_{k+id})}{\sum\limits_{i:\ 1\le k+id\le n} 1}, \quad k = 1, 2, \ldots, d$$

II. Egyedi szezonindexek korrigálása: $\hat{s}_k = \tilde{s}_k - \frac{1}{d}\sum\limits_{i=1}^{d} \tilde{s}_i, \quad k = 1, 2, \ldots, d$

2. Trend és szezonalitás eliminálása differenciálással – annyiszor alkalmazzuk a ∇ differencia operátort a folyamatra, amíg el nem tűnik a trend. Ezután megnézzük, hogy maradt-e még szezonalitás a reziduálisokban, és ha igen, akkor alkalmas d-vel alkalmazzuk ∇_d differencia operátort. A d kiválasztásában segítségünkre lehet a folyamat ábrája, illetve az ACF/PACF függvények.
3. A fentiek kombinálása. Gyakori, hogy a trendet differenciálással szűrjük ki, majd a szezonális hatásokat szezonindexek számításával.

Most foglaljuk össze eddigi idősor-modellezési ismereteinket! Amennyiben az erős időbeli összefüggőséget mutató tapasztalati mintánkat az időtartományban (time domain, szemben a gyakorisági tartománnyal – frequency domain) szeretnénk modellezni, akkor a következő lépéseket kell követni.

Az idősor-modellezés fő lépései (Box-Jenkins modellezés):

1. Az idősor ábrázolása vonaldiagrammal
   - ránézésre homogén, hasonlóan kinéző részekre bontani (amelyek elegendő mintaelemet tartalmaznak)
   - kiugró/hibás/hiányzó értékek kezelése: kihagyás/javítás/békén hagyás
2. Előzetes transzformáció, például ha exponenciálisan nő az ábra alapján, akkor érdemes logaritmust venni.
3. Trend komponens kiszűrése
4. Szezonális komponens kiszűrése
5. A megfelelő modelltípus, modellcsalád választása (rendszerint ARMA/ARIMA) után paraméterbecslés, a legjobb modell kiválasztása valamelyik információs kritérium alapján
6. Modelldiagnosztika – az becsült együtthatók szignifikánsak-e, illetve a modellből visszaszámolt reziduálisok fehér zaj folyamatot követnek-e.
7. Előrejelzés: általában ez a végső cél, szeretnénk meglévő adataink alapján az idősor jövőbeli viselkedésére minél jobb jóslást adni.

A fenti felsorolásban a 2. és 3. pontok során az idősorban lévő nemstacionárius tago(ka)t szedjük ki.