

Debreceni Egyetem
Matematikai Intézet

Az EM algoritmus

Témavezető:
Dr. Ispány Márton
egyetemi adjunktus

Készítette:
Béres Ildikó
alkalmazott matematikus

Debrecen
2007

Tartalomjegyzék

Bevezetés	ii
1. Fejezet	
A maximum likelihood módszer	1
1. Alapfogalmak	1
2. A maximum likelihood becslés	3
3. A Newton-Raphson módszer	7
2. Fejezet	
Az EM algoritmus	9
4. Az EM algoritmus megfogalmazása	9
5. Általánosított EM algoritmus	11
6. Az általánosított EM algoritmus növekvő tulajdonsága	12
7. Az EM algoritmus monotonitása	15
3. Fejezet	
Alkalmazások	17
8. Kétdimenziós normális eloszlású minta hiányzó adatokkal	17
9. Allél gyakoriság becslése	28
Összefoglalás	34
Irodalomjegyzék	35
Függelék A.	36
Függelék B.	40

Bevezetés

Ez a dolgozat az Expectation-Maximization algoritmussal, népszerű nevén az EM algoritmussal foglalkozik. Ez egy igen széles körben alkalmazott algoritmus, a maximum likelihood becslés iteratív számítására alkalmas. Főként olyan feladatoknál hasznos, ahol hiányzó adatok vannak. Szinte minden statisztikai témában előjön a hiányzó adatok problémája. Így nagy jelentősége van egy olyan módszernek, amely megoldást kínál a feladatra. Először 1977-ben publikálta Dempster, Laird és Rubin a Royal Statistical Society folyóiratában. Az EM algoritmus elnevezés is tőlük ered. Hasznos irodalomnak bizonyult még McLachlan és Krishnan könyve, amely átfogó képet ad az EM algoritmusról és annak különböző variációiról elméleti és gyakorlati értelemben is. A dolgozat célja az EM algoritmus és néhány tulajdonságának az ismertetése, valamint konkrét példákon keresztül bemutatni az algoritmus néhány alkalmazását az R statisztikai programcsomag segítségével.

Az első fejezetben a dolgozat megértéséhez szükséges alapfogalmakat ismertetjük. Ezután bevezetjük a maximum likelihood becslés fogalmát. A második fejezetben az EM algoritmust ismertetjük. Végül az alkalmazásokkal foglalkozunk.

Az alkalmazások bemutatásához az *R* statisztikai programcsomagot választottuk. Ez egy adatkezelésre, számításra és grafikai bemutatásra alkalmas programcsomag. A legtöbb szakember statisztikai rendszerként használja. Bárki számára elérhető és letölthető a *www.R – project.org* oldalon. Az *R*-t tekinthetjük az *S* nyelv ingyenes változatának, amelyet a Bell laboratóriumban fejlesztett ki Rick Becker, John Chambers és Allan Wilks. Használható *Windows* és *Unix* környezetben is. A dolgozatban a *Windows*-os változatot használtuk. Többek között az alábbi funkciókat tartalmazza:

- hatékony adatkezelés és tároló szolgáltatás,
- operátorok egy készlete számításokra tömbökben, sajátos mátrixokban,
- eszközök egy nagy, összefüggő egységbe gyűjtése adatelemzésekre,

- grafikai szolgáltatások adatelemzésekre és közvetlen bemutatásra,
- egy fejlett, egyszerű és hatékony programnyelv, mely magában foglal feltételeket, kikötéseket.

1. Fejezet

A maximum likelihood módszer

1. Alapfogalmak

Ebben a részben néhány statisztikai alapfogalmat ismertetünk. Arra törekszünk, hogy a dolgozat megértéséhez adjunk segítséget. Részletesebben a [3] könyvben olvashatunk erről a témáról.

1.1. DEFINÍCIÓ. (Statisztikai mező)

Egy $(\Omega, \mathcal{A}, \mathcal{P})$ hármast statisztikai mezőnek nevezünk, ha az Ω nemüres halmaz az elemi események halmaza, \mathcal{A} az események σ -algebrája, $\mathcal{P} = \{P\}$ pedig valószínűségi mértékek egy családja.

Ha $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ alakban írható, akkor $(\Omega, \mathcal{A}, \mathcal{P}_\theta), \theta \in \Theta$, paraméteres statisztikai mező. θ -t paraméternek, Θ -t paraméterhalmaznak nevezzük.

1.2. DEFINÍCIÓ. (Statisztikai minta)

Az X_1, X_2, \dots, X_n valószínűségi változókat statisztikai mintának (mintának) nevezzük az $(\Omega, \mathcal{A}, \mathcal{P})$ statisztikai mezőn, ha függetlenek, azonos eloszlásúak és $X_i : \Omega \rightarrow \mathbb{R}$.

Az $x_1 = X_1(\omega), \dots, x_n = X_n(\omega)$ szám n -est rögzített $\omega \in \Omega$ esetén minta realizációnak hívjuk.

A gyakorlatban statisztikai mező helyett a mintatéren dolgozunk:

$$(\Omega, \mathcal{A}, \mathcal{P}) \rightarrow (X^n, \mathcal{X}^n, \mathcal{Q})$$

ahol $X^n \subset \mathbb{R}^n$, \mathcal{X}^n pedig X^n Borel halmazait jelöli.

1.3. DEFINÍCIÓ. Statisztika alatt a minta egy mérhető függvényét értjük, vagyis ha $T : X^n \rightarrow \mathbb{R}^d$ egy mérhető függvény, akkor a $T(X_1, \dots, X_n)$ valószínűségi vektorváltozó egy statisztika.

Becslés

Legyen $(\Omega, \mathcal{A}, \mathcal{P}_\theta), \theta \in \Theta$, paraméteres statisztikai mező, X_1, \dots, X_n minta. Célunk a θ paraméter becslése. Feltételezzük, hogy θ -nak van egy igazi θ^* értéke. Erre a θ^* értékre próbálunk meg következtetni a mintából.

1.4. DEFINÍCIÓ. Legyen X_1, \dots, X_n minta az $(\Omega, \mathcal{A}, \mathcal{P}_\theta), \theta \in \Theta$, paraméteres statisztikai mezőn, $\Theta \in \mathbb{R}^d$. A $T : X^n \rightarrow \mathbb{R}^d, (X_1, \dots, X_n) \mapsto T(X_1, \dots, X_n)$ mérhető leképezést a θ paraméter egy becslésének nevezzük és $\hat{\theta}$ -pal jelöljük. Utalhatunk a becslésnél a mintaelemszámra is a $\hat{\theta}_n$ jelöléssel.

Becsülhetjük a θ paraméter egy $g(\theta)$ függvényét is, ekkor a becslés jele $\widehat{g(\theta)}$.

Milyen is legyen a jó becslés?

I. A becslés értékei az igazi paraméter értékek körül ingadozzanak.

1.5. DEFINÍCIÓ. Az $(\Omega, \mathcal{A}, \mathcal{P}_\theta), \theta \in \Theta$, paraméteres statisztikai mezőn a θ paraméter $\hat{\theta}_n$ becslését torzítatlannak nevezzük, ha

$$E_\theta \hat{\theta} = \theta \quad \forall \theta \in \Theta.$$

Ahol E_θ a \mathcal{P}_θ valószínűségi mérték szerinti várható érték képzést jelöli.

A θ paraméter $B(\theta) = E_\theta \hat{\theta} - \theta, \forall \theta \in \Theta$, függvényét torzításnak nevezzük.

II. A becslés szórása a lehető legkisebb legyen.

1.6. DEFINÍCIÓ. Legyen az $(\Omega, \mathcal{A}, \mathcal{P}_\theta), \theta \in \Theta$, statisztikai mezőn $\hat{\theta}_1$ és $\hat{\theta}_2$ két torzítatlan, véges szórású becslése a θ paraméternek. $\hat{\theta}_1$ hatásosabb mint $\hat{\theta}_2$, ha

$$D_\theta^2(\hat{\theta}_1) \leq D_\theta^2(\hat{\theta}_2) \quad \forall \theta \in \Theta.$$

D_θ^2 a θ paraméter szerinti szórásnégyzet: $D_\theta^2(X) = E_\theta(X - E_\theta(X))^2$.

A θ paraméter egy $\hat{\theta}$ becslése hatásos, ha hatásosabb minden más torzítatlan becslésnél.

1.1. TÉTEL. Ha van hatásos becslés, akkor az egyértelmű.

III. A becslés konvergáljon az igazi paraméter értékhez, ha minden mintaelemszám esetén értelmezhető.

1.7. DEFINÍCIÓ. Legyen X_1, \dots, X_n, \dots minta az $(\Omega, \mathcal{A}, \mathcal{P}_\theta), \theta \in \Theta$, statisztikai mezőn, $\hat{\theta}_n$ pedig becsléssorozat.

$\hat{\theta}_n$ konzisztens becslése a θ paraméternek, ha $\hat{\theta}_n \rightarrow \theta$ sztochasztikusan $n \rightarrow \infty$ mellett, vagyis $\forall \varepsilon > 0$ esetén

$$\lim_{n \rightarrow \infty} P_\theta(|\hat{\theta}_n - \theta| > \varepsilon) = 0, \quad \theta \in \Theta.$$

Ha $\hat{\theta}_n \rightarrow \theta$ majdnem mindenütt, $n \rightarrow \infty$, vagyis

$$P_\theta(\lim_{n \rightarrow \infty} \hat{\theta}_n = \theta) = 1, \quad \theta \in \Theta,$$

akkor $\hat{\theta}_n$ erősen konzisztens becslés.

1.8. DEFINÍCIÓ. Egy $P_\theta, \theta \in \Theta$, mértékcsaládot dominálnak nevezünk, ha a P_θ valószínűségi mértékek abszolút folytonosak egy az (Ω, \mathcal{A}) mérhető téren értelmezett μ, σ -véges mértékre nézve.

1.9. DEFINÍCIÓ. Ha a $P_\theta, \theta \in \Theta$, mértékcsalád dominált, akkor az $(\Omega, \mathcal{A}, P_\theta), \theta \in \Theta$, statisztikai mező dominált. $\Theta \subset \mathbb{R}^d$.

1.10. DEFINÍCIÓ. Legyen $(\Omega, \mathcal{A}, P_\theta), \theta \in \Theta$, dominált statisztikai mező. Ekkor az

$$f(x, \theta) = \frac{dP_\theta}{d\mu}(x)$$

Radon-Nikodym deriváltakat általánosított sűrűségfüggvényeknek nevezzük.

2. A maximum likelihood becslés

Tekintsük az X valószínűségi vektor változó által indukált dominált statisztikai mezőt. Legyen x_1, \dots, x_n minta X -re. Ekkor a függetlenség miatt az általánosított sűrűségfüggvény a $\prod_{i=1}^n f(x_i, \theta)$ alakban áll elő. Jelölje $\mathcal{X} \subset \mathbb{R}^n$ a mintateret. Az

$$L : \mathcal{X} \times \Theta \rightarrow \mathbb{R}_+ \quad L(x_1, \dots, x_n, \theta) = \prod_{i=1}^n f(x_i, \theta)$$

függvényt az x_1, \dots, x_n mintához tartozó likelihood függvénynek nevezzük a θ paraméter mellett.

A likelihood függvény diszkrét esetben az x_1, \dots, x_n minta együttes eloszlása, abszolút folytonos esetben pedig az együttes sűrűségfüggvény a θ paraméter mellett.

Célunk a θ paraméter becslése. Ezt a likelihood függvény maximalizálásával érjük el.

