Kárszámeloszlások modellezése

DIPLOMAMUNKA

Írta: Talabér Dóra Edit

Biztosítási és pénzügyi matematika MSc
Aktuárius szakirány

Témavezető:

Prokaj Vilmos
egyetemi docens
ELTE TTK Valószínűségelméleti és Statisztika Tanszék

Eötvös Loránd Tudományegyetem
Természettudományi Kar

Budapesti Corvinus Egyetem
Közgazdaságtudományi Kar

2014

Tartalomjegyzék

Köszönetnyilvánítás

Ezúton szeretném kifejezni köszönetemet témavezetőmnek, Prokaj Vilmosnak, amiért kérdéseimre mindig körültekintően válaszolt és segített a dolgozat tartalmi és stilisztikai hibáinak korrigálásában.

Szeretném megköszönni Szüleimnek a rengeteg biztatást és támogatást, amivel hozzájárulnak tanulmányaim sikerességéhez. Köszönöm, hogy mindenben mellettem állnak. Köszönettel tartozom Kiss Blankának és Gombár Tamásnak bátorításukért és a sok segítségért, melyet a mesterképzés ideje alatt kaptam.

Bevezetés

Kárszámok modellezésénél gyakran feltesszük, hogy egy szerződő különböző típusú kárainak száma egymástól független. A dolgozatban a kétváltozós Poisson eloszlás felhasználásával megvizsgálom, hogy ha a kötvénytulajdonosok kétféle biztosítással rendelkeznek, akkor fellelhető-e összefüggés a kétfajta kárszámuk között. A kétféle biztosítás, amit vizsgálni fogok a kötelező gépjármű-felelősségbiztosítás és ugyanazon gépjárműre kötött casco biztosítás. Természetesen lehetne más nem-életbiztosítási kárszámot is vizsgálni.

A dolgozat első fejezetében áttekintem azokat a nevezetes eloszlásokat, amelyek jól használhatóak kárszámok leírására. Röviden összefoglalom fontos tulajdonságaikat: milyen paraméterrel vagy paraméterekkel jellemezhetőek és azoknak mi a jelentésük.

A második fejezetben bemutatom a kétváltozós Poisson eloszlást, mely több tudományos területen játszik fontos szerepet. Biztosítási kárszámok eloszlását is gyakran szokták Poisson eloszlásúnak feltételezni. A kétféle kárszám közötti összefüggést nem csak a kétváltozós Poisson eloszlással, hanem annak egy módosításával is próbálom megragadni.

A dolgozat harmadik fejezetében az EM algoritmus ismertetésére kerül sor. Az EM algoritmus egy népszerű iteratív módszer eloszlások paramétereinek maximum likelihood becslésére. A likelihood függvény logaritmusának szélsőérték problémája bizonyos esetekben nagyon komplikált és nehezen kezelhető számításokhoz vezethet, és sokszor csak nagy számításkapacitások árán lehet megoldani. Előfordulhat az is, hogy analitikusan nem oldható meg a probléma, ilyenkor numerikus algoritmusok segítségével érhetünk célt. Bemutatom, hogy az EM algoritmus hogyan alkalmazható a paraméterek becslésére normális eloszlások keveréke esetén.

Egy általánosabb megközelítés, ha az egyes kárszámok paraméterét vagy paramétereit a kötvénytulajdonos és a gépjármű bizonyos tulajdonságaitól tesszük függővé. Például figyelembe vesszük a vezető nemét, életkorát, lakhelyét, az au-

tó használatának célját (magán vagy vállalati), típusát, stb. Az így kapott regressziós modell paramétereinek becslésére használható az EM algoritmus, ennek tárgyalására kerül sor a negyedik fejezetben.

1. fejezet

Kárszámeloszlások

Ebben a részben a legismertebb kárszámeloszlások bemutatására és néhány tulajdonságuk ismertetésére kerül sor. Ezek a binomiális, geometriai, negatív binomiális és Poisson eloszlás. Mivel az ilyen eloszlású valószínűségi változók csak nemnegatív egész értéket vesznek fel, ezért kárszámok modellezésére jól használhatóak. Mindezek előtt nézzünk néhány definíciót.

1.1. Definíció. *Az X valószínűségi változó momentumgeneráló függvénye a*

$$t \in \mathbb{R} \longmapsto M_X(t) = \mathrm{E}(e^{tX})$$

függvény.

Az értelmezési tartomány azokból a valós számokból áll, ahol a fenti várható érték létezik. Minden eloszlásra $M_X(0) = 1$. Ha M_X véges a 0 kis környezetében, akkor ott akárhányszor differenciálható. A deriváltak 0 helyen vett értéke az X valószínűségi változó momentumait adják. Például

$$M_X' = \mathrm{E}X,$$

általában pedig

$$M_X^{(k)} = \mathrm{E}(X^k).$$

1.2. Definíció. *Az X nemnegatív egész értékű valószínűségi változó generátorfüggvénye*

$$G_X(z) = \mathrm{E}(z^X) = \sum_{k=0}^{\infty} z^k \mathrm{P}(X = k).$$

A generátorfüggvény segítségével meg lehet határozni az eloszlás elemeit. A következő kapcsolat írja le a köztük lévő összefüggést:

$$\frac{G_X^{(k)}(0)}{k!} = \mathrm{P}(X = k).$$

Ehhez szükséges a generátorfüggvény konvergenciája, ami a zárt egységkörlapon mindig teljesül. Vagyis a generátorfüggvény egyértelműen meghatározza az eloszlást, és fordítva.

1.1. Nevezetes kárszámeloszlások

1.1.1. Binomiális eloszlás

Egy X diszkrét értékű nemnegatív valószínűségi változó binomiális eloszlást követ n és p paraméterekkel, ha

$$p_k = \mathrm{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

ahol n pozitív egész, $k = 0, 1, \ldots, n$ és $0 < p < 1$. Az X változó n darab egymástól független kísérlet során bekövetkezett sikeres kísérletek száma, egy siker bekövetkeztének valószínűsége p. A várható érték és a szórásnégyzet

$$\mathrm{E}X = np \quad \text{és} \quad \mathrm{D}^2X = np(1-p).$$

Az X változó szórásnégyzete mindig kisebb, mint a várható értéke. Az X momentumgeneráló függvénye

$$M_X(t) = \mathrm{E}(e^{tX}) = \sum_{k=0}^{\infty} e^{tk}\mathrm{P}(X = k) = \sum_{k=0}^{\infty} e^{tk}\binom{n}{k} p^k (1-p)^{n-k} =$$
$$= \sum_{k=0}^{\infty} \binom{n}{k} (e^t p)^k (1-p)^{n-k} = (e^t p - 1 - p)^n.$$

A generátorfüggvény

$$G_X(z) = \mathrm{E}(z^X) = (pz + 1 - p)^n.$$

1.1.2. Geometriai eloszlás

Egy nemnegatív diszkrét értékű X valószínűségi változó geometriai eloszlást követ p paraméterrel, ha

$$p_k = \mathrm{P}(X = k) = p(1-p)^k,$$

ahol $k = 0, 1, \ldots$ és $0 < p < 1$. Az X változó az egymástól függetlenül végzett kísérletek során az első sikeres kísérlet bekövetkezte előtti sikertelen kísérletek számát adja meg, ahol a siker valószínűsége p. A várható érték és a szórásnégyzet

$$\mathrm{E}X = \frac{1-p}{p} \quad \text{és} \quad \mathrm{D}^2X = \frac{1-p}{p^2}.$$

Az X változó szórásnégyzete mindig kisebb, mint a várható értéke. Az X momentumgeneráló függvénye

$$M_X(t) = \mathrm{E}(e^{tX}) = \sum_{k=0}^{\infty} e^{tk}\mathrm{P}(X=k) = \sum_{k=0}^{\infty} e^{tk}p(1-p)^k = \\ = p\sum_{k=0}^{\infty}(e^t(1-p))^k = \frac{p}{1-e^t(1-p)}.$$

A generátorfüggvény

$$G_X(z) = \mathrm{E}(z^X) = \frac{p}{1-z(1-p)}.$$

1.1.3. Negatív binomiális eloszlás

Egy X diszkrét értékű nemnegatív valószínűségi változó negatív binomiális eloszlást követ r és p paraméterekkel, ha

$$p_k = \mathrm{P}(X=k) = \binom{k+r-1}{r-1}p^r(1-p)^k,$$

ahol r pozitív egész és $0 < p < 1$. Az X változó az egymástól függetlenül végzett kísérletek során az r. sikeres kísérlet bekövetkezte előtti sikertelen kísérletek számát adja meg, ahol a siker valószínűsége p. A várható érték és a szórásnégyzet

$$\mathrm{E}X = \frac{r(1-p)}{p} \quad \text{és} \quad \mathrm{D}^2X = \frac{r(1-p)}{p^2}.$$

Az X változó szórásnégyzete mindig nagyobb, mint a várható értéke. Az X momentumgeneráló függvénye

$$M_X(t) = \mathrm{E}(e^{tX}) = \sum_{k=0}^{\infty} e^{tk}\mathrm{P}(X=k) = \sum_{k=0}^{\infty} e^{tk}\binom{k+r-1}{r-1}p^r(1-p)^k = \\ = \frac{p^r}{(1-e^t(1-p))^r}\sum_{k=0}^{\infty}\binom{k+r-1}{r-1}(e^t(1-p))^k(1-e^t(1-p))^r = \\ = \left(\frac{p}{1-e^t(1-p)}\right)^r.$$

A generátorfüggvény

$$G_X(z) = \mathrm{E}(z^X) = \left(\frac{p}{1-z(1-p)}\right)^r.$$

1.1.4. Poisson eloszlás

Egy X diszkrét értékű nemnegatív valószínűségi változó Poisson eloszlást követ λ paraméterrel, ha

$$p_k = \mathrm{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

ahol $\lambda > 0$. A várható érték és a szórásnégyzet

$$\mathrm{E}X = \lambda \quad \text{és} \quad \mathrm{D}^2 X = \lambda.$$

Kifejezi az adott idő alatt ismert valószínűséggel megtörténő események bekövetkezésének számát (például: egy telefonközpontba adott időszakban és időtartamban beérkezett telefonhívások száma, vagy egy radioaktív anyag adott idő alatt elbomló atomjainak száma). Az X változó szórásnégyzete egyenlő a várható értékével. Az X momentumgeneráló függvénye

$$M_X(t) = \mathrm{E}(e^{tX}) = \sum_{k=0}^{\infty} e^{tk} \mathrm{P}(X = k) = \sum_{k=0}^{\infty} e^{tk} \frac{\lambda^k}{k!} e^{-\lambda} =$$
$$= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^t \lambda)^k}{k!} = \exp\left\{-\lambda(1 - e^t)\right\}$$

A generátorfüggvény

$$G_X(z) = \mathrm{E}(z^X) = e^{-\lambda(1-z)}$$

1.2. Az (a, b, 0) eloszlás család

Az előző fejezetben bemutatott eloszlások tagjai egy rekurziós összefüggésnek is eleget tesznek. Így a binomiális, a negatív binomiális és a Poisson eloszlás az $(a, b, 0)$ eloszlás családhoz tartozik.

1.3. Definíció. *Egy η nemnegatív egész értékű valószínűségi változó $(a, b, 0)$ eloszlású, ha teljesíti a következő rekurziós feltételt*

$$p_k = \left(a + \frac{b}{k}\right) p_{k-1}, \qquad k = 1, 2, \ldots,$$

ahol a és b adott konstansok.

A következő tétel állítása szerint $(a, b, 0)$ eloszlás családba pontosan az előbbi három eloszlás tartozik bele.

1.4. Tétel. *Az η egész értékű nem negatív valószínűségi változó pontosan akkor tartozik az $(a, b, 0)$ eloszlás osztályba, ha binomiális, negatív binomiális vagy Poisson eloszlású.*

A bizonyítás részletes tárgyalására nem térünk ki. Arról, hogy az említett három eloszlás $(a, b, 0)$, egyszerű számítással megbizonyosodhatunk, ha felhasználjuk az 1.1. táblázatot. Ez épp azt mutatja, hogy milyen kapcsolat van az egyes eloszlások és az a és b paraméterek között. A tétel állításának másik iránya esetszétválasztással látható be.

	a	b	p_0
binomiális	$-\frac{p}{1-p}$	$(m+1)\frac{p}{1-p}$	$(1-p)^n$
negatív binomiális	$1-p$	$(r-1)(1-p)$	p^r
Poisson	0	λ	$e^{-\lambda}$

1.1. táblázat. A binomiális, negatív binomiális vagy Poisson eloszlások paramétereinek kapcsolata az a és b konstansokhoz

2. fejezet

Kétváltozós Poisson eloszlás

A kétváltozós diszkrét eloszlások több kutatási területen is fontos szerepet játszanak. Ilyenek például az egészségügyi, marketing vagy sport statisztikával foglalkozó területek. Egy orvosi gyógymód vizsgálatánál mérhetik ugyanazon személy szükséges kezeléseinek számát a műtét előtt és után, vagy két különböző betegség megjelenését. Marketingben például vizsgálják két termék vagy termékcsoport fogyasztói igényét. Sport mérkőzésen két egymás ellen játszó csapat által szerzett pontok vagy gólok számának modellezése. A kétváltozós eloszlások közül az egyik leggyakrabban használt a kétváltozós Poisson eloszlás, mely népszerű az ilyen típusú adatok modellezésében.