2.1. DEFINÍCIÓ. A θ paraméter maximum likelihood becslése az a $\hat{\theta} : \mathcal{X} \rightarrow \Theta$ becslés, melyre

$$L(x_1, \dots, x_n, \hat{\theta}(x_1, \dots, x_n)) \geq L(x_1, \dots, x_n, \theta), \quad \theta \in \Theta,$$

teljesül, azaz $\hat{\theta}$ globális maximumhelye az L likelihood függvénynek.

A maximum likelihood becslést a gyakorlatban differenciálással határozzuk meg, mert szélsőérték. Az alábbi egyenletet kell megoldanunk θ -ban:

$$\frac{\partial L(x_1, \dots, x_n, \theta)}{\partial \theta} = 0. \quad (2.1)$$

Az egyszerűbb számolásért a likelihood függvény logaritmusát vesszük. A logaritmus függvény szigorúan monoton növekvő, így a (2.1) egyenlettel ekvivalens a következő egyenlet:

$$\frac{\partial \ln L(x_1, \dots, x_n, \theta)}{\partial \theta} = 0. \quad (2.2)$$

Ez a likelihood egyenlet.

Megoldásai az úgynevezett stacionárius pontok.

Ha

$$\frac{\partial^2 \ln L(x_1, \dots, x_n, \theta)}{\partial \theta^2} = \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \ln f(x_i, \theta) < 0$$

egy $\hat{\theta}$ stacionárius pontban, akkor $\hat{\theta}$ maximumhely.

A következő tétel elég általános feltételek mellett garantálja, hogy a maximum likelihood becslés erősen konzisztens becslés.

2.1. TÉTEL. Legyen X_1, \dots, X_n minta $f(X, \theta)$ általánosított sűrűségfüggvénnyel az $(\Omega, \mathcal{A}, \mathcal{P}_\theta), \theta \in \Theta$ identifikálható, dominált statisztikai mezőn. A likelihood egyenletnek létezik olyan $\hat{\theta}$ gyöke, melyre $P(\hat{\theta} \rightarrow \theta^*) = 1$, ha

- (1) $\ln f(X, \theta)$ függvény folytonosan differenciálható az $U \subseteq \Theta$ nyílt intervallumon, mely tartalmazza a θ^* igazi paramétert.
- (2) $\exists E(\ln f(X, \theta) \mid \theta^*)$ és véges $\forall \theta \in U$ -ra.

BIZONYÍTÁS. Lásd [3].

□

Vizsgáljuk most a többdimenziós esetet. Legyen Y p -dimenziós valószínűségi vektorváltozó $g(y, \theta)$ általánosított sűrűségfüggvénnyel a $(\Omega, \mathcal{A}, \mathcal{P}_\theta), \theta \in \Theta$, dominált statisztikai mezőn, $\theta = (\theta_1, \dots, \theta_d)^T$. x_1, \dots, x_n egy n elemű megfigyelt minta valamely X valószínűségi vektorváltozóra $f(x, \theta)$ sűrűségfüggvénnyel. Legyen

$$y = (x_1^T, \dots, x_n^T)^T.$$

A likelihood függvény ebben az esetben:

$$L(\theta) = \prod_{i=1}^n f(x_i, \theta) = g(y, \theta).$$

A megoldandó likelihood egyenletrendszer:

$$\frac{\partial \ln L(\theta)}{\partial \theta_j} = 0, \quad j = 1, \dots, d.$$

2.2. TÉTEL. (Cramér): Legyen $(\Omega, \mathcal{A}, \mathcal{P}_\theta), \theta = (\theta_1, \dots, \theta_d)^T \in \Theta$, identifikálható, dominált statisztikai mező $f(X, \theta)$ általánosított sűrűségfüggvénnyel. Tegyük fel, hogy:

- (1) $\exists \frac{\partial}{\partial \theta} f(X, \theta), \frac{\partial^2}{\partial \theta^2} f(X, \theta), \frac{\partial^3}{\partial \theta^3} f(X, \theta)$ a θ^* igazi paraméter egy nyílt környezetében.
- (2) $\exists F_1, F_2$ és F_3 függvények, hogy F_1, F_2 μ szerint integrálhatóak és $E(F_3(x) | \theta^*) < \infty$ és

$$\left| \frac{\partial}{\partial \theta} f(X, \theta) \right| < F_1(X); \quad \left| \frac{\partial^2}{\partial \theta^2} f(X, \theta) \right| < F_2(X);$$

$$\left| \frac{\partial^3}{\partial \theta^3} \ln f(X, \theta) \right| < F_3(X) \quad \forall \theta \in U$$

- (3) Az

$$\mathcal{I}(\theta) = E\left(\left(\frac{\partial}{\partial \theta} \ln f(X, \theta)\right)^2 \mid \theta\right)$$

Fisher-féle információs mennyiség véges $\forall \theta \in U$ -ra és $\mathcal{I}(\theta^*)$ pozitív definit.

Ekkor ha n elég nagy, a likelihood egyenletnek van olyan $\hat{\theta}_n$ gyöke \mathcal{P}_{θ^*} szerint 1 valószínűséggel, melyre

(a) $\hat{\theta}_n \rightarrow \theta^*$ \mathcal{P}_{θ^*} -m.m. (konzisztencia)

(b) $\sqrt{n}(\hat{\theta}_n - \theta^*) \Rightarrow \mathcal{N}(0, \mathcal{I}^{-1}(\theta^*))$ (aszimptotikus normalitás)

BIZONYÍTÁS. Lásd [10].

□

2.2. DEFINÍCIÓ. A loglikelihood függvény θ szerinti másodrendű parciális deriváltjai által meghatározott mátrixot Hesse mátrixnak nevezzük. Jele:

$$H = \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta^T}.$$

Legyen

$$I(\theta, y) = -\frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta^T},$$

azaz a Hesse mátrix (-1) -szerese. Adott feltételek mellett a várt Fisher-féle információs mátrix:

$$\mathcal{I}(\theta) = E[S(Y, \theta)S^T(Y, \theta)|\theta] = -E[I(\theta, Y)|\theta],$$

ahol az

$$S(Y, \theta) = \frac{\partial \ln L(\theta)}{\partial \theta}$$

a loglikelihood függvény gradiens vektora, azaz a score függvény.

A $\hat{\theta}$ maximum likelihood becslés aszimptotikus kovariancia mátrixa megegyezik a várt $\mathcal{I}(\theta)$ információs mátrix inverzével, ami $\mathcal{I}(\hat{\theta})$ -vel közelíthető. Azaz $\hat{\theta}_i = (\hat{\theta})_i$ véletlen hibája (az úgynevezett standard error) így adódik:

$$\varepsilon(\hat{\theta}_i) \approx (\mathcal{I}^{-1}(\hat{\theta}))_{ii}^{1/2}, \quad i = 1, \dots, d.$$

A megfigyelt információs mátrix általában jobban kezelhető, mint a várt információs mátrix, mivel nem kell hozzá várható érték számítás.

Gyakran nem lehetséges analitikusan maximalizálni a likelihood függvényt. Ilyen esetekben számolhatjuk θ maximum likelihood becslését iteratívan, egy Newton-Raphson maximalizációs eljárás, vagy valamely variánsának használatával.

3. A Newton-Raphson módszer

A Newton-Raphson módszernél a likelihood egyenlet megoldásához az

$$S(y, \theta) = 0 \tag{3.1}$$

egyenletben közelítjük az $\ln L(\theta)$ loglikelihood függvény $S(y, \theta)$ gradiens vektorát θ_k körüli Taylor- sorfejtéssel. Azaz

$$S(y, \theta) \approx S(y, \theta_k) - I(\theta_k, y)(\theta - \theta_k).$$

Ha a jobb oldalt 0-val egyenlővé tesszük, kapunk egy új θ -t.

$$S(y, \theta_k) - I(\theta_k, y)(\theta - \theta_k) = 0 \quad / \cdot I^{-1}(\theta_k, y)$$

$$I^{-1}(\theta_k, y)S(y, \theta_k) - \theta + \theta_k = 0$$

$$I^{-1}(\theta_k, y)S(y, \theta_k) + \theta_k = \theta$$

Ezért θ következő iterációját így számíthatjuk ki:

$$\theta_{k+1} = \theta_k + I^{-1}(\theta_k, y)S(y, \theta_k). \tag{3.2}$$

Ha a loglikelihood függvény konkáv és egycsúcsú, akkor a θ_k sorozat θ maximum likelihood becsléséhez konvergál. Egy lépésben konvergál, ha a loglikelihood függvény θ egy másodfokú függvénye. Ha a loglikelihood függvény nem konkáv, akkor nem biztos, hogy θ_k θ -hoz konvergál tetszőleges kezdőértéknél. A Newton-Raphson módszerrel jó kezdőérték és $L(\theta)$ megfelelő tulajdonságai mellett olyan θ_k sorozatot kapunk, amely másodrendben konvergál a (3.1) egyenlet egy θ_* megoldásához.

3.1. DEFINÍCIÓ. *Legyen adott egy $\| \cdot \|$ norma Θ -n. Azt mondjuk, hogy a θ_k sorozat legalább másodrendben konvergál θ_* -hoz, ha létezik olyan h konstans, hogy*

$$\| \theta_{k+1} - \theta_* \| \leq h \| \theta_k - \theta_* \|^2,$$

ahol $k = 0, 1, 2, \dots$.

A Newton-Raphson módszer fő erőssége, hogy a másodrendű konvergencia nagyon gyors. De számos probléma adódhat a módszer alkalmazásakor. Először is minden egyes iterációnál ki kell számolni a $d \times d$ -s $I(\theta_k, y)$ információs mátrixot, (azaz a Hesse mátrix

(-1)-szeresét), továbbá egy d ismeretlenes lineáris egyenletrendszer megoldását. Így amit a Newton-Raphson módszer egy iterációjának számítása megkövetel, valószínűleg hatványozottan növekszik ahogy d egyre nagyobb lesz. Figyelembe kell venni a Hesse mátrix méretét is.

θ_k éppoly gyakran tart a nyeregpontokhoz és a lokális minimumhoz, mint a lokális maximumhoz. A Newton-Raphson módszer egyik előnye, hogy minden egyes k iterációjánál meg kell becsülni $I(\theta_k, y)$ -t, így közvetlenül biztosítja a határértékek kovariancia mátrixának egy becslését az $I^{-1}(\theta_k, y)$ megfigyelt információs mátrix inverze által.

2. Fejezet

Az EM algoritmus

4. Az EM algoritmus megfogalmazása

Az EM algoritmus a maximum likelihood becslés iteratív számítására alkalmas eljárás. Elsősorban olyan feladatoknál hasznos, ahol hiányzó adatok vannak. Adat például a megfigyelések rögzítése során bekövetkezett hiba miatt hiányozhat. Előfordulhat azonban, hogy adataink valamely feltételezett tapasztalatból erednek, ezért a valóságban sohasem figyelhetőek meg.

Az EM algoritmus nem igazi algoritmus, hanem egy metaalgoritmus. Minden egyes iterációja két lépésből áll. Az Expectation Maximization (EM) elnevezés is erre utal. Az első E várható érték lépésre a hiányzó adat kitöltéseként gondolhatunk. A második M maximalizációs lépéssel pedig megbecsüljük a paramétereket.

Legyen Y a megfigyelt, hiányos y adatok valószínűségi vektorváltozója $g(y, \theta)$ sűrűségfüggvénnyel, ahol $\theta = (\theta_1, \dots, \theta_d)^T$ ismeretlen paraméter a Θ paraméterteréren. A megfigyelt y adat vektorra hiányos adatként tekintünk, valamint az úgynevezett hiánytalan adatok egy megfigyelhető függvényeként. Jelölje x a teljes vagy úgynevezett hiánytalan adatok vektorát. Legyen z a pótdatot tartalmazó vektor, amire mint megfigyelhetetlen vagy hiányzó adatként hivatkozunk. Legyen t egy leképezés X -ből Y -ra: $t : X \rightarrow Y$; $t(x) = y$. Például ha $x = (y, z)$, akkor $t(y, z) = y$. Legyen $f(x, \theta)$ az x hiánytalan adat vektorok X valószínűségi változójának a sűrűségfüggvénye. Akkor a hiánytalan (teljes) adat loglikelihood függvénye θ -ra, abban az esetben, ha x teljesen megfigyelhető, a következőképpen adódik:

$$\ln L(x, \theta) = \ln f(x, \theta).$$

Formálisan két mintaterünk van, \mathcal{X} és \mathcal{Y} , valamint egy szürjektív leképezés \mathcal{X} -ből \mathcal{Y} -ra. Az \mathcal{X} -beli x teljes adat vektor megfigyelése helyett az \mathcal{Y} -beli $y = t(x)$ hiányos adat

vektort figyeljük meg. Ebből következik, hogy:

$$g(y, \theta) = \int_{\mathcal{X}(y)} f(x, \theta) dx,$$

ahol $\mathcal{X}(y) = \{x \in \mathcal{X} : t(x) = y\} \subset \mathcal{X}$, azaz $\mathcal{X}(y)$ az \mathcal{Y} inverzképe a t leképezés által.

Az EM algoritmus E lépésében az alábbi feltételes várható értéket számoljuk ki:

$$Q(\theta, \theta_k) = E(\ln L(x, \theta) \mid y, \theta_k).$$

Az M lépésben pedig válasszuk θ_{k+1} -et a $\theta \in \Theta$ paraméter olyan értékének, amely maximalizálja $Q(\theta, \theta_k)$ -t. Azaz

$$Q(\theta_{k+1}, \theta_k) \geq Q(\theta, \theta_k), \quad \forall \theta \in \Theta.$$

Ezt más módon is kifejezhetjük. Azt mondjuk, hogy θ_{k+1} az

$$\mathcal{M}(\theta_k) = \arg \max_{\theta} Q(\theta, \theta_k)$$

halmazhoz tartozik, amely azon pontok halmazát jelöli, melyek maximalizálják $Q(\theta, \theta_k)$ -t.

Az első iteráció:

E lépés:

$$Q(\theta, \theta_0) = E[\ln L(x, \theta) \mid y, \theta_0]$$

M lépés:

$$Q(\theta_1, \theta_0) \geq Q(\theta, \theta_0)$$

A második iterációban θ_0 -t θ_1 -el helyettesítjük:

E lépés:

$$Q(\theta, \theta_1) = E[\ln L(x, \theta) \mid y, \theta_1]$$

M lépés:

$$Q(\theta_2, \theta_1) \geq Q(\theta, \theta_1)$$

Addig folytatjuk az iterációt az E és M lépések váltogatásával, amíg az

$$L(\theta_{k+1}) - L(\theta_k)$$

különbség elegendően kicsi érték lesz.

Alapvetően az EM algoritmus lényege, hogy a $Q(\theta, \theta_k)$ maximalizálása a megfigyelt adatok $\ln g(y, \theta)$ loglikelihood függvényében növekedéshez vezet, azaz

$$L(\theta_{k+1}) \geq L(\theta_k) \quad k = 0, 1, 2, \dots$$

/Ezt az állítást a következő fejezetben részletesen tárgyaljuk. /

A fentiekből láthatjuk, hogy nem szükséges részletezni a pontos leképezést \mathcal{X} -ből \mathcal{Y} -ra, valamint a hiányos adat g sűrűségfüggvényének megfelelő reprezentációját sem a hiánytalan adat f sűrűségfüggvényének tagjaiban. Mindössze a teljes x adat vektor és a megfigyelt y adat vektor által adott X valószínűségi vektorváltozó feltételes sűrűségfüggvényét szükséges meghatároznunk. Ez az E lépés elvégzéséhez szükséges. Az x teljes adat vektor megválasztása nem egyértelmű, ezért úgy választjuk, hogy az E és M lépések minél egyszerűbbek legyenek. Praktikus úgy megadni x -et, hogy felgyorsítsa az algoritmusban a konvergenciát.

5. Általánosított EM algoritmus

A gyakorlatban előfordul, hogy nincs olyan θ paraméterérték, amely globálisan maximalizálja a $Q(\theta, \theta_k)$ függvényt. Ilyen szituációkra definiálta Dempster, Laird, Rubin az általánosított EM algoritmust (Generalized EM Algorithm=GEM algoritmus).

Ennek az algoritmusnak az M lépésében θ_{k+1} -et úgy választjuk, hogy $Q(\theta, \theta_k)$ maximalizálása helyett az alábbi egyenlőtlenséget teljesítse:

$$Q(\theta_{k+1}, \theta_k) \geq Q(\theta, \theta_k), \quad \forall \theta. \quad (5.1)$$

6. Az általánosított EM algoritmus növekvő tulajdonsága

6.1. DEFINÍCIÓ. Legyen $h : (a, b) \rightarrow \mathbb{R}$ függvény. Ha teljesül az alábbi egyenlőtlenség:

$$h(px_1 + qx_2) \leq ph(x_1) + qh(x_2), \quad (6.1)$$

ahol $\forall x_1, x_2 \in (a, b)$ és $\forall p, q \in [0, 1]$, melyre $p + q = 1$, akkor azt mondjuk, hogy h konvex (a, b) -n. A h függvény szigorúan konvex, ha (6.1)-ben szigorú egyenlőtlenség van.

Legyen $h(x)$ egy kétszer differenciálható függvény. $h(x)$ akkor és csak akkor konvex egy (a, b) intervallumon, ha $h''(x) \geq 0 \quad \forall x \in (a, b)$ esetén.

Ha $h''(x) > 0$, akkor $h(x)$ szigorúan konvex.

6.1. TÉTEL. Jensen-egyenlőtlenség.

Tegyük fel, hogy a ξ valószínűségi változó értékei az (a, b) nyílt intervallumba esnek. Ha $h(x)$ konvex (a, b) -n, és $\exists E(h(\xi))$ és $E(\xi)$, akkor

$$E(h(\xi)) \geq h(E(\xi)).$$

Ha $h(x)$ szigorúan konvex, akkor az egyenlőség akkor és csak akkor teljesül, ha $\xi = E(\xi)$ majdnem mindenütt.

BIZONYÍTÁS. Legyen $m = E(\xi)$; $x \in (a, b)$.

Taylor-sorfejtéssel m körül kapjuk, hogy:

$$h(x) = h(m) + h'(m)(x - m) + h''(v) \frac{(x - m)^2}{2} \geq h(m) + h'(x - m) \quad (6.2)$$

$\forall m < v < x$, $(v \in (a, b))$ esetén.

Helyettesítsük az x pontot a ξ valószínűségi változóval, és vegyük a várható értéket:

$$E(h(\xi)) \geq h(m) + h'(m)(E(\xi) - m) = h(m) = h(E(\xi))$$

Amennyiben $h(x)$ szigorúan konvex, $\frac{h''(v)(x-m)^2}{2} > 0$ minden $x \neq m$ -re. Így

$$h(x) = h(m) + h'(m)(x - m) + h''(v) \frac{(x - m)^2}{2} > h(m) + h'(x - m),$$

ahol $x \neq m$. Helyettesítsük az x pontot ismét a ξ valószínűségi változóval, és vegyük a várható értéket:

$$E(h(\xi)) > h(m) + h'(m)(E(\xi) - m) = h(m) = h(E(\xi))$$

A (6.2) beli egyenlőség akkor és csak akkor teljesül, ha $x = m$ majdnem mindenütt. Ebből látszik, hogy

$$E(h(\xi)) = h(E(\xi))$$

akkor és csak akkor, ha $\xi = E(\xi)$ majdnem mindenütt. \square

6.2. TÉTEL. Információs egyenlőtlenség.

Legyenek f és g sűrűségfüggvények egy μ mértékre nézve. Tételezzük fel, hogy $f > 0$ és $g > 0$ μ - majdnem mindenütt. Akkor

$$\int (\ln f) f d\mu \geq \int (\ln g) f d\mu.$$

Az egyenlőség akkor és csak akkor áll fenn, ha $f = g$ μ -majdnem mindenütt.

BIZONYÍTÁS.

$$\int (\ln f(x)) f(x) d\mu(x) - \int (\ln g(x)) f(x) d\mu(x) = \int \ln \frac{f(x)}{g(x)} f(x) d\mu(x) = \int -\ln \frac{g(x)}{f(x)} f(x) d\mu(x)$$

Mivel $-\ln(x)$ szigorúan konvex függvény $x \in (0, \infty)$, alkalmazhatjuk a Jensen-egyenlőtlenséget :

$$\int -\ln \frac{g(x)}{f(x)} f(x) d\mu(x) \geq -\ln \int \frac{g(x)}{f(x)} f(x) d\mu(x) = -\ln \int g(x) d\mu(x) = 0$$

Tehát

$$\int (\ln f(x)) f(x) d\mu(x) - \int (\ln g(x)) f(x) d\mu(x) \geq 0$$

Egyenlőség csak akkor lehet, ha $\frac{g}{f} = \int \frac{g}{f} d\mu$ μ -majdnem mindenütt, de

$$\int \frac{g}{f} f d\mu = 1 \Rightarrow \frac{g}{f} = 1 \text{ } \mu\text{-m.m.} \Rightarrow g = f \text{ } \mu\text{-m.m.}$$

\square

6.3. TÉTEL. Legyen $g(y, \theta)$ a megfigyelt Y adatok és $f(x, \theta)$ a teljes X adatok sűrűségfüggvénye. Ekkor

$$\ln L(\theta_{k+1}) \geq \ln L(\theta_k), \quad k = 0, 1, \dots$$

Ha $Q(\theta_{k+1}, \theta_k) > Q(\theta_k, \theta_k)$, akkor

$$\ln L(\theta_{k+1}) > \ln L(\theta_k), \quad k = 0, 1, \dots$$

BIZONYÍTÁS.

$$Q(\theta, \theta_k) = E[\ln f(X, \theta) | Y = y, \theta_k]$$

$$Q(\theta_k, \theta_k) = E[\ln f(X, \theta_k) | Y = y, \theta_k].$$

Belátjuk, hogy

$$Q(\theta_k, \theta_k) - \ln L(\theta_k) \geq Q(\theta, \theta_k) - \ln L(\theta)$$

minden θ -ra.

$$Q(\theta_k, \theta_k) - \ln L(\theta_k) = E[\ln f(X, \theta_k) | Y = y, \theta_k] - \ln g(y, \theta_k) = E \left[\ln \left[\frac{f(X, \theta_k)}{g(Y, \theta_k)} \right] | Y = y, \theta_k \right] =$$

Az információs egyenlőtlenség alapján:

$$\begin{aligned} &= \int \ln \frac{f(x(y), \theta_k)}{g(y, \theta_k)} \cdot \frac{f(x(y), \theta_k)}{g(y, \theta_k)} dx \geq \int \ln \frac{f(x(y), \theta)}{g(y, \theta)} \cdot \frac{f(x(y), \theta_k)}{g(y, \theta_k)} dx = \\ &= E \left[\ln \left[\frac{f(X, \theta)}{g(Y, \theta)} \right] | Y = y, \theta_k \right] = Q(\theta, \theta_k) - \ln g(y, \theta) = Q(\theta, \theta_k) - \ln L(\theta). \end{aligned}$$

Tehát

$$Q(\theta_k, \theta_k) - \ln L(\theta_k) \geq Q(\theta, \theta_k) - \ln L(\theta), \quad \forall \theta.$$