A most következő fejezetben áttekintjük a kétváltozós Poisson eloszlás néhány tulajdonságát. Azután a fejezet második részében a diagonálisan módosított Poisson eloszlás ismertetése következik. Ehhez a [9] könyvre és [1], [8] cikkekre támaszkodunk.

2.1. Kétváltozós Poisson eloszlás

A fejezet bevezetésében említett példákon túl a kétváltozós Poisson eloszlás használható kárszámok modellezésére is. Jelölje X és Y a károk számát, és legyen $N=X+Y$. Tegyük fel, hogy minden szerződő két különböző kárszámmal rendelkezik. Például egy kötelező gépjármű felelősség és egy casco biztosítási kárszámmal. A díj kiszámítása a kárszámok függetlensége mellett a következőképpen írható le, legyen

$$X \sim \text{Poisson}(\lambda_1) \quad \text{és} \quad Y \sim \text{Poisson}(\lambda_2)$$

függetlenek. Ebből következik, hogy

$$N \sim \text{Poisson}(\lambda_1 + \lambda_2).$$

A λ_1 és λ_2 paraméterek függnek a vezető és a gépjármű tulajdonságaitól. Ezek a paraméterek minden kötvénytulajdonosra becsülhetők. Feltételezzük, hogy a várható kárnagyság egy pénzegység, így a nettó díjelvvel számolt díj

$$\Pi = \text{E}(N) = \text{E}(X) + \text{E}(Y) = \lambda_1 + \lambda_2$$

alakba írható. Ezt az összeget még általában növelni szokták, hogy a biztosító átlagosan ne veszítsen pénzt. Ehhez többféle díjelvet is lehet használni, ezek közül egyik a szórásnégyzet díjelv. Ez hasonlóan az előző díjelvhez az összkárszám várható értékét tartalmazza, valamint egy kockázati tagot, ami a szórásnégyzettel arányos. Jelölje az így kapott díjat Π^*. Kihasználva, hogy a Poisson eloszlású valószínűségi változók esetében a várható érték megegyezik a szórásnégyzettel, kapjuk a

$$\Pi^* = \text{E}(N) + \alpha \text{D}^2(N) = (1+\alpha)(\lambda_1 + \lambda_2)$$

összefüggést.

A kétváltozós Poisson eloszlásban nem követeljük meg a függetlenségi feltételt. Az X és Y változókat a következőképpen írjuk fel:

$$\begin{aligned} X &= X_1 + X_3 \\ Y &= X_2 + X_3, \end{aligned}$$

ahol X_i-k független Poisson eloszlású valószínűségi változók $\lambda_i (i = 1, 2, 3)$ paraméterekkel.

Vizsgáljuk meg a két változó közötti kapcsolatot. Az X és Y konstrukciójából következik, hogy

$$\text{cov}(X, Y) = \text{cov}(X_1 + X_3, X_2 + X_3) = \text{cov}(X_3, X_3) = \lambda_3,$$

vagyis λ_3 paraméter a két változó közötti összefüggőség mértéke. A $\lambda_3 = 0$ esetben visszakapjuk a két független változós esetet. Mivel Poisson eloszlású független változók összege Poisson és $X_i (i = 1, 2, 3)$ változók függetlenek, ezért a marginális eloszlások is Poissonok lesznek, $X \sim \text{Poisson}(\lambda_1 + \lambda_3)$ és $Y \sim \text{Poisson}(\lambda_2 + \lambda_3)$. A változók közötti korreláció pedig

$$R(X, Y) = \frac{\lambda_3}{\sqrt{(\lambda_1 + \lambda_3)(\lambda_2 + \lambda_3)}}.$$

A nettó díjelvvel számolt díj emelkedik ahhoz képest, mint amit a változók függetlenségének estében kaptunk.

$$\Pi = \mathrm{E}(N) = \mathrm{E}(X+Y) = \lambda_1+\lambda_3+\lambda_2+\lambda_3$$

A kárszám szórásnégyzete

$$\mathrm{D}^2(N) = \mathrm{D}^2(X+Y) = (\lambda_1+\lambda_3)+(\lambda_2+\lambda_3)+2\lambda_3 = \lambda_1+\lambda_2+4\lambda_3$$

Kárszámok modellezésénél gyakran előforduló jelenség az ún. „overdispersion", ami azt jelenti, hogy a változó szórásnégyzete nagyobb, mint a várható értéke. Amíg a két kárszám (X és Y) független volt, addig ezt a jelenséget nem figyelhettük meg, hiszen az N szórásnégyzete és a várható értéke egymással egyenlő volt. A függetlenségi feltevés elhagyásával azonban látható, hogy a kárszám szórásnégyzete nagyobb, mint a várható értéke.

Az X és Y együttes eloszlása kétváltozós Poisson eloszlást követ.

$$(X,Y) \sim \mathrm{BP}(\lambda_1,\lambda_2,\lambda_3)$$

Az együttes eloszlás:

$$\mathrm{P}(X=x,Y=y) = e^{-(\lambda_1+\lambda_2+\lambda_3)}\frac{\lambda_1^x\lambda_2^y}{x!\,y!}\sum_{i=0}^{\min(x,y)}\binom{x}{i}\binom{y}{i}i!\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^i \tag{2.1}$$

Ennek belátásához írjuk fel a generátorfüggvényt. Az átalakítások során felhasználjuk az X_1, X_2, X_3 változók függetlenségét.

$$\begin{aligned} G_{X,Y}(z_1,z_2) &= \mathrm{E}(z_1^X z_2^Y) = \mathrm{E}(z_1^{X_1+X_3} z_2^{X_2+X_3}) = \mathrm{E}(z_1^{X_1} z_2^{X_2}(z_1z_2)^{X_3}) = \\ &= \exp\{\lambda_1(z_1-1)+\lambda_2(z_2-1)+\lambda_3(z_1z_2-1)\} = \\ &= \exp\{(\lambda_1+\lambda_3)(z_1-1)+(\lambda_2+\lambda_3)(z_2-1)+\lambda_3(z_1-1)(z_2-1)\} \end{aligned} \tag{2.2}$$

Felhasználva az exponenciális függvény hatványsorát a (2.2) kifejezés átalakítható

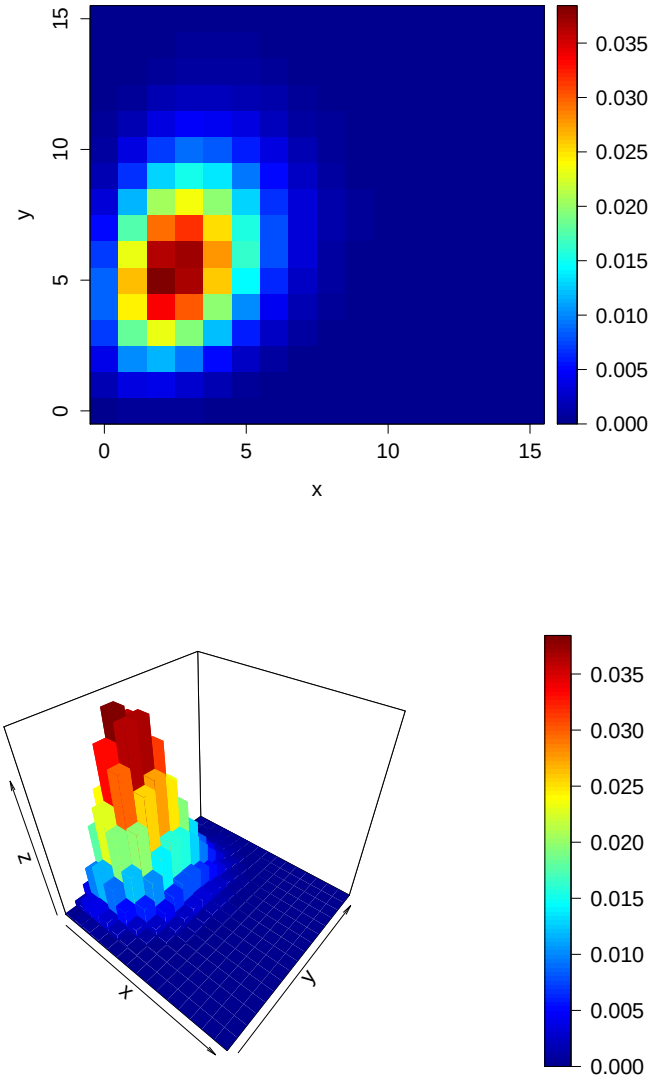
$$\begin{aligned} G_{X,Y}(z_1,z_2) &= \mathrm{E}(z_1^{X_1} z_2^{X_2}(z_1z_2)^{X_3}) = \\ &= \exp\{-(\lambda_1+\lambda_2+\lambda_3)\}\sum_{i=0}^{\infty}\frac{\lambda_1^i z_1^i}{i!}\sum_{j=0}^{\infty}\frac{\lambda_2^j z_2^j}{j!}\sum_{k=0}^{\infty}\frac{\lambda_3^k z_1^k z_2^k}{k!} = \\ &= \exp\{-(\lambda_1+\lambda_2+\lambda_3)\}\sum_{x,y}\sum_{i}\frac{\lambda_1^{x-i}\lambda_2^{y-i}\lambda_3^i}{(x-i)!\,(y-i)!\,i!}z_1^x z_2^y \end{aligned}$$

alakba. Innen kapjuk az együttes eloszlásfüggvény tagjait, ami a $z_1^x z_2^y$ együtthatója. Átalakítással megkapjuk az eloszlás fenti alakját.

$$\begin{aligned} \mathrm{P}(X=x,Y=y) &= e^{-(\lambda_1+\lambda_2+\lambda_3)}\sum_{i=0}^{\min(x,y)}\frac{\lambda_1^{x-i}\lambda_2^{y-i}\lambda_3^i}{(x-i)!\,(y-i)!\,i!} = \\ &= e^{-(\lambda_1+\lambda_2+\lambda_3)}\frac{\lambda_1^x\lambda_2^y}{x!\,y!}\sum_{i=0}^{\min(x,y)}\binom{x}{i}\binom{y}{i}i!\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^i \end{aligned}$$

Az eloszlás momentumgeneráló függvénye:

$$M_{X,Y}(t_1, t_2) = \exp\left\{\lambda_1(e^{t_1} - 1) + \lambda_2(e^{t_2} - 1) + \lambda_3(e^{t_1+t_2} - 1)\right\}$$



2.1. ábra. A BP(2, 5, 1) eloszlás 2- és 3-dimenziós ábrája a $[0, 15] \times [0, 15]$ tartományon.

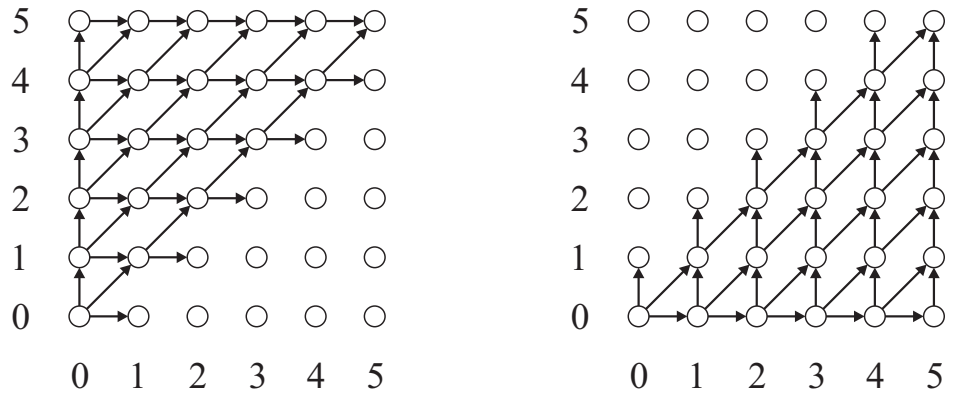
Az együttes eloszlás tagjai egy rekurziós formula segítségével is számolhatók.

Vezessük be a $f(x,y) = \mathrm{P}(X = x, Y = y)$ jelölést. Az eloszlás első néhány tagja

$$\begin{aligned} f(0,0) =& e^{-(\lambda_1+\lambda_2+\lambda_3)} \\ f(x,0) =& e^{-(\lambda_2+\lambda_3)}\mathrm{P}(X = x) = e^{-(\lambda_2+\lambda_3)}\frac{e^{-\lambda_1}\lambda_1^x}{x!} \\ f(0,y) =& e^{-(\lambda_1+\lambda_3)}\mathrm{P}(Y = y) = e^{-(\lambda_1+\lambda_3)}\frac{e^{-\lambda_2}\lambda_2^y}{y!} \end{aligned}$$

formában írható le. Az $x \geq 1$ és $x \geq 1$ esetben pedig a következő kapcsolat írható fel:

$$\begin{aligned} xf(x,y) &= \lambda_1 f(x-1,y) + \lambda_3 f(x-1,y-1), \\ yf(x,y) &= \lambda_2 f(x,y-1) + \lambda_3 f(x-1,y-1). \end{aligned}$$



2.2. ábra. A rekurziós összefüggéseket szemléltető ábra. A nyilak mutatják, hogy az egyes pontokban a valószínűség számításához mely értékek ismerete szükséges.

2.2. Diagonálisan módosított kétváltozós Poisson eloszlás

A kárszámok modellezésénél előfordul, hogy a kármentesség nagyobb valószínűségű, mint az az illesztett eloszlás szerint várható lenne. Ugyanez igaz lehet az (1,1) cellára is, hiszen egy baleset bekövetkezése maga után vonhatja a másik fajta kár bekövetkezését is. Ezért a következőkben definiáljuk a diagonálisan módosított kétváltozós Poisson eloszlást, ami ezeket a tulajdonságokat is hordozza.

Azt mondjuk, hogy N' = (X', Y') diagonálisan módosított kétváltozós Poisson eloszlású, ha egy kétváltozós Poisson $N = (X,Y)$ és egy (D,D) alakú „diagonális"

változó keveréke, valamely $p \in [0,1]$ keverési aránnyal. Kicsit pontosabban, legyen (X,Y) kétváltozós Poisson, D tetszőleges nemnegatív egész értékű változó és ξ tőlük független p paraméterű indikátor. Ekkor

$$X' = (1-\xi)X + \xi D, \qquad Y' = (1-\xi)Y + \xi D$$

diagonálisan módosított kétváltozós Poisson eloszlású.

$$f_{DM}(x,y) = \begin{cases} (1-p)f(x,y|\lambda_1,\lambda_2,\lambda_3) & \text{ha } x \neq y, \\ (1-p)f(x,y|\lambda_1,\lambda_2,\lambda_3) + pf_D(x|\theta) & \text{ha } x = y, \end{cases}$$

ahol f a kétváltozós Poisson eloszlás, f_D a $\{0,1,2,\ldots\}$ halmazon definiált $D(x;\theta)$ „diagonális" változó súlyfüggvénye θ paramétervektorral és $p \in [0,1]$. A $p=0$ esetben visszakapjuk a kétváltozós Poisson eloszlást. A $D(x;\theta)$ eloszlását választhatjuk Poisson eloszlásúnak, geometriai eloszlásúnak, vagy tetszőleges véges tagszámú eloszlás is lehet, ezt jelölje $Disc(J)$. Utóbbi esetben az eloszlás

$$f(x|\theta,J) = \begin{cases} \theta_x & \text{ha } x = 0,1,\ldots,J, \\ 0 & \text{különben,} \end{cases} \tag{2.3}$$

ahol $\sum_{x=0}^{J} \theta_x = 1$. Triviálisan következik, hogy ha $J=0$, akkor azt az esetet kapjuk, amikor csak a (0,0) cellát módosítjuk.

A diagonálisan módosított Poisson eloszlás abban tér el az előző alfejezetben ismertetett kétváltozós Poisson eloszlástól, hogy a marginális eloszlások nem Poissonok. Az X marginális eloszlása egy Poisson és egy diszkrét eloszlás keveréke.

$$f_{DM}(x) = (1-p)f_{Po}(x|\lambda_1+\lambda_3) + pf_D(x|\theta),$$

ahol $f_{Po}(.|\lambda)$ a λ paraméterű Poisson eloszlás. Az X változó marginális várható értéke és szórásnégyzete:

$$\begin{aligned} \mathrm{E}(X) &= (1-p)(\lambda_1+\lambda_3) + p\mathrm{E}_D(X) \\ \mathrm{D}^2(X) &= (1-p)[\lambda_1+\lambda_3+(\lambda_1+\lambda_3)^2] + p\mathrm{E}_D(X^2) - [(1-p)(\lambda_1+\lambda_3) + p\mathrm{E}_D(X)]^2 \end{aligned}$$

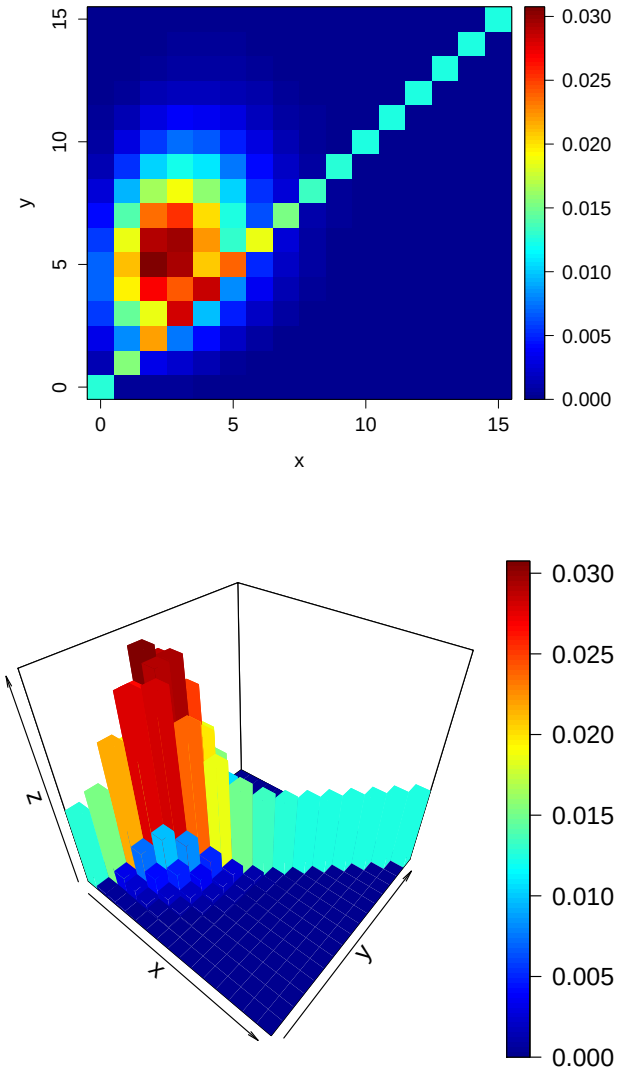
Ehhez a következő összefüggéseket használjuk fel:

$$\begin{aligned} \mathrm{D}^2(X) &= \mathrm{E}(X^2) - (\mathrm{E}(X))^2 \\ \mathrm{E}(X^2) &= (1-p)\mathrm{E}_{Po}(X^2) + p\mathrm{E}_D(X^2) = (1-p)[\lambda_1+\lambda_3+(\lambda_1+\lambda_3)^2] + p\mathrm{E}_D(X^2) \end{aligned}$$

A diagonális változó eloszlásától függően X várható értéke és szórásnégyzete között mindkét reláció előfordulhat.

I. $\mathrm{E}(X) \geq \mathrm{D}^2(X)$
Legyen $D(x;\theta) \in Disc(1)$, $\theta = (0,1)^T$, $\lambda_1 + \lambda_3 = 1$ és $p = 0{,}5$.
Ekkor $\mathrm{E}(X) = 1$ és $\mathrm{D}^2(X) = 0{,}5$.

II. $\mathrm{E}(X) \leq \mathrm{D}^2(X)$
Legyen $D(x;\theta) \in Disc(0)$ $\theta = 1$, $\lambda_1 + \lambda_3 = 1$ és $p = 0{,}5$.
Ekkor $\mathrm{E}(X) = 0{,}5$ és $\mathrm{D}^2(X) = 0{,}75$.



2.3. ábra. A BP(2, 5, 1) és a $Disc(15)$ $p = 0.2$ keverési arányú eloszlás 2- és 3-dimenziós ábrája a $[0,15] \times [0,15]$ tartományon, ahol $D(x;\theta)$ egyenletes diszkrét eloszlású [0,15] intervallumon.

3. fejezet

Az EM algoritmus

Az EM algoritmus egy népszerű módszer eloszlások paramétereinek maximum likelihood becslésére. A megfigyelt mintaelemeket jelölje $x_1, x_2, \ldots, x_n$, melyek ugyanabból az $f(x|\theta)$ eloszlásból származnak. Feltéve, hogy függetlenek a mintaelemek, a likelihood függvény

$$L(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

Erre úgy gondolhatunk, mint a θ paraméter függvénye adott $x_1, x_2, \ldots, x_n$ esetén. Célunk, hogy megtaláljuk azt a θ^* paramétert, ami maximalizálja L-et.

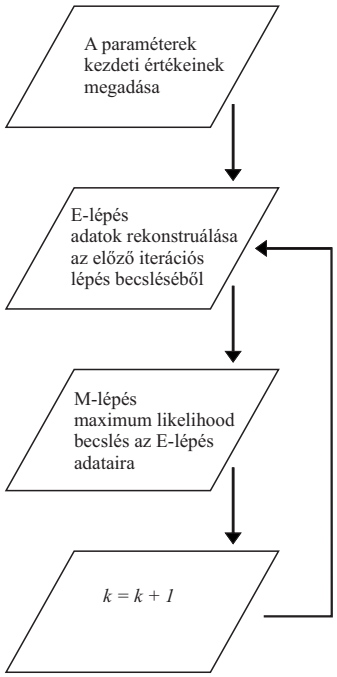
$$\theta^* = \arg\max_{\theta} L(x|\theta).$$

A maximum likelihood becslés még teljes adatszerkezetekre is sokszor bonyolult, előfordulhat, hogy csak közelítő eljárással lehet kiszámolni. Az EM algoritmus akkor is alkalmazható, ha nem áll rendelkezésre minden adat, cenzoráltak a mérések, a modell látens változót tartalmaz vagy keverékfelbontás paramétereit szeretnénk becsülni.

Az algoritmus két lépésből áll, melyek közül az első egy feltételes várható értékkel megpróbálja a hiányzó adatokat pótolni, a második az így kiegészített adatok likehihood függvényét maximalizálja. Van olyan eset, amikor az adatok kiegészítése csupán technikai, így könnyebben végrehajtható a maximum likelihood becslés. Az így nyert maximum általános feltételek mellett azonban a hiányos adatok likelihood függvényének is maximuma. Az eljárás során tehát az alábbi lépések felváltva ismétlődnek:

I. E-lépés (expectation): az iteráció előző lépésében a paraméterre adott becslése alapján feltételes várható értékkel rekonstruáljuk az adatokat

II. M-lépés (maximization): az adatok maximum likelihood függvényének maximumhelyének meghatározása a paraméter függvényében



3.1. ábra. Az EM algoritmus lépéseinek sematikus ábrája. A kezdeti paraméterérték megválasztása után az E- és M-lépések addig ismétlődnek, míg két egymás utáni iterációs lépés után a becslések különbsége megfelelően kicsi nem lesz.

Az EM algoritmus a tudomány különböző ágában nagyon népszerű és gyakran használt módszer. Széles körben alkalmazott területek a genetika, az ökonometria, valamint klinikai és szociológiai tanulmányokban is felhasználják. Az EM algoritmus keverék eloszlások paramétereinek becslésére is jól használható módszer. Az algoritmust képrekonstrukciós tomografikai eljárások során gyakran használják paraméterek maximum likelihood becslésére. Elterjedt alkalmazási terület például a beszédfelismeréshez alkalmazott rejtett Markov modellek is.