Vagyis

$$\ln L(\theta) - Q(\theta, \theta_k) \geq \ln L(\theta_k) - Q(\theta_k, \theta_k), \quad \forall \theta. \quad (6.3)$$

Ezért az

$$\ln L(\theta) - Q(\theta, \theta_k)$$

különbség $\theta = \theta_k$ esetén éri el a minimumát. Ha θ_{k+1} -et (5.1) szerint választjuk, és (6.3)-ben θ helyére írjuk, akkor azt kapjuk, hogy

$$\begin{aligned} \ln L(\theta_{k+1}) &= Q(\theta_{k+1}, \theta_k) + [\ln L(\theta_{k+1}) - Q(\theta_{k+1}, \theta_k)] \geq \\ &\geq Q(\theta_k, \theta_k) + [\ln L(\theta_k) - Q(\theta_k, \theta_k)] = \ln L(\theta_k) \end{aligned}$$

Vagyis a $Q(\theta_k, \theta_k)$ növelése a megfigyelt adatok $\ln g(y, \theta)$ loglikelihood függvényében is növekedéshez vezet. \square

7. Az EM algoritmus monotonitása

Dempster, Laird és Rubin (1977) megmutatta, hogy a hiányos adatra vonatkozó $L(\theta)$ likelihood függvény nem csökken egy EM iteráció után, vagyis

$$L(\theta_{k+1}) \geq L(\theta_k) \quad k = 0, 1, 2, \dots \quad (7.1)$$

Ennek belátásához vegyük X feltételes sűrűségfüggvényét $Y = y$ mellett:

$$\frac{f(x, \theta)}{g(y, \theta)}.$$

Ekkor a loglikelihood függvény

$$\begin{aligned} \ln L(\theta) &= \ln(g(y, \theta)) = \ln \left(f(x, \theta) \cdot \frac{g(y, \theta)}{f(x, \theta)} \right) = \ln \left(f(x, \theta) : \frac{f(x, \theta)}{g(y, \theta)} \right) = \\ &= \ln(f(x, \theta)) - \ln \left(\frac{f(x, \theta)}{g(y, \theta)} \right) = \ln(L(x, \theta)) - \ln \left(\frac{f(x, \theta)}{g(y, \theta)} \right). \end{aligned}$$

Vegyük mindkét oldal várható értékét az X feltételes sűrűségfüggvényre $Y = y$ mellett:

$$\ln L(\theta) = E[\ln L(x, \theta | y, \theta_k)] - E \left[\ln \frac{f(x, \theta)}{g(y, \theta)} \middle| y, \theta_k \right].$$

Jelölje $H(\theta, \theta_k)$ az $E \left[\ln \frac{f(x, \theta)}{g(y, \theta)} \middle| y, \theta_k \right]$ várható értéket. Azt kapjuk, hogy

$$\ln L(\theta) = Q(\theta, \theta_k) - H(\theta, \theta_k).$$

Ezt felhasználva $(\ln L(\theta_{k+1}) - \ln L(\theta_k))$ -ra azt kapjuk, hogy

$$\ln L(\theta_{k+1}) - \ln L(\theta_k) = (Q(\theta_{k+1}, \theta_k) - Q(\theta_k, \theta_k)) - (H(\theta_{k+1}, \theta_k) - H(\theta_k, \theta_k)).$$

(7.1) akkor teljesül, ha $\ln L(\theta_{k+1}) - \ln L(\theta_k) \geq 0$, vagyis

$$Q(\theta_{k+1}, \theta_k) - Q(\theta_k, \theta_k) \geq H(\theta_{k+1}, \theta_k) - H(\theta_k, \theta_k). \quad (7.2)$$

Tudjuk, hogy

$$Q(\theta_{k+1}, \theta_k) - Q(\theta_k, \theta_k) \geq 0,$$

mivel θ_{k+1} -et úgy választjuk, hogy

$$Q(\theta_{k+1}, \theta_k) \geq Q(\theta, \theta_k) \quad \forall \theta \in \Theta - ra.$$

Ezért ha az alábbi formula igaz, akkor (7.2) teljesül.

$$H(\theta_{k+1}, \theta_k) - H(\theta_k, \theta_k) \leq 0. \quad (7.3)$$

Bármely θ -ra

$$H(\theta, \theta_k) - H(\theta_k, \theta_k) = E \left[\ln \left\{ \frac{f(x, \theta)}{g(y, \theta)} : \frac{f(x, \theta_k)}{g(y, \theta_k)} \right\} \middle| y, \theta_k \right]$$

A logaritmus függvény konkávitása miatt felhasználhatjuk a Jensen-egyenlőtlenséget.

$$\begin{aligned} E \left[\ln \left\{ \frac{f(x, \theta)}{g(y, \theta)} : \frac{f(x, \theta_k)}{g(y, \theta_k)} \right\} \middle| y, \theta_k \right] &\leq \ln \left[E \left\{ \frac{f(x, \theta)}{g(y, \theta)} : \frac{f(x, \theta_k)}{g(y, \theta_k)} \right\} \middle| y, \theta_k \right] = \\ &= \ln \int_{\mathcal{X}(y)} \frac{f(x, \theta)}{g(y, \theta)} dx = 0, \end{aligned}$$

Ezzel igazoltuk a (7.3) egyenlőtlenséget. Tehát beláttuk, hogy az $L(\theta)$ likelihood függvény nem csökken egy EM iteráció után. Növekszik a likelihood, amennyiben

$$Q(\theta_{k+1}, \theta_k) > Q(\theta, \theta_k).$$

Ezért ha az $\{L(\theta_k)\}$ likelihood értékek sorozata korlátos, akkor ez a sorozat monotonon konvergál egy L_* -hoz.

(7.1) egy következménye az EM algoritmus egyértelműsége. Ha a $\hat{\theta}$ maximum likelihood becslés globálisan maximalizálja $L(\theta)$ -t, ki kell hogy elégítse az alábbi egyenlőtlenséget:

$$Q(\hat{\theta}, \hat{\theta}) \geq Q(\theta, \hat{\theta}) \quad \forall \theta. \quad (7.4)$$

Máskülönben

$$Q(\hat{\theta}, \hat{\theta}) < Q(\theta_0, \theta)$$

teljesül valamely θ_0 -ra, beleértve, hogy

$$L(\theta_0) > L(\hat{\theta}),$$

ami ellentmond annak a ténynek, hogy $\hat{\theta}$ az $L(\theta)$ globális maximalizálója.

Látható a (7.4) egyenlőtlenség deriváltjából, hogy $\hat{\theta}$ gyöke az alábbi egyenletnek:

$$\left[\frac{\partial Q(\theta, \hat{\theta})}{\partial \theta} \right]_{\theta=\hat{\theta}} = 0.$$

3. Fejezet

Alkalmazások

8. Kétdimenziós normális eloszlású minta hiányzó adatokkal

Legyen az $U = (U_1, U_2)^T$ valószínűségi vektorváltozó kétdimenziós normális eloszlású

$$E(U) = m = (m_1, m_2)^T$$

várható értékkel és

$$\text{var}(U) = D = \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix}$$

szórás mátrix-szal, azaz

$$U \sim \mathcal{N}_2(m, D).$$

Ekkor a sűrűségfüggvény:

$$f(u) = \frac{1}{(2\pi)(\det D)^{\frac{1}{2}}} e^{-\frac{1}{2}(u-m)^T D^{-1}(u-m)}.$$

Vegyünk U -ból egy n elemű mintát. Tegyük fel, hogy r megfigyelésnél nem hiányzik adat, r_1 megfigyelésnél hiányzik az első változó, r_2 megfigyelésnél pedig a második változó. $n = r + r_1 + r_2$. A $\theta = (m_1, m_2, d_{11}, d_{12}, d_{22})^T$ paraméter becslése a célunk.

Jelölje a teljesen megfigyelt adatokat

$$u_j = (u_{1j}, u_{2j})^T \quad j = 1, \dots, r.$$

Jelölje u_{2j} , $j = r+1, \dots, r+r_1$, azokat a megfigyeléseket, ahol az első u_{1j} változó hiányzik, és u_{1j} , $j = r+r_1+1, \dots, n$, azokat a megfigyeléseket, ahol a második u_{2j} változó hiányzik. Ekkor az x teljes adat vektor:

$$x = (u_1^T, \dots, u_n^T)^T.$$

A megfigyelt, hiányos adatok vektora:

$$y = (u_1^T, \dots, u_r^T, v^T)^T,$$

ahol

$$v = (u_{2,r+1}, \dots, u_{2,r+r_1}, u_{1,r+r_1+1}, \dots, u_{1,n})^T.$$

A z hiányzó adat vektor:

$$z = (u_{1,r+1}, \dots, u_{1,r+r_1}, u_{2,r+r_1+1}, \dots, u_{2,n})^T.$$

Az E lépésben az x teljes adat vektor loglikelihood függvényének feltételes várható értékét számoljuk ki:

$$Q(\theta, \theta_k) = E(\ln L(x, \theta) \mid y, \theta_k)$$

Az x teljes adat likelihood függvénye:

$$L(x, \theta) = \prod_{j=1}^n \frac{1}{(2\pi)(\det D)^{\frac{1}{2}}} e^{-\frac{1}{2}(u_j - m)^T D^{-1}(u_j - m)}$$

A loglikelihood függvény:

$$\begin{aligned} \ln L(x, \theta) &= \sum_{j=1}^n \ln \frac{1}{(2\pi)(\det D)^{\frac{1}{2}}} e^{-\frac{1}{2}(u_j - m)^T D^{-1}(u_j - m)} = \\ &= -n \ln(2\pi) - \frac{1}{2} n \ln(\det D) - \frac{1}{2} \sum_{j=1}^n (u_j - m)^T D^{-1}(u_j - m). \end{aligned}$$

Ismert, hogy

$$\begin{aligned} \det D &= d_{11}d_{22} - d_{12}^2, \\ D^{-1} &= \frac{1}{d_{11}d_{22} - d_{12}^2} \begin{pmatrix} d_{22} & -d_{12} \\ -d_{12} & d_{11} \end{pmatrix}. \end{aligned}$$

Így

$$\begin{aligned} &-\frac{1}{2} \sum_{j=1}^n (u_j - m)^T D^{-1}(u_j - m) = \\ &= -\frac{1}{2} \frac{1}{d_{11}d_{22} - d_{12}^2} \sum_{j=1}^n \left(\begin{pmatrix} u_{1j} & u_{2j} \end{pmatrix} - \begin{pmatrix} m_1 & m_2 \end{pmatrix} \right) \begin{pmatrix} d_{22} & -d_{12} \\ -d_{12} & d_{11} \end{pmatrix} \left(\begin{pmatrix} u_{1j} \\ u_{2j} \end{pmatrix} - \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} \right) \\ &= -\frac{1}{2} \frac{1}{d_{11}d_{22} - d_{12}^2} (A + B + C), \end{aligned}$$

ahol

$$\begin{aligned}
A &= \sum_{j=1}^n \begin{pmatrix} u_{1j} & u_{2j} \end{pmatrix} \begin{pmatrix} d_{22} & -d_{12} \\ -d_{12} & d_{11} \end{pmatrix} \begin{pmatrix} u_{1j} \\ u_{2j} \end{pmatrix} = \\
&= d_{22} \sum_{j=1}^n u_{1j}u_{1j} + d_{11} \sum_{j=1}^n u_{2j}u_{2j} - 2d_{12} \sum_{j=1}^n u_{1j}u_{2j}, \\
B &= -2 \sum_{j=1}^n \begin{pmatrix} u_{1j} & u_{2j} \end{pmatrix} \begin{pmatrix} d_{22} & -d_{12} \\ -d_{12} & d_{11} \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} = \\
&= -2(d_{22}m_1 - d_{12}m_2) \sum_{j=1}^n u_{1j} + 2(d_{11}m_2 - d_{12}m_1) \sum_{j=1}^n u_{2j}, \\
C &= n \left(\begin{pmatrix} m_1 & m_2 \end{pmatrix} \begin{pmatrix} d_{22} & -d_{12} \\ -d_{12} & d_{11} \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} \right) = \\
&= n(d_{22}m_1^2 - 2d_{12}m_1m_2 - d_{11}m_2^2).
\end{aligned}$$