3.1. Konkrét példa

Az algoritmus szemléltetésére először nézzünk egy egyszerű példát, amely a [12] könyvből származik. Tegyük fel, hogy egy képen kétféle mintázat - világos és sötét - figyelhető meg. A sötét minták alakjuk szerint további két csoportba oszthatók:

kör alakú vagy szögletes. Szeretnénk megállapítani egy sötét alakzat előfordulási valószínűségét. Tudjuk még, hogy az alakzatok eloszlása trinomiális eloszlású. Jelölje X_1 a sötét kör alakú, X_2 a sötét szögletes és X_3 a világos minták darabszámát. Legyen $\mathbf{x} = (x_1, x_2, x_3)$ egy véletlen minta. A trinomiális eloszlás

$$\mathrm{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{n!}{x_1!\, x_2!\, x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3}$$

képlettel adható meg, ahol $n = x_1 + x_2 + x_3$ és $p_1 + p_2 + p_3 = 1$. Az egyszerűség kedvéért ennél többet is feltételezünk, az eloszlásról azt tudjuk, hogy

$$\mathrm{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{n!}{x_1!\, x_2!\, x_3!} \left(\frac{1}{4}\right)^{x_1} \left(\frac{1}{4} + \frac{p}{4}\right)^{x_2} \left(\frac{1}{2} - \frac{p}{4}\right)^{x_3}$$

alakú, ahol most a három ismeretlen paraméter (p_1, p_2 és p_3) helyett a p ismeretlen paraméter van. Felírva a log-likelihood függvényt, majd megoldva a $\frac{\mathrm{d}}{\mathrm{d}p}\left(\log f(\mathbf{x}|\mathbf{p})\right) = 0$ egyenletet, p paraméterre a következő becslés adódik:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}p}\left(\log\left(\frac{n!}{x_1!\, x_2!\, x_3!}\right) + x_1 \log\left(\frac{1}{4}\right) + x_2\left(\frac{1}{4} + \frac{p}{4}\right) + x_3\left(\frac{1}{2} - \frac{p}{4}\right)\right) &= 0 \\
\frac{x_2}{\frac{1}{4} + \frac{p}{4}} \cdot \frac{1}{4} + \frac{x_3}{\frac{1}{2} - \frac{p}{4}} \cdot \left(-\frac{1}{4}\right) &= 0 \\
\frac{x_2}{1+p} - \frac{x_3}{2-p} &= 0 \\
p = \frac{2x_2 - x_3}{x_2 + x_3}. \qquad (3.1)
\end{aligned}$$

Tegyük fel, hogy azt meg tudjuk különböztetni, hogy milyen színűek az alakzatok, viszont ha fekete alakzatot látunk annak az alakját (kör alakú vagy szögletes) nem tudjuk eldönteni. Jelölje y_1 és y_2 a megfigyelt fekete és fehér alakzatok számát, a megfelelő véletlen változókat Y_1 és Y_2. Így felírható az $y_1 = x_1 + x_2$ és $y_2 = x_3$ összefüggés. Cél az, hogy a megfigyelt y_1 és y_2 értékek mellett becslést adjunk a p paraméterre. A direkt megközelítést most nem tudjuk alkalmazni, mert x_2 értéket nem tudtuk megfigyelni, csak az $y_1 = x_1 + x_2$ érték áll a rendelkezésünkre.

Szükségünk lesz Y_1, Y_2 eloszlására, először ezt számítjuk ki.

$$\begin{aligned}
g(y_1, y_2|p) = \mathrm{P}(Y_1 = y_1, Y_2 = y_2) = \mathrm{P}(X_1 + X_2 = y_1, X_3 = y_2) = \\
= \sum_{i=0}^{y_1} \mathrm{P}(X_1 = i, X_2 = y_1 - i, X_3 = y_2) = \\
= \frac{(y_1 + y_2)!}{y_1!\, y_2!} p_3^{y_2} \sum_{i=0}^{y_1} \frac{y_1!}{i!\,(y_1 - i)!} p_1^i p_2^{y_1 - i} = \\
= \frac{(y_1 + y_2)!}{y_1!\, y_2!} (p_1 + p_2)^{y_1} p_3^{y_2} = \frac{(y_1 + y_2)!}{y_1!\, y_2!} \left(\frac{1}{2} + \frac{p}{4}\right)^{y_1} \left(\frac{1}{2} - \frac{p}{4}\right)^{y_2}
\end{aligned}$$

Látható, hogy az eloszlás binomiális eloszlás. Ebben a p paraméter maximum likelihood becslése a

$$p^* = \arg\max_p g(y_1, y_2|p)$$

megoldása, ami ebben az esetben könnyen számítható. Azokban az esetekben, amikor a direkt megközelítés nehezen számolható, lehet az EM algoritmust használni. Egy iterációs lépés lényege, hogy mivel x_1 és x_2 értéke nem ismert, először ezek feltételes várható értékét számítjuk ki. Majd ezeket használjuk p paraméter becslésére. Számítsuk ki X_1 és X_2 feltételes várható értékeket, ehhez szükség lesz $\mathrm{P}(X_1|Y_1)$ eloszlásra.

$$\begin{aligned}\mathrm{P}(X_1 = x_1|Y_1 = y_1, Y_2 = y_2) &= \frac{\mathrm{P}(X_1 = x_1, X_2 = y_1 - x_1, X_3 = y_2)}{\mathrm{P}(X_1 + X_2 = y_1, X_3 = y_2)} = \\ &= \left(\frac{(y_1+y_2)!}{x_1!\,(y_1-x_1)!\,x_3!} p_1^{x_1} p_2^{y_1-x_1} p_3^{y_2}\right) / \left(\frac{(y_1+y_2)!}{y_1!\,y_2!}(p_1+p_2)^{y_1} p_3^{y_2}\right) = \\ &= \frac{y_1!}{x_1!\,(y_1-x_1)!} p_1^{x_1} p_2^{y_1-x_1} \cdot \frac{1}{(p_1+p_2)^{y_1}}\end{aligned}$$

A feltételes várható érték

$$\begin{aligned}\mathrm{E}(X_1|X_1 + X_2 = y_1, X_3 = y_2) &= \sum_{i=0}^{y_1} i \cdot \frac{y_1!}{i!\,(y_1-i)!} p_1^i p_2^{y_1-i} \cdot \frac{1}{(p_1+p_2)^{y_1}} = \\ &= y_1 \frac{p_1}{p_1+p_2} = y_1 \frac{\frac{1}{4}}{\frac{1}{2}+\frac{p}{4}} \qquad (3.2)\end{aligned}$$

$$\mathrm{E}(X_2|X_1 + X_2 = y_1, X_3 = y_2) = y_1 \frac{p_2}{p_1+p_2} = y_1 \frac{\frac{1}{4}+\frac{p}{4}}{\frac{1}{2}+\frac{p}{4}} \qquad (3.3)$$

Nézzük meg, hogyan működik az EM algoritmus a konkrét példában. Tegyük fel, hogy már eljutottunk k iterációs lépésig, és ismert $x_1^{(k)}$, $x_2^{(k)}$, valamint $p^{(k)}$.

I. E-lépés: Kiszámítjuk az x feltételes várható értékét az y megfigyelt adat és a p paraméter aktuális becsült értéke mellett. Jelen esetben ez (3.2) szerint

$$x_1^{(k+1)} = \mathrm{E}(x_1|y_1, y_2, p^{(k)}) = y_1 \frac{\frac{1}{4}}{\frac{1}{2}+\frac{p^{(k)}}{4}},$$

hasonlóan megkapjuk (3.3) alapján, hogy

$$x_2^{(k+1)} = \mathrm{E}(x_2|y_1, y_2, p^{(k)}) = y_1 \frac{\frac{1}{4}+\frac{p^{(k)}}{4}}{\frac{1}{2}+\frac{p^{(k)}}{4}}.$$

Ebben a példában x_3 becslésére nincs szükség, mert y_2 megfigyelt érték x_3-mal egyenlő.

II. M-lépés: Az E-lépésben kapott $x_1^{(k+1)}$ és $x_2^{(k+1)}$ értékeket behelyettesítve (3.1) képletbe $p^{(k+1)}$-re kapjuk, hogy

$$p^{(k+1)} = \frac{2x_2^{(k+1)} - x_3}{x_2^{(k+1)} + x_3}$$

A példa szemléltetése a `trinom` függvény segítségével. A függvény megtalálható a Függelékben. Tegyük fel, hogy 1000 mintaelemből 672 fekete és 328 fehér. (Valójában a fekete minták közül $x_1 = 250$ és $x_2 = 422$, ezt előre nem tudjuk.) Kezdőértéknek $p = 0.5$ választva az algoritmus iterációs lépéseit követhetjük nyomon a 3.1 táblázatban. Összesen 10 iterációs lépést végezve p paraméterre 0.688 értéket kapunk eredményül.

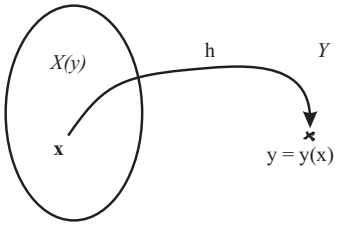
k	$x_1^{(k)}$	$x_2^{(k)}$	$p^{(k)}$
1	268.8000	403.2000	0.6542670
2	253.1772	418.8228	0.6824183
3	250.5202	421.4798	0.6870893
4	250.0847	421.9153	0.6878518
5	250.0138	421.9862	0.6879759
6	250.0022	421.9978	0.6879961
7	250.0004	421.9996	0.6879994
8	250.0001	421.9999	0.6879999
9	250.0000	422.0000	0.6880000
10	250.0000	422.0000	0.6880000

3.1. táblázat. Az EM algoritmus alkalmazása trinomiális eloszlású mintára.

3.2. Az EM algoritmus néhány tulajdonsága

Ebben a részben [10] és [11] könyveket követem. Jelölje X és Y a két mintateret. Legyen $\mathbf{y} \in \mathbb{R}^m$ egy megfigyelés Y-ból, erre hiányos adatként tekintünk. Az X-ből származó teljes, de nem megfigyelhető adatok vektorát $\mathbf{x} \in \mathbb{R}^n$ jelöli, ahol $m < n$. Legyen $h : X \to Y$ egy leképezés, melyre $h(\mathbf{x}) = \mathbf{y}$. Vagyis h a teljes és a megfigyelt adatok közötti kapcsolatot írja le. Jelölje $f_X(\mathbf{x}|\theta) = f(\mathbf{x}|\theta)$ a teljes adatok eloszlását, ahol $\theta \in \Theta \subset \mathbb{R}^r$ paramétervektor. Ekkor a hiányos adatok eloszlása

$$g(\mathbf{y}|\theta) = \int_{X(\mathbf{y})} f(\mathbf{x}|\theta)\mathrm{d}\mathbf{x},$$



3.2. ábra. Az X és Y közötti összefüggés h függvény segítségével. X a nem megfigyelhető adtok halmaza, Y a megfigyelhető adtok halmaza. $\mathbf{y}$ jelöli a megfigyelt mintát.

ahol $X(\mathbf{y}) = \{\mathbf{x} \in X : h(\mathbf{x}) = \mathbf{y}\} \subset X$, vagyis az $\mathbf{y}$ h általi inverzképe.

Célunk, hogy megtaláljuk azt a θ paramétervektort, ami maximalizálja a $\log f(\mathbf{x}|\theta)$ függvényt. Mivel azonban $\mathbf{x}$-et nem ismerjük, helyette a $\log f(\mathbf{x}|\theta)$ feltételes várható értékét számítjuk ki az ismert $\mathbf{y}$ és θ aktuális becslése mellett. Egy iterációs lépésben két lépést valósítunk meg. Tegyük fel, hogy már elvégeztünk k iterációs lépést, és θ-ra a $\theta^{(k)}$ becslést kaptuk. Ezután:

E-lépés: Kiszámítjuk

$$Q(\theta, \theta^{(k)}) = \mathrm{E}(\log f(\mathbf{x}|\theta)|\mathbf{y}, \theta^{(\mathbf{k})})$$

értéket. Ennél a lépésnél $\theta^{(k)}$ ismert érték.

M-lépés: Legyen $\theta^{(k+1)}$ a $Q(\theta, \theta^{(k)})$ függvény θ szerinti maximumhelye:

$$\theta^{(k+1)} = \arg\max_{\theta} Q(\theta, \theta^{(k)})$$

Az eljárás során kiindulunk egy $\theta^{(0)}$ kezdőértékből, majd az E- és M-lépéseket addig ismételjük, amíg olyan k-hoz érünk, amire $\left\|\theta^{(k)} - \theta^{(k-1)}\right\| < \varepsilon$ már teljesül, valamilyen előre adott ε-ra.

Egy fontos kérdés, az algoritmus konvergenciája. Az EM algoritmus esetében igaz lesz, hogy minden egyes iterációban a becsült paraméter növeli a likelihood függvény értékét, vagy legalábbis nem csökkenti, amíg (lokális) maximumot el nem ér. Legyen

$$l(\theta) = \log g(\mathbf{y}|\theta)$$

a megfigyelt adatok log-likelihood függvénye.