Azt kapjuk, hogy

$$\begin{aligned}
\ln L(x, \theta) &= -n \ln(2\pi) - \frac{1}{2}n \ln(d_{11}d_{22} - d_{12}^2) - \frac{1}{2} \frac{1}{d_{11}d_{22} - d_{12}^2} \cdot \\
&\cdot [d_{22} \sum_{j=1}^n u_{1j}u_{1j} + d_{11} \sum_{j=1}^n u_{2j}u_{2j} - 2d_{12} \sum_{j=1}^n u_{1j}u_{2j} - \\
&- 2(d_{22}m_1 - d_{12}m_2) \sum_{j=1}^n u_{1j} - 2(d_{11}m_2 - d_{12}m_1) \sum_{j=1}^n u_{2j} + \\
&+ n(d_{22}m_1^2 - 2d_{12}m_1m_2 - d_{11}m_2^2)].
\end{aligned}$$

Legyen

$$\begin{aligned}
 T_1 &= \sum_{j=1}^n u_{1j}, \\
 T_2 &= \sum_{j=1}^n u_{2j}, \\
 T_{11} &= \sum_{j=1}^n u_{1j}u_{1j}, \\
 T_{12} &= \sum_{j=1}^n u_{1j}u_{2j}, \\
 T_{22} &= \sum_{j=1}^n u_{2j}u_{2j}.
 \end{aligned}$$

Akkor

$$\begin{aligned}
 \ln L(x, \theta) &= -n \ln(2\pi) - \frac{1}{2}n \ln(d_{11}d_{22} - d_{12}^2) - \frac{1}{2} \frac{1}{d_{11}d_{22} - d_{12}^2} \cdot \\
 &\quad \cdot [d_{22}T_{11} + d_{11}T_{22} - 2d_{12}T_{12} - \\
 &\quad - 2(d_{22}m_1 - d_{12}m_2)T_1 - 2(d_{11}m_2 - d_{12}m_1)T_2 + \\
 &\quad + n(d_{22}m_1^2 - 2d_{12}m_1m_2 - d_{11}m_2^2)]
 \end{aligned}$$

Látható, hogy $L(x, \theta)$ a reguláris exponenciális eloszláscsaládhoz tartozik az alábbi elégséges statisztikával:

$$T = (T_1, T_2, T_{11}, T_{12}, T_{22})^T.$$

Az E lépéshez, azaz, hogy kiszámoljuk $Q(\theta, \theta_k) = E(\ln L(x, \theta) \mid y, \theta_k)$ -t, ismernünk kell az elégséges statisztikák feltételes várható értékét. Ehhez a következőkre van szükségünk:

$$\begin{aligned}
 E(U_{1j} \mid u_{2j}, \theta_k) & \quad j = r + 1, \dots, r + r_1 \\
 E(U_{1j}^2 \mid u_{2j}^2, \theta_k) & \quad j = r + 1, \dots, r + r_1 \\
 E(U_{2j} \mid u_{1j}, \theta_k) & \quad j = r + r_1 + 1, \dots, n \\
 E(U_{2j}^2 \mid u_{1j}^2, \theta_k) & \quad j = r + r_1 + 1, \dots, n.
 \end{aligned}$$

8.1. TÉTEL. Legyen $\xi \sim \mathcal{N}_n(m, D)$ n dimenzós normális eloszlású m várható értékkel és D szórás mátrix-szal. Vegyük $\xi = (\xi_1, \dots, \xi_n)^T$ két vektorra való felbontását: $n = n_1 + n_2$; $\xi_1 = (\xi_1, \dots, \xi_{n_1})^T$; $\xi_2 = (\xi_{n_1+1}, \dots, \xi_{n_1+n_2})^T$. Ennek megfelelően:

$$m = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}; \quad D = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix}.$$

Tegyük fel, hogy D_{11} invertálható. Legyen $\xi_{2.1} = \xi_2 - D_{21}D_{11}^{-1}\xi_1$. Ekkor ξ_1 és $\xi_{2.1}$ függetlenek és rendre n_1 és n_2 dimenziós normális eloszlásúak:

$$\xi_1 \sim \mathcal{N}_{n_1}(m_1, D_{11}); \quad \xi_{2.1} \sim \mathcal{N}_{n_2}(m_{2.1}, D_{22.1}),$$

ahol

$$m_{2.1} = m_2 - D_{21}D_{11}^{-1}m_1; \quad D_{22.1} = D_{22} - D_{21}D_{11}^{-1}D_{12}.$$

BIZONYÍTÁS. Lásd [2]. □

8.2. TÉTEL. Vegyük a (8.1) tétel jelöléseit és feltételeit. Akkor ξ_2 feltételes eloszlása a $\xi_1 = w_1$ feltétel mellett n_2 dimenziós normális eloszlás, melynek várható értéke: $D_{21}D_{11}^{-1}w_1 + m_{2.1}$, szórás mátrixa: $D_{22.1}$.

BIZONYÍTÁS. Lásd [2]. □

A (8.1) és (8.2) alapján U_2 feltételes eloszlása az $U_1 = u_1$ feltétel mellett r_2 dimenziós normális eloszlás. A várható értéke:

$$d_{12}d_{11}^{-1}u_1 + m_2 - d_{12}d_{11}^{-1}m_1 = m_2 + d_{12}d_{11}^{-1}(u_1 - m_1).$$

Szórásnégyzete:

$$d_{22} - d_{12}^2d_{11}^{-1}.$$

$$\begin{aligned} E(U_{2j} | u_{1j}, \theta_k) &= \hat{u}_{2j} = \hat{m}_2 + \frac{\hat{d}_{12}}{\hat{d}_{11}}(u_{1j} - \hat{m}_1) & j = r + r_1 + 1, \dots, n \\ E(U_{2j}^2 | u_{1j}, \theta_k) &= \hat{u}_{2j}^2 + \hat{d}_{22} - \frac{\hat{d}_{12}^2}{\hat{d}_{11}} & j = r + r_1 + 1, \dots, n \end{aligned}$$

U_1 feltételes eloszlása az $U_2 = u_2$ feltétel mellett r_1 dimenziós normális eloszlás. A várható értéke:

$$m_1 + \frac{d_{12}}{d_{22}}(u_2 - m_2).$$

Szórásnégyzete:

$$d_{22} - \frac{d_{12}^2}{d_{11}}.$$

$$E(U_{1j} | u_{2j}, \theta_k) = \hat{u}_{1j} = \hat{m}_1 + \frac{\hat{d}_{12}}{\hat{d}_{22}}(u_{2j} - \hat{m}_2) \quad j = r+1, \dots, r+r_1$$

$$E(U_{1j}^2 | u_{2j}, \theta_k) = \hat{u}_{1j}^2 + \hat{d}_{11} - \frac{\hat{d}_{12}^2}{\hat{d}_{22}} \quad j = r+1, \dots, r+r_1$$

Akkor az elégséges statisztikák E lépéshez szükséges feltételes várható értéke:

$$\hat{T}_1 = E(T_1 | u_{2j}, \theta_k) = \sum_{j=1}^r u_{1j} + \sum_{j=r+1}^{r+r_1} \left(\hat{m}_1 + \frac{\hat{d}_{12}}{\hat{d}_{22}}(u_{2j} - \hat{m}_2) \right) + \sum_{j=r+r_1+1}^n u_{1j}$$

$$\hat{T}_2 = E(T_2 | u_{1j}, \theta_k) = \sum_{j=1}^{r+r_1} u_{2j} + \sum_{j=r+r_1+1}^n \left(\hat{m}_2 + \frac{\hat{d}_{12}}{\hat{d}_{11}}(u_{1j} - \hat{m}_1) \right)$$

$$\begin{aligned} \hat{T}_{11} &= E(T_{11} | u_{2j}, \theta_k) = \sum_{j=1}^r u_{1j}u_{1j} + \sum_{j=r+1}^{r+r_1} \left(\left(\hat{m}_1 + \frac{\hat{d}_{12}}{\hat{d}_{22}}(u_{2j} - \hat{m}_2) \right)^2 + \hat{d}_{11} - \frac{\hat{d}_{12}^2}{\hat{d}_{22}} \right) + \\ &+ \sum_{j=r+r_1+1}^n u_{1j}u_{1j} \end{aligned}$$

$$\hat{T}_{22} = E(T_{22} | u_{1j}, \theta_k) = \sum_{j=1}^{r+r_1} u_{2j}u_{2j} + \sum_{j=r+r_1+1}^n \left(\left(\hat{m}_2 + \frac{\hat{d}_{12}}{\hat{d}_{11}}(u_{1j} - \hat{m}_1) \right)^2 + \hat{d}_{22} - \frac{\hat{d}_{12}^2}{\hat{d}_{11}} \right)$$

$$\begin{aligned} \hat{T}_{12} &= E(T_{12} | u_{1j}, u_{2j}, \theta_k) = \sum_{j=1}^r u_{1j}u_{2j} + \sum_{j=r+1}^{r+r_1} \left(\left(\hat{m}_1 + \frac{\hat{d}_{12}}{\hat{d}_{22}}(u_{2j} - \hat{m}_2) \right) u_{2j} \right) + \\ &+ \sum_{j=r+r_1+1}^n \left(u_{1j} \left(\hat{m}_2 + \frac{\hat{d}_{12}}{\hat{d}_{11}}(u_{1j} - \hat{m}_1) \right) \right) \end{aligned}$$

$$\begin{aligned}
Q(\theta, \theta_k) &= E(\ln L(x, \theta) \mid y, \theta_k) = -n \ln(2\pi) - \frac{1}{2}n \ln(\widehat{d}_{11}\widehat{d}_{22} - \widehat{d}_{12}^2) - \frac{1}{2} \frac{1}{\widehat{d}_{11}\widehat{d}_{22} - \widehat{d}_{12}^2} \cdot \\
&\cdot \left[\widehat{d}_{22} \left(\sum_{j=1}^r u_{1j}u_{1j} + \sum_{j=r+1}^{r+r_1} \left((\widehat{m}_1 + \frac{\widehat{d}_{12}}{\widehat{d}_{22}}(u_{2j} - \widehat{m}_2))^2 + \widehat{d}_{11} - \frac{\widehat{d}_{12}^2}{\widehat{d}_{22}} \right) + \sum_{j=r+r_1+1}^n u_{1j}u_{1j} \right) + \right. \\
&+ \widehat{d}_{11} \left(\sum_{j=1}^{r+r_1} u_{2j}u_{2j} + \sum_{j=r+r_1+1}^n \left((\widehat{m}_2 + \frac{\widehat{d}_{12}}{\widehat{d}_{11}}(u_{1j} - \widehat{m}_1))^2 + \widehat{d}_{22} - \frac{\widehat{d}_{12}^2}{\widehat{d}_{11}} \right) \right) - \\
&- 2\widehat{d}_{12} \left[\sum_{j=1}^r u_{1j}u_{2j} + \sum_{j=r+1}^{r+r_1} \left((\widehat{m}_1 + \frac{\widehat{d}_{12}}{\widehat{d}_{22}}(u_{2j} - \widehat{m}_2))u_{2j} \right) + \right. \\
&+ \left. \sum_{j=r+r_1+1}^n (u_{1j}(\widehat{m}_2 + \frac{\widehat{d}_{12}}{\widehat{d}_{11}}(u_{1j} - \widehat{m}_1))) \right] - \\
&- 2(\widehat{d}_{22}\widehat{m}_1 - \widehat{d}_{12}\widehat{m}_2) \left(\sum_{j=1}^r u_{1j} + \sum_{j=r+1}^{r+r_1} \left(\widehat{m}_1 + \frac{\widehat{d}_{12}}{\widehat{d}_{22}}(u_{2j} - \widehat{m}_2) \right) + \sum_{j=r+r_1+1}^n u_{1j} \right) - \\
&- 2(\widehat{d}_{11}\widehat{m}_2 - \widehat{d}_{12}\widehat{m}_1) \left(\sum_{j=1}^{r+r_1} u_{2j} + \sum_{j=r+r_1+1}^n \left(\widehat{m}_2 + \frac{\widehat{d}_{12}}{\widehat{d}_{11}}(u_{1j} - \widehat{m}_1) \right) \right) + \\
&+ n(\widehat{d}_{22}\widehat{m}_1^2 - 2\widehat{d}_{12}\widehat{m}_1\widehat{m}_2 - \widehat{d}_{11}\widehat{m}_2^2)]
\end{aligned}$$

Az M lépésben $Q(\theta, \theta_k)$ maximalizálásával a következőt kapjuk $\widehat{\theta}$ -ra :

$$\begin{aligned}
m_1 &= \frac{\widehat{T}_1}{n} \\
m_2 &= \frac{\widehat{T}_2}{n} \\
d_{11} &= \frac{\widehat{T}_{11}}{n} - \frac{\widehat{T}_1\widehat{T}_1}{n^2} \\
d_{22} &= \frac{\widehat{T}_{22}}{n} - \frac{\widehat{T}_2\widehat{T}_2}{n^2} \\
d_{12} &= \frac{\widehat{T}_{12}}{n} - \frac{\widehat{T}_1\widehat{T}_2}{n^2}
\end{aligned}$$

Nézzük meg egy konkrét példán keresztül. Az *R* statisztikai programcsomag *BSDA* csomagjában található *Fabric* nevű adathalmazzal dolgozunk.

```
> Fabric
      Type With Without
1      1   12      8
2      2    3      4
3      3   12     15
4      4   16     14
5      5    4      6
6      6   24     21
7      7   11     10
8      8   17     15
9      9   19     22
10     10    8      7
```

Az adatok 10 különböző ruha lágyságára vonatkoznak öblítővel (*with* oszlop), illetve anélkül (*without* oszlop) mosva.

A *with* oszlopot vegyük az első változónak. Tegyük fel, hogy hiányzik az 6., 7. és 8. érték. A *without* oszlop pedig legyen a második változó. És tekintsük hiányszónak a 9. és 10. értéket.

```
> U<-matrix(c(12,3,12,16,4,0,0,0,19,8,8,4,15,14,6,21,10,15,0,0),
+ nrow=2,ncol=10,byrow=TRUE);
> U
```

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]   12   3  12  16   4   0   0   0  19    8
[2,]    8   4  15  14   6  21  10  15   0    0
```

A kezdő értékeket válasszuk a következőképpen. A várható értékekre vegyük az átlagot, így $m_1 = 10.6$, $m_2 = 11.6$. Helyettesítsük a hiányzó értékeket az átlaggal, és úgy számoljunk szórást és kovarianciát.