3.1. Tétel. *Az EM algoritmus iterációs lépései során kapott $\theta^{(k)}$ és $\theta^{(k+1)}$ becslésekre teljesül, hogy $l(\theta^{(k+1)}) \geq l(\theta^{(k)})$.*

Bizonyítás. Vezessük be $\mathbf{x}$ feltételes eloszlásra $(\mathbf{y}, \theta)$ feltétel mellett az

$$k(\mathbf{x}|\mathbf{y}, \theta) = \frac{f(\mathbf{x}|\theta)}{g(\mathbf{y}|\theta)},$$

jelölést, és legyen

$$H(\theta, \theta') = \mathrm{E}(\log k(\mathbf{x}|\mathbf{y}, \theta)|\mathbf{y}, \theta').$$

Ekkor a log-likelihood függvény felírható

$$l(\theta) = \log g(\mathbf{y}|\theta) = \log f(\mathbf{x}|\theta) - \log k(\mathbf{x}|\mathbf{y}, \theta)$$

alakba. Vegyük mindkét oldal feltételes várható értékét $(\mathbf{y}, \theta^{(k)})$ feltétel mellett, ekkor az

$$\begin{aligned} l(\theta) &= \mathrm{E}(\log f(\mathbf{x}|\theta))|\mathbf{y}, \theta^{(k)}) - \mathrm{E}(\log k(\mathbf{x}|\mathbf{y}, \theta))|\mathbf{y}, \theta^{(k)}) = \\ &= Q(\theta, \theta^{(k)}) - H(\theta, \theta^{(k)}) \end{aligned} \tag{3.4}$$

összefüggést kapjuk. Azt szeretnénk belátni, hogy $l(\theta^{(k+1)}) \geq l(\theta^{(k)})$, vagyis

$$l(\theta^{(k+1)}) - l(\theta^{(k)}) \geq 0$$

A fenti (3.4) összefüggés szerint

$$\begin{aligned} l(\theta^{(k+1)}) - l(\theta^{(k)}) &= \\ &= \left(Q(\theta^{(k+1)}, \theta^{(k)}) - H(\theta^{(k+1)}, \theta^{(k)})\right) - \left(Q(\theta^{(k)}, \theta^{(k)}) - H(\theta^{(k)}, \theta^{(k)})\right) = \\ &= \left(Q(\theta^{(k+1)}, \theta^{(k)}) - Q(\theta^{(k)}, \theta^{(k)})\right) - \left(H(\theta^{(k+1)}, \theta^{(k)}) - H(\theta^{(k)}, \theta^{(k)})\right) \end{aligned}$$

A $\theta^{(k+1)}$ M-lépésbeli választása miatt

$$Q(\theta^{(k+1)}, \theta^{(k)}) \geq Q(\theta, \theta^{(k)})$$

teljesül minden $\theta \in \Theta$. Kell még, hogy minden $\theta \in \Theta$ fennáll, hogy $H(\theta, \theta^{(k)}) - H(\theta^{(k)}, \theta^{(k)})$ kifejezés nem pozitív.

$$\begin{aligned} H(\theta, \theta^{(k)}) - H(\theta^{(k)}, \theta^{(k)}) &= \mathrm{E}\left(\log k(\mathbf{x}|\mathbf{y}, \theta)|\mathbf{y}, \theta^{(k)}\right) - \mathrm{E}\left(\log k(\mathbf{x}|\mathbf{y}, \theta^{(k)})|\mathbf{y}, \theta^{(k)}\right) = \\ &= \mathrm{E}\left(\log \frac{k(\mathbf{x}|\mathbf{y}, \theta)}{k(\mathbf{x}|\mathbf{y}, \theta^{(k)})}|\mathbf{y}, \theta^{(k)}\right) \leq \\ &\leq \log \mathrm{E}\left(\frac{k(\mathbf{x}|\mathbf{y}, \theta)}{k(\mathbf{x}|\mathbf{y}, \theta^{(k)})}|\mathbf{y}, \theta^{(k)}\right) = \\ &= \log \int_X k(\mathbf{x}|\mathbf{y}, \theta) \mathrm{d}\mathbf{x} = \\ &= 0 \end{aligned}$$

□

A bizonyítás során felhasználtuk a Jensen-egyenlőtlenséget. A tétel ellenére lehetséges, hogy az EM algoritmus nem a globális maximumát találja meg a likelihood függvénynek. Mivel több lokális maximum is lehet, ezért lehetséges, hogy az algoritmus által talált maximum nem globális maximum. Több maximum esetén a kezdeti θ paramétertől függ, hogy hova konvergál az algoritmus.

3.3. Keverék eloszlások paraméterbecslése

Az EM algoritmus alkalmazásának egyik népszerű területe a keverék eloszlások paramétereinek maximum likelihood becslése. Tegyük fel, hogy az eloszlás a következő alakban áll elő:

$$g(y_j|\Psi) = \sum_{i=1}^{g} \pi_i g_i(y_j|\theta_i),$$

ahol a $\Psi = (\pi_1, \pi_2, \ldots, \pi_k, \theta_1, \theta_2, \ldots \theta_g)$ vektor a súlyokat és az eloszlások paramétereit tartalmazza, $\sum_{i=1}^{g} \pi_i = 1$ és g_i az i. eloszlás sűrűségfüggvénye θ_i paraméterrel, esetleg paramétervektorral. Vagyis az eloszlás g darab eloszlás konvex kombinációja, ahol g_i súlya π_i.

A megfigyelt adatok log-likelihood függvénye ekkor felírható

$$\log L(\mathbf{y}|\Psi) = \log \prod_{j=1}^{n} g(y_j|\Psi) = \sum_{j=1}^{n} \log g(y_j|\Psi) = \sum_{j=1}^{n} \log \left(\sum_{i=1}^{g} \pi_i g_i(y_j|\theta_i) \right) \tag{3.5}$$

alakba. A Ψ paraméter becslése a

$$\frac{\partial \log L(y|\Psi)}{\partial \Psi} = 0$$

likelihood egyenlethez vezet, ami bizonyos esetekben nehezen számolható.

Vezessük be a z_{ij} változókat, amelyek azt az információt hordozzák, hogy melyik mintaelem melyik eloszlásból származik. Tehát legyen

$$z_{ij} = \begin{cases} 1 & \text{ha } y_j \text{ a } g_i \text{ eloszlásból való,} \\ 0 & \text{egyébként.} \end{cases}$$

A $Z = [z_{ij}]$ mátrix elemei nem megfigyelhetőek. A Z mátrix elemei arról informálnak, hogy melyik eloszlásból való az i-edik megfigyelés, ezért Z minden oszlopának pontosan egy eleme lesz 1, a többi elem abban az oszlopban 0. Legyen $\mathbf{x} = (\mathbf{y}^T, \mathbf{z}^T)^T$ teljes adat, $\mathbf{y}$ megfigyelésvektor és Z_j indikátorváltozó, amely a fent leírt információkat tartalmazza. Ekkor az $\mathbf{x}$ likelihood függvénye a következő alakban írható fel.

$$L(\mathbf{x}|\Psi) = \prod_{j=1}^{n} \prod_{i=1}^{g} (\pi_i g_i(y_j|\theta_i))^{z_{ij}}$$

Innen a log-likelihood függvény

$$\log L(x|\Psi) = \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}(\log\pi_i + \log g_i(y_j|\theta_i)) \tag{3.6}$$

Szeretnénk a π_i együtthatókat és a θ_i paramétereket becsülni. Ehhez használjuk az EM algoritmust. Mivel a z_{ij} értékek ismeretlenek, ezért az algoritmus első lépésében, az E-lépésben, a z_j feltételes várható értékét számítjuk ki a teljes adat log-likelihood függvénye mellett. Ehhez felhasználjuk a megfigyelt $\mathbf{y}$ vektort és a Ψ aktuális értékét. Jelölje $\Psi^{(0)}$ a Ψ kezdeti értékét. Ezután az EM algoritmus első iterációs lépésének E-lépésében szükséges a teljes adat log-likelihood függvényének számítása az adott $\mathbf{y}$ és $\Psi^{(0)}$ mellett. Ezt írhatjuk

$$Q(\Psi, \Psi^{(0)}) = \mathrm{E}(\log L(x|\Psi)|\mathbf{y}, \Psi^{(0)})$$

alakba. Ugyanígy az algoritmus $(k+1)$-edik iterációjában az E-lépéshez szükség lesz $Q(\Psi, \Psi^{(k)})$-ra, amiben $\Psi^{(k)}$ a Ψ értéke az algoritmus k-adik iterációs lépése után. Mivel a teljes adat log-likelihood függvénye lineáris a nem megfigyelhető z_{ij} változókban, ezért a $(k+1)$-edik iterációs lépésben a nekik megfelelő Z_{ij} véletlen változók feltételes várható értékét kell kiszámolni a megfigyelt $\mathbf{y}$ mellett.

$$z_{ij}^{(k+1)} = \mathrm{E}(Z_{ij}|\mathbf{y}, \Psi^{(k)}) = \mathrm{P}(Z_{ij}|\mathbf{y}, \Psi^{(k)}) = \frac{\pi_i^{(k)} g_i(y_j|\theta_i^{(k)})}{\sum_{h=1}^{g} \pi_h^{(k)} g_h(y_j|\theta_h^{(k)})} \tag{3.7}$$

Ez a mennyiség annak a valószínűségét adja meg, hogy a minta j-edik eleme, az y_j megfigyelés a keverék eloszlás i-edik komponenséből való. Ezt felhasználva a log-likelihood feltételes várható értékére

$$\begin{aligned} Q(\Psi, \Psi^{(k)}) =& \mathrm{E}(\log L(x|\Psi)|\mathbf{y}, \Psi^{(k)}) = \\ &= \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}^{(k+1)}(\log\pi_i + \log g_i(y_j|\theta_i)) \end{aligned}$$

összefüggés adódik.

Az algoritmus ugyanezen iterációs lépésénél az M-lépés a $Q(\Psi, \Psi^{(k)})$ kifejezés globális maximalizálását végzi el. A $\pi_i^{(k+1)}$ súlyok és a $\theta_i^{(k+1)}$ eloszlás paraméterek kiszámítása egymástól függetlenül történik. Ha a z_{ij} értékek megfigyelhetőek lennének, akkor a π_i maximum likelihood becslése

$$\pi_i^* = \frac{1}{n}\sum_{j=1}^{n} z_{ij}$$

lenne. Mivel ezeket nem ismerjük, ezért helyettesítjük (3.7) összefüggésben kapott várható értékkel. Így π_i-re a következő

$$\pi_i^{(k+1)} = \frac{1}{n}\sum_j^n z_{ij}^{(k+1)} = \frac{1}{n}\sum_{j=1}^n \frac{\pi_i^{(k)} g_i(y_j|\theta_i^{(k)})}{\sum_{h=1}^g \pi_h^{(k)} g_h(y_j|\theta_h^{(k)})}$$

becslés adódik. A θ_i becsléséhez a likelihood függvény maximumát kell megtalálni.

$$\theta_i^{(k+1)} = \arg\max_{\theta_i} \sum_{j=1}^n z_{ij}^{(k+1)} \log g_i(y_j|\theta_i)$$

3.4. EM algoritmus normális eloszlások keverékének paraméterbecslésére

A következőkben nézzünk egy speciális esetet az EM algoritmus alkalmazására. Tegyük fel, hogy a mintánk g darab normális eloszlás keverékéből származik, vagyis az Y eloszlás az $Y_1, Y_2, \dots Y_g$ normális eloszlások konvex kombinációja.

$$\begin{aligned} Y_1 &\sim N(\mu_1, \sigma_1^2) \\ Y_2 &\sim N(\mu_2, \sigma_2^2) \\ &\vdots \\ Y_g &\sim N(\mu_g, \sigma_g^2) \\ Y &= \sum_{i=1}^g \pi_i Y_i, \end{aligned}$$

ahol $\sum_{i=1}^g \pi_i = 1$. Ψ tartalmazza az összes ismeretlen paramétert, így

$$\Psi = (\pi_1, \dots, \pi_g, \theta_1, \dots, \theta_g),$$

ahol $\theta_i = (\mu_i, \sigma_i)$ és Y sűrűségfüggvénye

$$g_Y(y|\Psi) = \sum_{i=1}^g \pi_i g_{Y_i}(y|\theta_i)$$

ahol

$$g_{Y_i}(y|\theta_i) = \left(2\pi\sigma_i^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y-\mu_i)^2/\sigma_i^2\right\}.$$