```
> Um<-matrix(c(12,3,12,16,4,10.6,10.6,10.6,19,8,8,4,15,14,6,21,10,15,11.6
+ ,11.6),
+ nrow=2,ncol=10,byrow=TRUE);
> var(Um[1,])
[1] 23.524
> var(Um[2,])
[1] 24.65289
> cov(Um[1,],Um[2,])
[1] 11.93378
```

A FÜGGELÉK A-ban található az első néhány iteráció az *R* programcsomagban. A többi iterációt ezekhez hasonlóan számolhatjuk ki. A következő táblázat tartalmazza az eredményeket.

iteráció	m_1	m_2	d_{11}	d_{22}	d_{12}
1.	10.60000	11.60000	23.52400	24.65289	11.93378
2.	11.12216	11.91424	28.62488	27.79890	19.45915
3.	11.45466	12.00614	30.46534	28.47855	23.13750
4.	11.64735	12.01190	31.37718	28.45714	24.95358
5.	11.76796	11.99705	32.02557	28.29004	25.93594
6.	11.84798	11.97995	32.54719	28.11656	26.50493
7.	11.90275	11.96506	32.96662	27.96868	26.84911
8.	11.94086	11.95318	33.29559	27.85166	27.06434
9.	11.96764	11.94411	33.54745	27.76269	27.20273
10.	11.98655	11.93733	33.73653	27.69662	27.29375
11.	11.99995	11.93235	33.87638	27.64828	27.35471
12.	12.00945	11.92872	33.97867	27.61325	27.39612
13.	12.01619	11.92610	34.05287	27.58802	27.42456
14.	12.02097	11.92422	34.10638	27.56993	27.44425
15.	12.02436	11.92287	34.14478	27.55700	27.45796
16.	12.02677	11.92190	34.17227	27.54777	27.46757
17.	12.02848	11.92122	34.19189	27.54120	27.47431
18.	12.02969	11.92073	34.20587	27.53653	27.47906
19.	12.03055	11.92038	34.21582	27.53321	27.48241
20.	12.03116	11.92013	34.22289	27.53084	27.48478
21.	12.03160	11.91995	34.22792	27.52917	27.48645
22.	12.03190	11.91983	34.23149	27.52798	27.48764
23.	12.03212	11.91974	34.23403	27.52713	27.48848
24.	12.03228	11.91968	34.23583	27.52653	27.48907
25.	12.03239	11.91963	34.23711	27.52610	27.48949

TÁBLÁZAT 1

iteráció	m_1	m_2	d_{11}	d_{22}	d_{12}
26.	12.03246	11.91960	34.23801	27.52580	27.48979
27.	12.03252	11.91958	34.23866	27.52559	27.49000
28.	12.03256	11.91956	34.23911	27.52544	27.49016
29.	12.03259	11.91955	34.23944	27.52533	27.49026
30.	12.03261	11.91954	34.23967	27.52525	27.49034
31.	12.03262	11.91954	34.23983	27.52520	27.49039
32.	12.03263	11.91953	34.23994	27.52516	27.49043
33.	12.03264	11.91953	34.24003	27.52513	27.49046
34.	12.03264	11.91953	34.24008	27.52511	27.49048
35.	12.03265	11.91953	34.24013	27.52510	27.49049
36.	12.03265	11.91953	34.24015	27.52509	27.49050
37.	12.03265	11.91952	34.24018	27.52508	27.49051
38.	12.03265	11.91952	34.24019	27.52508	27.49051
39.	12.03265	11.91952	34.24020	27.52507	27.49051
40.	12.03265	11.91952	34.24021	27.52507	27.49052
41.	12.03265	11.91952	34.24021	27.52507	27.49052

TÁBLÁZAT 2

9. Allél gyakoriság becslése

Sok helyzetben szükség van arra, hogy tudjuk milyen a vércsoportunk. Amikor valaki vért kap vagy vért ad, meghatározzák a vércsoportját. Mint tudjuk három allél létezik, az A, a B és a 0. Minden ember két allélt örököl, az egyiket az anyától, a másikat az apától. Így négy fenotípus létezik: az A, a B, az AB és a 0. Az A és B allélok genetikailag dominánsak a 0-val szemben. Így az a személy aki az egyik szülőtől A, a másiktól 0 allélt örököl, vagyis A/0 genotípusú, ugyanúgy A fenotípusú, mint az, aki mindkét szülőtől A allélt örököl, vagyis A/A genotípusú. Hasonlóképpen mind a B/0 és a B/B genotípus a B fenotípusnak felel meg.

Vegyünk egy n elemű mintát. Jelölje n_A az A, n_B a B, n_{AB} az AB és n_O a 0 fenotípusúak számát. Továbbá $n_{A/A}$, $n_{A/O}$, $n_{B/B}$, $n_{B/O}$, $n_{A/B}$, $n_{O/O}$ a genotípusok számát. Ekkor n_A , n_B , n_{AB} , n_O a megfigyelt Y adatok, az $n_{A/A}$, $n_{A/O}$, $n_{B/B}$, $n_{B/O}$, $n_{A/B}$, $n_{O/O}$ genotípusok pedig a hiánytalan X adatok. Célunk a három különböző allél gyakoriságának a becslése. Jelöljük ezeket p_A , p_B és p_O -al.

Tudjuk, hogy az allél gyakoriság nemnegatív, és eleget tesz a $p_A + p_B + p_O = 1$ feltételnek. Továbbá a populáció genetikai klasszikus Hardy-Weinberg törvénye kimondja, hogy amennyiben különbözik a két allél egy genotípusban, úgy annak a genotípusnak a gyakorisága megegyezik a megfelelő allélok gyakoriságának kétszeres szorzatával. Például az A/A genotípus gyakorisága p_A^2 , az A/0 genotípus gyakorisága pedig $2p_A p_O$. Vagyis

$$p_{A/A} = p_A^2$$

$$p_{A/O} = 2p_A p_O.$$

Továbbá

$$p_{B/B} = p_B^2$$

$$p_{B/O} = 2p_B p_O$$

$$p_{A/B} = 2p_A p_B$$

$$p_{O/O} = p_O^2.$$

Annak a valószínűsége, hogy $n_{A/A}$ számú A/A genotípus fordul elő az n elemű mintánkban $\underbrace{p_A^2 \cdot \dots \cdot p_A^2}_{n_{A/A} \text{ db}} = (p_A^2)^{n_{A/A}}$. Hasonlóképpen az $n_{A/0}, n_{B/B}, n_{B/0}, n_{A/B}, n_{0/0}$ számú A/0, B/B, B/0, A/B és 0/0 genotípusok gyakorisága rendre $(2p_A p_0)^{n_{A/0}}, (p_B^2)^{n_{B/B}}, (2p_B p_0)^{n_{B/0}}, (2p_A p_B)^{n_{A/B}}, (p_0^2)^{n_{0/0}}$. Így a hiánytalan X adat likelihood függvénye:

$$f(X, p) = (p_A^2)^{n_{A/A}} \cdot (2p_A p_0)^{n_{A/0}} \cdot (p_B^2)^{n_{B/B}} \cdot (2p_B p_0)^{n_{B/0}} \cdot (2p_A p_B)^{n_{A/B}} \cdot (p_0^2)^{n_{0/0}} \cdot \binom{n}{n_{A/A} n_{A/0} n_{B/B} n_{B/0} n_{A/B} n_{0/0}}.$$

Logaritmusát véve megkapjuk a loglikelihoodjét:

$$\begin{aligned} \ln f(X, p) &= n_{A/A} \ln p_A^2 + n_{A/0} \ln 2p_A p_0 + n_{B/B} \ln p_B^2 + \\ &+ n_{B/0} \ln(2p_B p_0) + n_{A/B} \ln(2p_A p_B) + n_0 \ln p_0^2 + \\ &+ \ln \binom{n}{n_{A/A} n_{A/0} n_{B/B} n_{B/0} n_{A/B} n_{0/0}}. \end{aligned}$$

Vegyük ennek a feltételes várható értékét a megfigyelt n_A, n_B, n_{AB}, n_0 (Y) adatokra. Az általános paramétervektor $\hat{p} = (\hat{p}_A, \hat{p}_B, \hat{p}_0)^T$. Ez az EM algoritmus E lépése.

$$E(n_{AB} | Y) = n_{AB}$$

$$E(n_0 | Y) = n_0$$

mivel n_{AB} és n_0 Y -mérhető. A függetlenség miatt pedig

$$E(n_{A/A} | Y) = E n_{A/A} = n_A \frac{\hat{p}_A^2}{\hat{p}_A^2 + 2\hat{p}_A \hat{p}_0}$$

$$E(n_{A/0} | Y) = E n_{A/0} = n_A \frac{2\hat{p}_A \hat{p}_0}{\hat{p}_A^2 + 2\hat{p}_A \hat{p}_0}$$

$$E(n_{B/B} | Y) = E n_{B/B} = n_B \frac{\hat{p}_B^2}{\hat{p}_B^2 + 2\hat{p}_B \hat{p}_0}$$

$$E(n_{B/0} | Y) = E n_{B/0} = n_B \frac{2\hat{p}_B \hat{p}_0}{\hat{p}_B^2 + 2\hat{p}_B \hat{p}_0}.$$

Az egyszerűség kedvéért nevezzük el a fenti várható értékeket az alábbi módon:

$$E(n_{A/A} | Y) = \hat{n}_{A/A}$$

$$E(n_{A/0} | Y) = \hat{n}_{A/0}$$

$$E(n_{B/B} | Y) = \hat{n}_{B/B}$$

$$E(n_{B/0} | Y) = \hat{n}_{B/0}$$

Tehát a loglikelihood várható értéke:

$$\begin{aligned} E(\ln f(X, p) | Y) = Q(p, \hat{p}) &= \hat{n}_{A/A} \ln p_A^2 + \hat{n}_{A/0} \ln 2p_A p_0 + \hat{n}_{B/B} \ln p_B^2 + \\ &+ \hat{n}_{B/0} \ln(2p_B p_0) + n_{AB} \ln(2p_A p_B) + n_0 \ln p_0^2 + \\ &\ln \left(\frac{n}{\hat{n}_{A/A} \hat{n}_{A/0} \hat{n}_{B/B} \hat{n}_{B/0} n_{A/B} \hat{n}_{0/0}} \right). \end{aligned}$$

A következő lépés, hogy $Q(p, \hat{p})$ -t maximalizáljuk. A Lagrange-féle multiplikátoros módszerrel könnyen megtalálhatjuk a szélsőértékeket. A Lagrange-függvény a mi esetünkben a következő:

$$L(p, \lambda) = Q(p, \hat{p}) + \lambda(p_A + p_B + p_0 - 1).$$

Az elsőrendű parciális deriváltaknak a szélsőérték helyen el kell tűnniük. Ha létezik megoldás, akkor az egyértelműen szélsőértéket ad. Az alábbi egyenletrendszert kell megoldanunk:

$$\frac{\partial}{\partial p_A} L(p, \lambda) = 0$$

$$\frac{\partial}{\partial p_B} L(p, \lambda) = 0$$

$$\frac{\partial}{\partial p_0} L(p, \lambda) = 0$$

$$\frac{\partial}{\partial \lambda} L(p, \lambda) = 0.$$

Megoldása:

$$\begin{aligned}\frac{\partial}{\partial p_A} L(p, \lambda) &= \hat{n}_{A/A} \frac{1}{p_A^2} 2p_A + \hat{n}_{A/0} \frac{1}{2p_A p_0} 2p_0 + n_{AB} \frac{1}{2p_A p_B} 2p_B + \lambda = \\ &= \frac{2\hat{n}_{A/A}}{p_A} + \frac{\hat{n}_{A/0}}{p_A} + \frac{n_{AB}}{p_A} + \lambda\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial p_B} L(p, \lambda) &= \hat{n}_{B/B} \frac{1}{p_B^2} 2p_B + \hat{n}_{B/0} \frac{1}{2p_B p_0} 2p_0 + n_{AB} \frac{1}{2p_A p_B} 2p_A + \lambda = \\ &= \frac{2\hat{n}_{B/B}}{p_B} + \frac{\hat{n}_{B/0}}{p_B} + \frac{n_{AB}}{p_B} + \lambda\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial p_0} L(p, \lambda) &= \hat{n}_{A/0} \frac{1}{2p_A p_0} 2p_A + \hat{n}_{B/0} \frac{1}{2p_B p_0} 2p_B + n_0 \frac{1}{p_0^2} 2p_0 + \lambda = \\ &= \frac{\hat{n}_{A/0}}{p_0} + \frac{\hat{n}_{B/0}}{p_0} + \frac{2n_0}{p_0} + \lambda\end{aligned}$$

$$\frac{\partial}{\partial \lambda} L(p, \lambda) = p_A + p_B + p_0 - 1$$

Egyenletrendszerünk:

$$\frac{2\hat{n}_{A/A}}{p_A} + \frac{\hat{n}_{A/0}}{p_A} + \frac{n_{AB}}{p_A} + \lambda = 0 \quad (9.1)$$

$$\frac{2\hat{n}_{B/B}}{p_B} + \frac{\hat{n}_{B/0}}{p_B} + \frac{n_{AB}}{p_B} + \lambda = 0 \quad (9.2)$$

$$\frac{\hat{n}_{A/0}}{p_0} + \frac{\hat{n}_{B/0}}{p_0} + \frac{2n_0}{p_0} + \lambda = 0 \quad (9.3)$$

$$p_A + p_B + p_0 - 1 = 0. \quad (9.4)$$

(9.1) - (9.2):

$$\frac{2\hat{n}_{A/A}}{p_A} + \frac{\hat{n}_{A/0}}{p_A} + \frac{n_{AB}}{p_A} = \frac{2\hat{n}_{B/B}}{p_B} + \frac{\hat{n}_{B/0}}{p_B} + \frac{n_{AB}}{p_B} \quad (9.5)$$

(9.2)-(9.3):

$$\frac{2\hat{n}_{B/B}}{p_B} + \frac{\hat{n}_{B/0}}{p_B} + \frac{n_{AB}}{p_B} = \frac{\hat{n}_{A/0}}{p_0} + \frac{\hat{n}_{B/0}}{p_0} + \frac{2n_0}{p_0} \quad (9.6)$$

(9.5):

$$\frac{2\hat{n}_{A/A} + \hat{n}_{A/0} + n_{AB}}{2\hat{n}_{B/B} + \hat{n}_{B/0} + n_{AB}} = \frac{p_A}{p_B} = \frac{1 - p_B - p_0}{p_B} \quad (9.7)$$

(9.6)-ből:

$$p_B = p_0 \frac{2\hat{n}_{B/B} + \hat{n}_{B/0} + n_{AB}}{\hat{n}_{A/0} + \hat{n}_{B/0} + 2n_0} \quad (9.8)$$

(9.7)-ben p_B -t (9.8) szerint helyettesítve:

$$\begin{aligned}
\frac{2\hat{n}_{A/A} + \hat{n}_{A/0} + n_{AB}}{2\hat{n}_{B/B} + \hat{n}_{B/0} + n_{AB}} &= \frac{1 - p_0 \frac{2\hat{n}_{B/B} + \hat{n}_{B/0} + n_{AB}}{\hat{n}_{A/0} + \hat{n}_{B/0} + 2n_0} - p_0}{p_0 \frac{2\hat{n}_{B/B} + \hat{n}_{B/0} + n_{AB}}{\hat{n}_{A/0} + \hat{n}_{B/0} + 2n_0}} \\
p_0 \frac{2\hat{n}_{A/A} + \hat{n}_{A/0} + n_{AB}}{\hat{n}_{A/0} + \hat{n}_{B/0} + 2n_0} &= 1 - p_0 \frac{2\hat{n}_{B/B} + \hat{n}_{B/0} + n_{AB}}{\hat{n}_{A/0} + \hat{n}_{B/0} + 2n_0} - p_0 \\
p_0 \frac{2\hat{n}_{A/A} + \hat{n}_{A/0} + 2n_{AB} + 2\hat{n}_{B/B} + \hat{n}_{B/0}}{\hat{n}_{A/0} + \hat{n}_{B/0} + 2n_0} &= 1 - p_0 \\
p_0 \frac{2\hat{n}_{A/A} + 2\hat{n}_{A/0} + 2n_{AB} + 2\hat{n}_{B/B} + 2\hat{n}_{B/0} + 2n_0}{\hat{n}_{A/0} + \hat{n}_{B/0} + 2n_0} &= 1
\end{aligned}$$

$$\begin{aligned}
p_0 &= \frac{\hat{n}_{A/0} + \hat{n}_{B/0} + 2n_0}{2(\hat{n}_{A/A} + \hat{n}_{A/0} + n_{AB} + \hat{n}_{B/B} + \hat{n}_{B/0} + n_0)} \\
p_0 &= \frac{\hat{n}_{A/0} + \hat{n}_{B/0} + 2n_0}{2n} \tag{9.9}
\end{aligned}$$

$$\begin{aligned}
p_B &= \frac{\hat{n}_{A/0} + \hat{n}_{B/0} + 2n_0}{2n} \cdot \frac{2\hat{n}_{B/B} + \hat{n}_{B/0} + n_{AB}}{\hat{n}_{A/0} + \hat{n}_{B/0} + 2n_0} \\
p_B &= \frac{2\hat{n}_{B/B} + \hat{n}_{B/0} + n_{AB}}{2n} \tag{9.10}
\end{aligned}$$

$$\begin{aligned}
p_A &= 1 - p_B - p_0 = 1 - \frac{2\hat{n}_{B/B} + \hat{n}_{B/0} + n_{AB}}{2n} - \frac{\hat{n}_{A/0} + \hat{n}_{B/0} + 2n_0}{2n} = \\
&= 1 - \frac{2\hat{n}_{B/B} + \hat{n}_{B/0} + n_{AB} + \hat{n}_{A/0} + \hat{n}_{B/0} + 2n_0}{2n} = \\
&= \frac{2\hat{n}_{A/A} + 2\hat{n}_{A/0} + 2n_{AB} + 2\hat{n}_{B/B} + 2\hat{n}_{B/0} + 2n_0 - 2\hat{n}_{B/B} - 2\hat{n}_{B/0} - n_{AB} - \hat{n}_{A/0} - 2n_0}{2n}
\end{aligned}$$

$$p_A = \frac{2\hat{n}_{A/A} + \hat{n}_{A/0} + n_{AB}}{2n} \tag{9.11}$$

Nézzük meg most egy konkrét példán keresztül a három allél gyakoriságának becslését. A [11] dolgozatban található az alföldi lakosság körében vizsgált 8045 személy vércsoportja. A vizsgált személyek között 3315 A vércsoportú, 1577 B vércsoportú, 2365 0 vércsoportú és 788 AB vércsoportú volt. Az allél gyakoriság becslése az R programcsomagban a FÜGGELÉK B-ben található.

Az iteráció eredményeit az alábbi táblázatban összegeztem:

iteráció	\hat{p}_0	\hat{p}_A	\hat{p}_B
0.	0.3	0.45	0.25
1.	0.48008864	0.3433011	0.1758125
2.	0.5286676	0.3091991	0.1621332
3.	0.5383631	0.3016202	0.0.1600167
4.	0.5402452	0.3000878	0.1596669
5.	0.5406093	0.2997863	0.1596044
6.	0.5406799	0.2997275	0.1595927
7.	0.5406936	0.299716	0.1595904
8.	0.5406962	0.2997138	0.15959
9.	0.5406967	0.2997134	0.1595899
10.	0.5406968	0.2997133	0.1595899
11.	0.5406968	0.2997133	0.1595899

TÁBLÁZAT 3

Jól megfigyelhető, a konvergencia gyorsasága.

Összefoglalás

A dolgozat első felében az EM algoritmust és annak egy tulajdonságát ismertettük, a második felében pedig két konkrét példán keresztül mutattuk be az algoritmus alkalmazását az R statisztikai programcsomag segítségével. Ennek a dolgozatnak nem volt célja, és a terjedelme sem engedi, az algoritmus minden tulajdonságának ismertetését. Az EM algoritmusnak számtalan variációja ismert, melyek ebben a dolgozatban nem kaptak helyet, de számos információt találhat az érdeklődő olvasó az erről a témáról írt szakirodalmakban. Az algoritmus alkalmazásainak csak nagyon kis részét érintettük. Little és Rubin (1987) részletesen foglalkozott a témával.

Az EM algoritmus két lépésből áll. Az első E lépésnél kitöltjük a hiányzó adatot, és vesszük az így kapott teljes adat ismeretlen paraméter melletti loglikelihood függvényének a feltételes várható értékét: $Q(\theta, \theta_k) = E(\ln L(x, \theta) \mid y, \theta_k)$. A második M lépésben pedig maximalizáljuk az E lépésben kapott várható értéket. Az E és M lépéseket újra és újra végrehajtjuk, mindig az újonnan kapott becsléssel helyettesítve a régit, ameddig meg nem látjuk a konvergenciát. Az EM algoritmus legfontosabb tulajdonsága, hogy a megfigyelt adatok loglikelihood függvényében növekedéshez vezet.

A példák megoldása során arra jutottunk, hogy az algoritmus nem feltétlenül egyszerű és gyors. Ezek a tényezők nagyban függenek az adathalmaztól és az induló érték megválasztásától, mely a statisztikusra van bízva.

Irodalomjegyzék

- [1] Dempster, A. P., Laird, N. M., Rubin, D. B., *Maximum likelihood from incomplete data via the EM algorithm (with discussion)*, Journal of the Royal Statistical Society B **39** (1977), 1-38.
- [2] Fazekas István, *Valószínűségszámítás*, Kossuth Egyetemi Kiadó, Debrecen, 2000.
- [3] Fazekas István szerk. , *Bevezetés a matematikai statisztikába*, Kossuth Egyetemi Kiadó, Debrecen, 2000.
- [4] Járai Antal, *Mérték és integrál*, Nemzeti Tankönyvkiadó, Budapest, 2002.
- [5] Lajkó Károly, *Analízis II.*, Egyetemi Jegyzet, DE, Debrecen, 2001.
- [6] Lajkó Károly, *Analízis III.*, Egyetemi Jegyzet, DE, Debrecen, 2001.
- [7] Little, RJA., Rubin DB., *Statistical Analysis with missing Data*, Wiley, New York, 1987.
- [8] Lange, Kenneth, *Numerical Analysis for Statisticians*, Springer, New York, 1999.
- [9] McLachlan, Geoffrey J., Krishnan, Thriyambakam, *The EM Algorithm and Extensions*, John Wiley & Sons, Inc., New York, 1997.
- [10] Móri F. Tamás, Székely J. Gábor, *Többváltozós statisztikai analízis*, Műszaki Könyvkiadó, Budapest, 1986.
- [11] <http://www.nemennyi.net/konyvtar/01-AVERCSOP/refined/01-AVERCSOP.pdf>

FÜGGELÉK A