Az n elemű minta minta elemei $y_1, y_2, \dots, y_n$ függetlenek, akkor a log-likelihood függvény a (3.5) egyenlet szerint

$$\begin{aligned} \log L(y|\Psi) &= \sum_{j=1}^n \log\left(\sum_{i=1}^g \pi_i g_{Y_i}(y_j|\theta_i)\right) = \\ &= \sum_{j=1}^n \log\left(\sum_{i=1}^g \pi_i \left(2\pi\sigma_i^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y_j-\mu_i)^2/\sigma_i^2\right\}\right) \end{aligned}$$

lesz. Ennek maximalizálása a logaritmusfüggvény miatt nehéz, ezért bevezetjük a z_{ij} változókat. z_{ij} értéke 1, ha y_j a g_{Y_i} eloszlásból való, egyébként 0. Ekkor a teljes adat $\mathbf{x} = (\mathbf{y}^T, \mathbf{z}^T)^T$ log-likelihood függvény a (3.6) egyenlethez hasonlóan

$$\begin{aligned} \log L(x) &= \sum_{j=1}^{n} \sum_{i=1}^{g} z_{ij} \log g_{Y_i}(y_j|\theta_i) = \\ &= -\frac{1}{2} n \log(2\pi) - \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{g} z_{ij} \left(\log\sigma_i^2 + (y_j - \mu_i)^2/\sigma_i^2 \right) \end{aligned}$$

Az algoritmus $(k+1)$-edik iterációjának E-lépésében mivel a z_{ij} értékek nem ismertek, ezért elsőként azokat a várható értékükkel helyettesítjük. Ezt a (3.7) egyenletben szereplő képlet szerint tesszük.

$$z_{ij}^{(k+1)} = \frac{\pi_i^{(k)} g_i(y_j|\theta_i^{(k)})}{\sum_{h=1}^{g} \pi_h^{(k)} g_h(y_j|\theta_h^{(k)})}$$

Ezután az M-lépésben maximalizáljuk a log-likelihood függvényt. Normális eloszlás esetén a maximum likelihood becslés μ-re a mintaelemek átlaga, σ-ra a tapasztalati szórás. Most is ezt használjuk, csak a mintaelemeket súlyozva számítjuk be a paraméterek becslésébe.

$$\begin{aligned} \mu_i^{(k+1)} &= \frac{\sum_{j=1}^{n} z_{ij}^{(k)} y_j}{\sum_{j=1}^{n} z_{ij}^{(k)}} \\ \sigma_i^{(k+1)} &= \sqrt{\frac{\sum_{j=1}^{n} z_{ij}^{(k)} (y_j - \mu_i^{(k)})^2}{\sum_{j=1} z_{ii}^{(k)}}} \end{aligned}$$

és

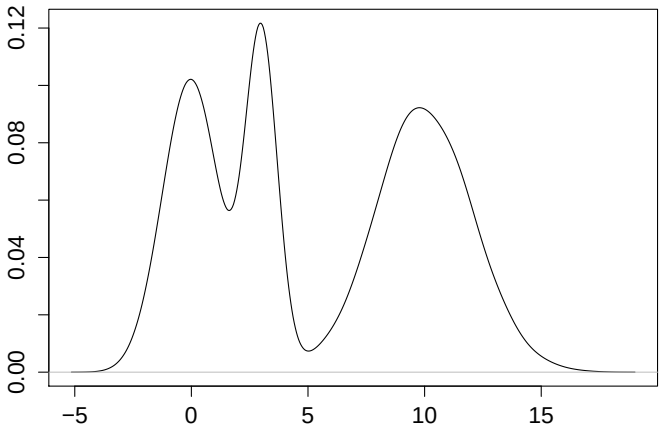
$$\pi_i^{(k+1)} = \frac{1}{n} \sum_{j=1}^{n} z_{ij}^{(k)} \qquad (i = 1, \ldots, g)$$

Most következik egy numerikus példa az algoritmus szemléltetésére. A `mixgen`[1] függvénnyel 10000 elemű mintát generáltam három normális eloszlás keverékéből. Ezek a következő paraméterekkel rendelkeznek:

$N(0,1)$	$N(10,2)$	$N(3,0.1)$
$\mu_1 = 0$	$\mu_2 = 10$	$\mu_3 = 3$
$\sigma_1 = 1$	$\sigma_2 = 2$	$\sigma_3 = 0.1$
$\pi_1 = 0.3$	$\pi_2 = 0.5$	$\pi_3 = 0.2$

[1] ld. Függelék

A 3.3 ábrán látható a minta sűrűségfüggvénye. Ebből szeretnénk visszabecsülni a paramétereket. Ehhez az R statisztikai programban általam megírt `emtnorm`[2] függvényt használom. Azt feltételezzük, hogy tudjuk, hogy három normális eloszlás keverékéből kaptuk meg a mintát. Tehát szeretnénk meghatározni μ, σ és π paramétervektorok értékét.



3.3. ábra. A generált minta sűrűségfüggvénye.

A kezdőérték megválasztásában semmilyen szakirodalombeli állításra nem tudtam támaszkodni, a súlyokra az egyenlő arány, a várható értékekre a mintaátlag kézenfekvő megoldásnak tűnt. (A szórást viszont már nem lehetett ugyanolyan értékekre állítani, mivel az algoritmus abban az esetben nem csinál semmit.) Így a következő értékek mellett döntöttem:

$$\pi = (1/3, 1/3, 1/3)$$
$$\mu = (\bar{y}, \bar{y}, \bar{y})$$
$$\sigma = (1, 2, 3)$$

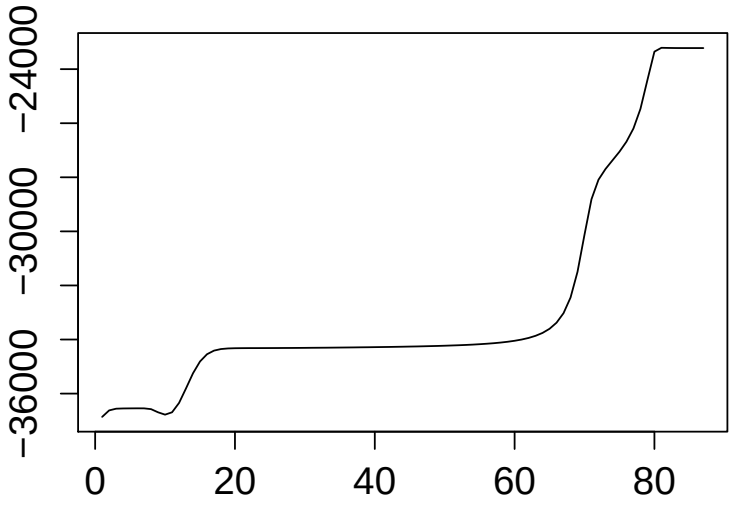
Az algoritmus legfeljebb 100 iterációs lépést végez. Előbb leáll, ha az egymást követő iterációs lépésekben kapott becslések különbsége minden paramétervektor esetén kisebb, mint 1e-04 és a log-likelihood változása kisebb, mint 0.5.

$$\begin{array}{lll} \mu_1 = -0.0329 & \mu_2 = 3.0011 & \mu_3 = 9.9890 \\ \sigma_1 = 0.9925 & \sigma_2 = 0.0998 & \sigma_3 = 1.9874 \\ \pi_1 = 0.3106 & \pi_2 = 0.2022 & \pi_3 = 0.4872 \end{array}$$

A log-likelihood függvény maximuma -23219.59, az iteráció során bekövetkező változás a 3.4. ábrán látható. Az algoritmus összesen 87 iterációs lépést végzett.

[2]ld. Függelék

A valós értékek ismeretében össze tudjuk hasonlítani a kapott eredményt. Megállapítható, hogy minden paraméter becslése legalább egy tizedesjegyig pontos. Pontosabb eredményt kaphatunk, ha az algoritmus leállási feltételéhez szigorúbb feltételeket kötünk, vagy más kezdeti paraméter értékeket állítunk be.



3.4. ábra. Az EM algoritmus iterációs lépéseinek során a log-likelihood függvény maximumának változása.

4. fejezet

Poisson regressziós modellek

A kétváltozós Poisson eloszlás általánosabb megközelítésében a λ paraméterek minden megfigyelésre különbözhetnek. Pontosabban, nem csak a λ_1, λ_2 és λ_3 értékek térnek el egymástól, hanem egy-egy paraméter a megfigyelt egyén sajátos tulajdonságaitól függ. A gyakorlati alkalmazásban a paraméterek becsléséhez felhasználnak különböző magyarázó változókat, ez pontosabb és összetettebb modellt eredményez.

Ebben a fejezetben áttekintem, hogy hogyan működik az EM algoritmus a Poisson regressziós modell paramétereinek becslésére. Ehhez a [8] cikkben található eredményeket használom fel. Majd valós adatok felhasználásával kétváltozós Poisson és kétváltozós diagonálisan módosított Poisson eloszlást illesztettem az R programban.

4.1. Kétváltozós Poisson regressziós modell

Tegyük fel, hogy az i. kötvénytulajdonos első típusú kár kárszámát X_i, második típusú kár kárszámát Y_i jelöli. A kétváltozós Poisson regressziós modell a következőképpen definiálható:

$$\begin{aligned}(X_i, Y_i) &\sim BP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}),\\ \log(\lambda_{1i}) &= w_{1i}^T \beta_1,\\ \log(\lambda_{2i}) &= w_{2i}^T \beta_2,\\ \log(\lambda_{3i}) &= w_{3i}^T \beta_3,\end{aligned} \tag{4.1}$$

ahol $i = 1, \ldots, n$ a megfigyelt kötvénytulajdonosok száma, $w_{\kappa i}$ a magyarázó változók vektora, β_κ a megfelelő magyarázó változók együtthatóvektora ($\kappa = 1, 2, 3$).

A modellezéshez használt $w_{\kappa i}$ magyarázó vektorok nem feltétlenül ugyanazokat a tulajdonságokat tartalmazzák mindhárom λ esetén. Általában azt fel szokták tenni, hogy a λ_3 paraméter állandó, így a modell könnyebben magyarázható.

4.2. EM algoritmus kétváltozós Poisson modellre

Nézzük hogyan működik az EM algoritmus a kétváltozós Poisson regressziós modellre. Jelölje az i. szerződő nem megfigyelhető változóit X_{1i}, X_{2i} és X_{3i}, míg a megfigyelhető adatok legyenek

$$X_i = X_{1i} + X_{3i}$$
$$Y_i = X_{2i} + X_{3i}$$

Ezek a jelölések kicsit eltérnek az előző fejezetben használttól, mivel most minden szerződőhöz két megfigyelésünk van: a kétféle típusú kárszám X_i és Y_i. Ha a nem megfigyelhető változókat ismernénk, akkor egyszerűen tudnánk illeszteni a kétváltozós Poisson modellt az X_1, X_2, X_3 változókra. Mivel ezeket nem ismerjük, így elsőként az algoritmus E-lépésében az X_{1i}, X_{2i} és X_{3i} változókat kell megbecsülni a feltételes várható értékükkel, majd ezeket felhasználva illeszteni a Poisson modellt.

Legyen a kezdeti paramétervektor $\omega = (\beta_1^{(0)}, \beta_2^{(0)}, \beta_3^{(0)})$, a magyarázó változók együtthatóinak vektora. Az adatok likelihood függvénye a következőképpen írható fel

$$L(\omega) = \left(\prod_{i=1}^{n} \frac{e^{-\lambda_{1i}} \lambda_{1i}^{x_{1i}}}{x_{1i}!}\right) \left(\prod_{i=1}^{n} \frac{e^{-\lambda_{2i}} \lambda_{1i}^{x_{2i}}}{x_{2i}!}\right) \left(\prod_{i=1}^{n} \frac{e^{-\lambda_{3i}} \lambda_{1i}^{x_{3i}}}{x_{3i}!}\right).$$

Így a minta log-likelihood függvénye

$$l(\omega) = -\sum_{i=1}^{n}\sum_{\kappa=1}^{3} \lambda_{\kappa i} + \sum_{i=1}^{n}\sum_{\kappa=1}^{3} x_{\kappa i} \log(\lambda_{\kappa i}) - \sum_{i=1}^{n}\sum_{\kappa=1}^{3} \log(x_{\kappa i}!),$$

ahol a λ paraméterek (4.1) egyenlet szerint adottak.

Tegyük fel, hogy az iteráció során már k lépést végeztünk, rendelkezésünkre állnak $\omega^{(k)}$, $\lambda_{1i}^{(k)}$, $\lambda_{2i}^{(k)}$, $\lambda_{3i}^{(k)}$ értékek. Az algoritmus E-lépésében kiszámítjuk X_{3i} feltételes várható értékét minden $i = 1, 2, \ldots, n$-re. Jelölje ezt s_i.

$$s_i = \mathrm{E}(X_{3i} | X_i, Y_i, \omega^{(k)}) =$$
$$= \begin{cases} \lambda_{3i}^{(k)} \frac{f_{BP}\left(x_i - 1, y_i - 1 | \lambda_{1i}^{(k)}, \lambda_{2i}^{(k)}, \lambda_{3i}^{(k)}\right)}{f_{BP}\left(x_i, y_i | \lambda_{1i}^{(k)}, \lambda_{2i}^{(k)}, \lambda_{3i}^{(k)}\right)} & \text{ha } \min(x_i, y_i) > 0 \\ 0 & \text{ha } \min(x_i, y_i) = 0, \end{cases}$$

ahol f_{BP} jelöli a (2.1) képlettel megadott kétváltozós Poisson eloszlást.

Az M-lépésben legyenek

$$\begin{aligned}
\beta_1^{(k+1)} &= \hat{\beta}(\mathbf{x}-\mathbf{s}, W_1),\\
\beta_2^{(k+1)} &= \hat{\beta}(\mathbf{y}-\mathbf{s}, W_2),\\
\beta_3^{(k+1)} &= \hat{\beta}(\mathbf{s}, W_3),\\
\lambda_{1i}^{(k+1)} &= \exp\left(W_{1i}^T \hat{\beta}_1^{(k+1)}\right),\\
\lambda_{2i}^{(k+1)} &= \exp\left(W_{2i}^T \hat{\beta}_2^{(k+1)}\right),\\
\lambda_{3i}^{(k+1)} &= \exp\left(W_{3i}^T \hat{\beta}_3^{(k+1)}\right),
\end{aligned}$$

ahol

- $\mathbf{s}=(s_1, s_2, \ldots, s_n)^T$ $n\times 1$ vektor, melynek elemeit az E-lépésben számítottuk ki,
- $\hat{\beta}(\mathbf{x}, \mathbf{W})$ a Poisson modell maximum likelihood becslése $\mathbf{x}$ vektor maximumhelyen és adott $\mathbf{W}$ adatmátrixra.
- a $\mathbf{W}_\kappa$ mátrix egy $n \times p_\kappa$ méretű mátrix, és $\mathbf{W}_{\kappa i}^T$ ennek a mátrixnak az i-edik sora $(i = 1, \ldots, n)$. Ha feltesszük, hogy az egyes λ paramétereket ugyanabból az adatmátrixból számítjuk, jelölje ezt a mátrixot $\mathbf{W}$, akkor a β paramétervektorra a
$$\beta^{(k+1)} = \hat{\beta}(u, \mathbf{W})$$
becslést kapjuk, ahol $\mathbf{u}^T = \left(\mathbf{x}^T - \mathbf{s}^T, \mathbf{y}^T - \mathbf{s}^T, \mathbf{s}^T\right)$.

4.3. EM algoritmus diagonálisan módosított kétváltozós Poisson modellre

A diagonálisan módosított Poisson eloszlás két független változó keveréke, egy kétváltozós Poisson eloszlású és egy diagonális változóé, ezért szükség van látens változók bevezetésére. Az előző fejezetben láttuk az EM algoritmus működését keverék eloszlások paraméterbecslésére. Vezessük be a V_i $(i=1, \ldots, n)$ változókat, melyek 0 és 1 értéket vesznek fel aszerint, hogy az i. megfigyelés melyik eloszlásból való.

$$v_i = \begin{cases} 1 & \text{ha a megfigyelés a diagonális eloszlásból jön,}\\ 0 & \text{egyébként.}\end{cases}$$

A log-likelihood függvény a következő alakot ölti

$$l(\omega,p,\theta)=\sum_{i=1}^{n} v_i\left\{\log p+\log f_D(x_i;\theta)\right\}+$$
$$+\sum_{i=1}^{n}(1-v_i)\left\{\log(1-p)-\sum_{i=1}^{n}\sum_{\kappa=1}^{3}\lambda_{\kappa i}+\sum_{i=1}^{n}\sum_{\kappa=1}^{3}x_{\kappa i}\log(\lambda_{\kappa i})-\sum_{i=1}^{n}\sum_{\kappa=1}^{3}\log(x_{\kappa i}!)\right\}.$$

Az EM algoritmus E-lépésében először a V_i változók feltételes várható értékeket számoljuk ki, majd ezután az X_{3i} feltételes várható értékeket. Tegyük fel, hogy már rendelkezésünkre állnak a k-adik iterációs lépés becslései $\omega^{(k)}$ (β együtthatók értékei), $\lambda_{1i}^{(k)}$, $\lambda_{2i}^{(k)}$, $\lambda_{3i}^{(k)}$, $p^{(k)}$ és $\theta^{(k)}$ (diszkrét eloszlás paramétervektora) minden $i=1,\ldots,n$ esetén, ekkor V_i feltételes várható értéket a következő összefüggéssel kapjuk:

$$v_i=\mathrm{E}\left(V_i|X=x_i,Y=y_i,\omega^{(k)},p^{(k)},\theta^{(k)}\right)=$$
$$=\begin{cases}\frac{p^{(k)}f_D\left(x_i|\theta^{(k)}\right)}{p^{(k)}f_D\left(x_i|\theta^{(k)}\right)+(1-p^{(k)})f_{BP}\left(x_i,y_i|\lambda_{1i}^{(k)},\lambda_{2i}^{(k)},\lambda_{3i}^{(k)}\right)} & \text{ha } x_i=y_i,\\ 0 & \text{ha } x_i\neq y_i,\end{cases}$$

ahol $f_D(x|\theta)$ a diszkrét eloszlás súlyfüggvénye, amely a $\{0,1,2,\ldots\}$ halmazon definiált $D(x;\theta)$ „diagonális” változó súlyfüggvénye θ paramétervektorral, 2.fejezet szerinti jelöléseknek megfelelően. Ezután következik az X_{3i} feltételes várható értékek kiszámítása minden $i=1,\ldots,n$ értékre.

$$s_i=\mathrm{E}(X_{3i}|X_i,Y_i,\omega^{(k)})=$$
$$=\begin{cases}\lambda_{3i}^{(k)}\frac{f_{BP}\left(x_i-1,y_i-1|\lambda_{1i}^{(k)},\lambda_{2i}^{(k)},\lambda_{3i}^{(k)}\right)}{f_{BP}\left(x_i,y_i|\lambda_{1i}^{(k)},\lambda_{2i}^{(k)},\lambda_{3i}^{(k)}\right)} & \text{ha } \min(x_i,y_i)>0\\ 0 & \text{ha } \min(x_i,y_i)=0,\end{cases}$$

M-lépés:

$$p^{(k+1)}=\frac{1}{n}\sum_{i=1}^{n}v_i,$$
$$\beta_1^{(k+1)}=\hat{\beta}_{\tilde{v}}(\mathbf{x}-\mathbf{s},W_1),$$
$$\beta_2^{(k+1)}=\hat{\beta}_{\tilde{v}}(\mathbf{y}-\mathbf{s},W_2),$$
$$\beta_3^{(k+1)}=\hat{\beta}_{\tilde{v}}(\mathbf{s},W_3),$$

$$\begin{aligned}
\theta^{(k+1)} &= \hat{\theta}_{v,D}, \\
\lambda_{1i}^{(k+1)} &= \exp\left(W_{1i}^T \hat{\beta}_1^{(k+1)}\right), \\
\lambda_{2i}^{(k+1)} &= \exp\left(W_{2i}^T \hat{\beta}_2^{(k+1)}\right), \\
\lambda_{3i}^{(k+1)} &= \exp\left(W_{3i}^T \hat{\beta}_3^{(k+1)}\right),
\end{aligned}$$

ahol

– $\mathbf{x}$, $\mathbf{y}$, $\mathbf{s}$, $\mathbf{v}$, $\tilde{\mathbf{v}}$ $n \times 1$ méretű vektorok, rendre x_i, y_i, s_i, v_i, $\tilde{v}_i = 1 - v_i$ elemekkel $i = 1 \ldots, n$,

– $\hat{\beta}_{\tilde{v}}(\mathbf{y}, \mathbf{W})$ a Poisson modell súlyozott maximum likelihood becslése $\mathbf{y}$ vektor maximumhelyen adott $\mathbf{W}$ adatmátrixra és $\mathbf{v}$ súlyvektorra,

– $\hat{\theta}_{v,D}$ a $D(x;\theta)$ eloszlás θ paraméterének súlyozott maximum likelihood becslése v súlyvektor mellett,

– a W_κ $(\kappa = 1, 2, 3)$ adatmátrix az előző pontban definiált.

A diagonális módosító eloszlás választása alapján $\theta^{(k+1)}$ becslését meg lehet adni zárt alakban:

- Geometriai eloszlás: $f(x|\theta) = (1-\theta)^x \theta, 0 \leq \theta \leq 1, x = 0, 1, \ldots$

$$\theta^{(k+1)} = \frac{\sum_{i=1}^n v_i}{\sum_{i=1}^n v_i x_i + \sum_{i=1}^n v_i}$$

- Poisson eloszlás: $f(x|\theta) = e^{-\theta}\theta^x / x!\,, 0 \leq \theta, x = 0, 1, \ldots$

$$\theta^{(k+1)} = \frac{\sum_{i=1}^n v_i x_i}{\sum_{i=1}^n v_i}$$

- Diszkrét eloszlás (2.3) pontban definiáltak szerint:

$$\begin{aligned}
\theta_j^{(k+1)} &= \frac{\sum_{i=1}^n I(X_i = Y_i = j) v_i}{\sum_{i=1}^n v_i}, \\
\theta_0^{(k+1)} &= 1 - \sum_{j=1}^J \theta_j^{(k+1)},
\end{aligned}$$

  ahol $I(x)$ az indikátor függvény, melynek értéke 1, ha x igaz, egyébként 0.

4.4. Elemzés

A következő részben egy konkrét példán keresztül mutatom be, hogyan alkalmazható az EM algoritmus kétváltozós Poisson és diagonálisan módosított kétváltozós Poisson eloszlások paramétereinek becslésére. Valós adatok nyilvánosan nem elérhetőek, ezért Lluíz Bermúdez cikkében [1] található adatokat, valamint az internetről ingyenesen letölthető *bivpois* nevű csomagot használom az R programban.

A cikkbeli elemzés nem terjed ki az egész portfólióra, összesen 80994 kötvénytulajdonos tartozik a vizsgált mintába. Ők mind magán jellegű célokra használják gépkocsijukat, és legalább három éve ugyanahhoz a biztosítótársasághoz szerződtek. Az adatok egy spanyol biztosítótársaság kötvénytulajdonosaira vonatkoznak 1995-ből.

A két kárszám X és Y legyenek a kötelező gépjármű-felelősségbiztosításból származó és a nem kötelező gépjármű-felelősségbiztosításból eredő, de gépjármű biztosításba tartozó biztosítások kárszámai. Ez utóbbiba biztosításba beletartozik például az orvosi elsősegélynyújtás, jogi segítség vagy az orvosi ellátás költsége. Az első biztosítás nem tartalmaz fedezetet lopáskárra, rongálásra, tűzkárra.

		Y							
		0	1	2	3	4	5	6	7
X	0	71087	3722	807	219	51	14	4	0
	1	3022	686	184	71	26	10	3	1
	2	574	138	55	15	8	4	1	1
	3	149	42	21	6	6	1	0	1
	4	29	15	3	2	1	1	0	0
	5	4	1	0	0	0	0	2	0
	6	2	1	0	1	0	0	0	0
	7	1	0	0	1	0	0	0	0
	8	0	0	1	0	0	0	0	0

4.1. táblázat. A táblázatban látható, hogy a két kárszám mentén hogyan oszlanak meg a kötvénytulajdonosok. Kiugróan magas azok száma, akiknek mindkét kárszáma nulla.

A 4.1 táblázat mutatja a kötvénytulajdonosok eloszlását. Látható, hogy a túlnyomó részüknek, 71087 embernek mindkét kárszáma nulla. Akár az első, akár a második kárszám mentén nézzük a kárszám növekedését, látható, hogy egyre

kevesebben tartoznak az egyes csoportokba. Kivételt képez az a néhány egyén, akiknek valamelyik kárszáma kiugróan magas.

Először azt számoltam ki, hogy mi lesz a λ_1 és λ_2 maximum likelihood becslése, ha függetlenséget feltételezünk a két változó között. Tudjuk, hogy Poisson eloszlású minta esetén a paraméter maximum likelihood becslése a minta átlaga.

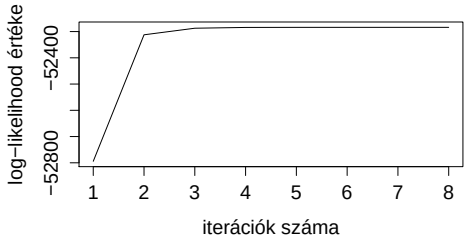
$$\hat{\lambda} = \frac{\sum_{i=1}^{N} k_i}{N} = \bar{X},$$

ahol k_i jelöli az i-edik szerződés kárszámát, N a szerződések számát. Ezt az összefüggést felhasználva a paraméterekre a $\lambda_1 = 0.0810$ és $\lambda_2 = 0.1024$ a becslés adódott.

A két változó közötti kovariancia 0.0269, a korreláció 0.1866. Feltételezve ezért az összefüggőséget kétváltozós Poisson eloszlást illesztettem az adatokra. Ekkor a paraméterekre a következő becslést kaptam:

$$\lambda_1 = 0.06700382$$
$$\lambda_2 = 0.08840047$$
$$\lambda_3 = 0.01396514$$

Ezeket az értékeket használva a paraméterekre a 4.2 táblázatban látható, hogy milyen becsült eloszlást várunk a kárszámokra. Az iterációs lépések során a log-likelihood érték változása láthatóak 4.1 ábrán.



4.1. ábra. A log-likelihood értékek változása az iterációs lépések során kétváltozós Poisson eloszlása esetén.

Mivel a kártalanok száma kiugróan magas, ezért érdemes megnézni, hogy a diagonálisan módosított Poisson hogyan illeszkedik az adatokra. Ehhez többféle $D(x;\theta)$ diagonálisan módosító változót is kipróbáltam.

		Y							
		0	1	2	3	4	5	6	7
X	0	68375	6044	267	8	0	0	0	0
	1	4581	1360	102	4	0	0	0	0
	2	153	78	13	1	0	0	0	0
	3	3	2	1	0	0	0	0	0
	4	0	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	0	0
	6	0	0	0	0	0	0	0	0
	7	0	0	0	0	0	0	0	0
	8	0	0	0	0	0	0	0	0

4.2. táblázat. A λ_1, λ_2 és λ_3 paraméterek becsült értékét használva itt látható, hogy kétváltozós Poisson eloszlást feltételezve a két kárszám mentén hogyan oszlanak meg a kötvénytulajdonosok.

Elsőként $Disc(0)$, majd $Disc(1)$ diagonális módosító változót választottam. A második esetben teljesen ugyanazt az eredményt kaptam, tehát az EM algoritmus becslése szerint az (1,1) cellára a diagonális változó 0 súlyt helyez, vagyis tulajdonképpen visszakapjuk a $Disc(0)$ esetet. Emellett viszont eggyel több paramétert kellett becsülni, ami lassítja a számításokat.

Diagonális módosító változónak Poisson vagy geometriai eloszlásút választva szintén a $Disc(0)$ esetben kapott eredményt kaptam vissza. A pontos eredmények a 4.3 táblázatban összefoglalva láthatók. A p paraméter a keverési arányt mondja meg, vagyis hogy mekkora súlya van a diagonális módosító változónak az eloszlásban. A θ paraméter a 2.3 képlet jelölése szerint. Utolsó oszlopban az iterációs lépések száma. Az iterációs lépések addig ismétlődnek, amíg a log-likelihood relatív változása kisebb, mint 1e-08, vagy a lépésszám eléri a háromszázat.

	λ_1	λ_2	λ_3	p	θ	it
BP	0.0670	0.0884	0.0140	-	-	8
$Disc(0)$	0.3853	0.4871	1.7e-06	0.79	1	124
$Disc(1)$	0.3853	0.4871	1.73e-06	0.79	(1,0)	139
Poisson	0.3853	0.4871	1.84e-06	0.79	0	146
geometriai	0.3853	0.4871	1.8e-06	0.79	1	139

4.3. táblázat.

Lineáris regresszióban az adatpontok illeszkedésének jóságát a determinációs együtthatóval mérhetjük. Az R^2 egy széles körben elfogadott mérőszáma ennek. Nem lineáris regresszióban például az Akaike-féle információs kritériummal (AIC) vagy a bayesi információs kritériummal (BIC) mérhető a modell illeszkedésének jósága. Kiszámításuk a következő képlettel adható meg:

$$AIC = -2l + 2(k+1),$$
$$BIC = -2l + k \log n,$$

ahol l a log-likelihood függvény maximuma, k a modellben becsült paraméterek száma, n a megfigyelések száma. Újabb paraméterek hozzávételével a log-likelihood maximuma növelhető. Mindkét mutató a paraméterek számától függő kifejezéssel bünteti a sok paramétert tartalmazó modellt. A BIC-nél ez a büntetés nagyobb, mint AIC esetén. Amennyiben n értéke nagy, akkor a BIC-t érdemes használni, kis megfigyelésszám esetén esetén az AIC-t. Két modell összehasonlítása előtt eldöntjük, hogy melyik mérőszámot szeretnénk használni. Azután a kapott eredmény alapján a kisebb AIC (vagy az előzetes döntéstől függően BIC) értékű modellt fogadjuk el jobban illeszkedőnek.

Az előző modellek esetén az AIC és BIC értékek a következők:

	l	AIC	BIC
BP	-52283.93	104574	104604
$Disc(0)$	-48630.52	97269	97309
$Disc(1)$	-48630.52	97271	97321
Poisson	-48630.52	97271	97321
geometriai	-48630.52	97271	97321

4.4. táblázat.

Mind az AIC, mind a BIC értéke alapján az látszik, hogy a diagonálisan módosított modell jobban illeszkedik. A 4.5 táblázatban látható az így kapott eredmény.

		Y							
		0	1	2	3	4	5	6	7
X	0	71094	3463	843	137	17	2	0	0
	1	2739	1334	325	53	6	1	0	0
	2	528	257	63	10	1	0	0	0
	3	68	33	8	1	0	0	0	0
	4	7	3	1	0	0	0	0	0
	5	1	0	0	0	0	0	0	0
	6	0	0	0	0	0	0	0	0
	7	0	0	0	0	0	0	0	0
	8	0	0	0	0	0	0	0	0

4.5. táblázat. Itt látható, hogy diagonálisan módosított Poisson eloszlást (módosítás a (0,0) cellán) feltételezve a két kárszám mentén hogyan oszlanak meg a kötvénytulajdonosok.

A modell tovább javítható, ha magyarázó változókat vezetünk be. Ekkor a kötvénytulajdonosok és a gépjármű bizonyos jól kezelhető tulajdonságainak segítségével még pontosabb becslést tudunk kapni a paraméterekre.

Irodalomjegyzék

[1] Lluís Bermúdez i Morata: A priori ratemaking using bivariate Poisson regression models, *Insurance: Mathematics and Economics*, Vol. 44, 2009, pp. 135–141.

[2] Jeff A. Bilmes: A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, 1998

[3] A. P. Dempster, N. M. Laird, D. B. Rubin: Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 39, No. 1., 1977, pp. 1–38.

[4] http://en.wikipedia.org

[5] Edward W. Frees: *Regression Modeling with Actuarial and Financial Applications*, International Series on Actuarial Science, Cambridge University Press, 2010

[6] Trevor Hastie, Robert Tibshirani, Jerome Friedman: *The Elements of Statistical Learning Data Mining, Inference, and Prediction*, Second Edition, Springer Series in Statistics, 2009

[7] Dimitris Karlis, Ioannis Ntzoufras: Analysis of sports data by using bivariate Poisson models, *The Statistician* 52, Part 3, 2003, pp. 381–393

[8] Dimitris Karlis, Ioannis Ntzoufras: Bivariate Poisson and Diagonal Inflated Bivariate Poisson Regression Models in R, *Journal of Statistical Software*, Volume 14, Issue 10, 2005, pp. 1–36.

[9] S. Kocherlakota, K. Kocherlakota: *Bivariate Discrete Distributions*, New York: Marcel and Dekker, Statistics a Series of Textbooks and Monographs, 1992

[10] Geoffrey McLachlan, Thriyambakam Krishnan: *The EM Algorithm and Extensions*, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., 2008

[11] Geoffrey McLachlan, David Peel: *Finite Mixture Models*, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., 2000

[12] Todd K. Moon, Wynn C. Stirling: *Mathematical Methods and Algorithms for Signal Processing*, Prentice-Hall, Inc. Upper Saddle River. New Jersey 07458, 1999

[13] Richard A. Redner, Homer F. Walker: Mixture Densities, Maximum Likelihood and the Em Algorithm, *Society for Industrial and Applied Mathematics*, Vol. 26, No. 2, 1984, pp. 195–239

[14] Jozef L. Teugels (Ed.), Bjørn Sundt (Ed.): *Encyclopedia of Actuarial Science*, John Wiley & Sons, Inc., 2004

[15] Yiu-Kuen Tse: *Nonlife Actuarial Models Theory, Methods and Evaluation*, International Series on Actuarial Science, Cambridge University Press, 2009

[16] Panagiotis Tsiamyrtzis, Dimitris Karlis: Strategies for Efficient Computation of Multivariate Poisson Probabilities, *Communications in Statistics, Simulation and Computation*, Vol. 33, No. 2, 2004, pp. 271–292

Függelék

```
trinom <- function(y1, y2, p0, it){
  p <-c()
  x1 <- c()
  x2 <- c()
  x1[1] = y1*0.25/(0.5 + 0.25*p0)
  x2[1] = y1*(0.25 + 0.25*p0)/(0.5 + 0.25*p0)
  p[1] = (2*x2[1] - y2)/(x2[1] + y2)
  for (k in 1:(it-1)){
    x1[k+1] = y1*0.25/(0.5 + 0.25*p[k])
    x2[k+1] = y1*(0.25 + 0.25*p[k])/(0.5 + 0.25*p[k])
    p[k+1] = (2*x2[k+1] - y2)/(x2[k+1] + y2)
  }
  list(x1, x2, p)
}
```

```
mixgen <- function(g=3, n=10000, pi=c(0.3, 0.2, 0.5),
mu=c(0, 3, 10), sigma=c(1, 0.1, 1)){

  # kevert normalis eloszlasu minta generalasa
  # g az eloszlasok szama, n a minta elemszama
  # pi sulyvektor, mu a varhato ertekek vektora,
   sigma a szorasok vektora

  p <- rep(0, g)
  y <- rep(0, n)
  for(i in 2:g){
    p[i] = p[i-1] + pi[i-1]
  }
  r = runif(n, 0, 1)
```

```
  for(j in 1:n){
    c = sum(r[j]>p)
    y[j] = rnorm(1, mean = mu[c], sd = sigma[c])
  }
  y
}


delta <- function(x, y, eps) max(abs(x-y)/abs(x))<eps

emtnorm <- function(y, g=3, it=40, eps=1e-4){
  # kevert normalis eloszlasbol szarmazo minta parametereinek
    es a keveresi sulyok becslese
  # y: megfigyelesvektor
  # g: hany darab eloszlas kevereke
  # it: iteracios lepesek szama

  n <- length(y)
  pi <- matrix( rep( c(1/g,0),c(g,g*(it-1)) ),nrow=g )
  # minden iteracios lepesnel eltaroljuk, hogy milyen keveresi
    aranyt feltetelezunk, kezdetben mindent azonos sullyal veszunk

  mu <- matrix( rep( c(mean(y),0) , c(g ,g*(it-1)) ),nrow=g )
  sigma <- matrix( rep( c(1:g,0) , c(rep(1,g),g*(it-1)) ),nrow=g)
  # iteracios lepesenkent a mu parameter es sigma becsult erteke

  z <- matrix(rep(0,(g*n)),nrow=g,byrow=T) # segedmatrix
  f <- matrix(rep(0,(g*n)),nrow=g,byrow=T) # segedmatrix
  l = rep(0,it)  # loglikelihood ertekek

  k <- 1
  repeat{
    k<-k+1

    # E-lepes: Z matrix kitoltese
```

```
    v <- matrix(rep(0,g*n),nrow=g)
    for (i in 1:g){
      v[i,] = pi[i,k-1]*dnorm(y, mean = mu[i,k-1],
                          sd = sigma[i,k-1], log = FALSE)
    }
    for(j in 1:n){
      z[,j] = v[,j]/sum(v[,j]) # osztas az oszloposszeggel
    }

    # M-lepes: mu, sigma, pi ujrabecslese
    for (i in 1:g){
      mu[i,k] = (z[i,]\%*\%y)/sum(z[i,])
      sigma[i,k] = sqrt((z[i,]\%*\%(y-mu[i,k-1])^2)/sum(z[i,]))
      pi[i,k] = 1/n * sum(z[i,])
      f[i,] = dnorm(y, mean = mu[i,k], sd = sigma[i,k], log = TRUE)
    }

    # loglikelihood szamitasa
    l[k] = sum(t(z)\%*\%log(pi[,k])) + sum(hadamard.prod(z,f))

    #leallas
    if((delta(mu[,k],mu[,k-1],eps) &&
        delta(sigma[,k],sigma[,k-1],eps) &&
        delta(pi[,k],pi[,k-1],eps) &&
        (l[k]-l[k-1]<0.5) ) || (k+1>it)) break()
  }

  result <-list(pi = pi[,k], mu = mu[,k], sigma = sigma[,k],
                loglike = l[k], max=l[2:k])
  result
}
```