```

> m1<-10.6; m2<-11.6; d11<-23.524; d22<-24.65289; d12<-11.93378;n<-10;
> r<-5; r1<-3;
> m11<-((sum(U[1,1:r]))+(sum(m1+(d12/d22)*(U[2,(r+1):(r+r1)]-m2)))+
+ (sum(U[1,(r+r1+1):n])))/n
> m21<-((sum(U[2,1:(r+r1)]))+(sum(m2+(d12/d11)*(U[1,(r+r1+1):n]-m1)))/n
> d111<-((sum(U[1,1:r]*U[1,1:r]))+
+ (sum((m1+(d12/d22)*(U[2,(r+1):(r+r1)]-m2))^2+d11-(d12^2/d22)))+
+ (sum(U[1,(r+r1+1):n]*U[1,(r+r1+1):n]))/n-m11*m11;
> d221<-((sum(U[2,1:(r+r1)]*U[2,1:(r+r1)])))+
+ (sum((m2+(d12/d11)*(U[1,(r+r1+1):n]-m1))^2+d22-(d12^2/d11)))/n-
+ m21*m21;
> d121<-((sum(U[1,1:r]*U[2,1:r]))+
+ (sum((m1+(d12/d22)*(U[2,(r+1):(r+r1)]-m2))*U[2,(r+1):(r+r1)])))+
+ (sum(U[1,(r+r1+1):n]*(m2+(d12/d11)*(U[1,(r+r1+1):n]-m1))))/n-m11*m21;
> m11; m21; d111; d221; d121;
[1] 11.12216
[1] 11.91424
[1] 28.62488
[1] 27.7989
[1] 19.45915
> m12<-((sum(U[1,1:r]))+(sum(m11+(d121/d221)*(U[2,(r+1):(r+r1)]-m21)))+
+ (sum(U[1,(r+r1+1):n])))/n
> m22<-((sum(U[2,1:(r+r1)]))+(sum(m21+(d121/d111)*
+ (U[1,(r+r1+1):n]-m11))))/n
> d112<-((sum(U[1,1:r]*U[1,1:r]))+

```



```

+ (sum((m11+(d121/d221)*(U[2,(r+1):(r+r1)]-m21))^2+d111-(d121^2/d221)))+
+ (sum(U[1,(r+r1+1):n]*U[1,(r+r1+1):n]))/n-m12*m12;
> d222<-((sum(U[2,1:(r+r1)]*U[2,1:(r+r1)]))+
+ (sum((m21+(d121/d111)*(U[1,(r+r1+1):n]-m11))^2+d221-(d121^2/d111)))/n-
+ m22*m22;
> d122<-((sum(U[1,1:r]*U[2,1:r]))+
+ (sum((m11+(d121/d221)*(U[2,(r+1):(r+r1)]-m21))*U[2,(r+1):(r+r1)]))+
+ (sum(U[1,(r+r1+1):n]*(m21+(d121/d111)*(U[1,(r+r1+1):n]-m11))))/n-
+ m12*m22;
> m12; m22; d112; d222; d122;
[1] 11.45466
[1] 12.00614
[1] 30.46534
[1] 28.47855
[1] 23.13750
> m13<-((sum(U[1,1:r]))+(sum(m12+(d122/d222)*(U[2,(r+1):(r+r1)]-m22)))+
+ (sum(U[1,(r+r1+1):n])))/n
> m23<-((sum(U[2,1:(r+r1)]))+sum(m22+(d122/d112)*
+ (U[1,(r+r1+1):n]-m12)))/n
> d113<-((sum(U[1,1:r]*U[1,1:r]))+
+ (sum((m12+(d122/d222)*(U[2,(r+1):(r+r1)]-m22))^2+d112-(d122^2/d222)))+
+ (sum(U[1,(r+r1+1):n]*U[1,(r+r1+1):n]))/n-m13*m13;
> d223<-((sum(U[2,1:(r+r1)]*U[2,1:(r+r1)]))+
+ (sum((m22+(d122/d112)*(U[1,(r+r1+1):n]-m12))^2+d222-(d122^2/d112)))/n-
+ m23*m23;
> d123<-((sum(U[1,1:r]*U[2,1:r]))+
+ (sum((m12+(d122/d222)*(U[2,(r+1):(r+r1)]-m22))*U[2,(r+1):(r+r1)]))+
+ (sum(U[1,(r+r1+1):n]*(m22+(d122/d112)*(U[1,(r+r1+1):n]-m12))))/n-
+ m13*m23;
> m13; m23; d113; d223; d123;

```

```

[1] 11.64735
[1] 12.01190
[1] 31.37718
[1] 28.45714
[1] 24.95358
> m14<-((sum(U[1,1:r]))+(sum(m13+(d123/d223)*(U[2,(r+1):(r+r1)]-m23)))+
+ (sum(U[1,(r+r1+1):n])))/n
> m24<-((sum(U[2,1:(r+r1)]))+(sum(m23+(d123/d113)*
+ (U[1,(r+r1+1):n]-m13))))/n
> d114<-((sum(U[1,1:r]*U[1,1:r]))+
+ (sum((m13+(d123/d223)*(U[2,(r+1):(r+r1)]-m23))^2+d113-(d123^2/d223)))+
+ (sum(U[1,(r+r1+1):n]*U[1,(r+r1+1):n])))/n-m14*m14;
> d224<-((sum(U[2,1:(r+r1)]*U[2,1:(r+r1)])))+
+ (sum((m23+(d123/d113)*(U[1,(r+r1+1):n]-m13))^2+d223-(d123^2/d113)))/n-
+ m24*m24;
> d124<-((sum(U[1,1:r]*U[2,1:r]))+
+ (sum((m13+(d123/d223)*(U[2,(r+1):(r+r1)]-m23))*U[2,(r+1):(r+r1)])))+
+ (sum(U[1,(r+r1+1):n]*(m23+(d123/d113)*(U[1,(r+r1+1):n]-m13))))/n-
+ m14*m24;
> m14; m24; d114; d224; d124;
[1] 11.76796
[1] 11.99705
[1] 32.02557
[1] 28.29004
[1] 25.93594
> m15<-((sum(U[1,1:r]))+(sum(m14+(d124/d224)*(U[2,(r+1):(r+r1)]-m24)))+
+ (sum(U[1,(r+r1+1):n])))/n
> m25<-((sum(U[2,1:(r+r1)]))+(sum(m24+(d124/d114)*
+ (U[1,(r+r1+1):n]-m14))))/n
> d115<-((sum(U[1,1:r]*U[1,1:r]))+

```

```

+ (sum((m14+(d124/d224)*(U[2,(r+1):(r+r1)]-m24))^2+d114-(d124^2/d224)))+
+ (sum(U[1,(r+r1+1):n]*U[1,(r+r1+1):n]))/n-m15*m15;
> d225<-((sum(U[2,1:(r+r1)]*U[2,1:(r+r1)]))+
+ (sum((m24+(d124/d114)*(U[1,(r+r1+1):n]-m14))^2+d224-(d124^2/d114)))/n-
+ m25*m25;
> d125<-((sum(U[1,1:r]*U[2,1:r]))+
+ (sum((m14+(d124/d224)*(U[2,(r+1):(r+r1)]-m24))*U[2,(r+1):(r+r1)]))+
+ (sum(U[1,(r+r1+1):n]*(m24+(d124/d114)*(U[1,(r+r1+1):n]-m14))))/n-
+ m15*m25;
> m15; m25; d115; d225; d125;
[1] 11.84798
[1] 11.97995
[1] 32.54719
[1] 28.11656
[1] 26.50493

```

FÜGGELÉK B

```
> n<-8045; nA<-3315; nB<-1577; n0<-2365; nAB<-788;
> p0A<-0.45; p0B<-0.25; p00<-0.3;
> n0AA<-((nA*p0A^2)/(p0A^2+(2*p0A*p00)))
> n0A0<-((nA*2*p0A*p00)/(p0A^2+(2*p0A*p00)))
> n0BB<-((nB*(p0B)^2)/(p0B^2+(2*p0B*p00)))
> n0B0<-((nB*2*p0B*p00)/((p0B)^2+(2*p0B*p00)))
> p10<-((n0A0+n0B0+(2*n0))/(2*n))
> p1B<-(((2*n0BB)+n0B0+nAB)/(2*n))
> p1A<-(((2*n0AA)+n0A0+nAB)/(2*n))
> p10;p1B;p1A
[1] 0.4808864
[1] 0.1758125
[1] 0.3433011
> n1AA<-((nA*(p1A)^2)/((p1A)^2+(2*p1A*p10)))
> n1A0<-((nA*2*p1A*p10)/((p1A)^2+(2*p1A*p10)))
> n1BB<-((nB*(p1B)^2)/((p1B)^2+(2*p1B*p10)))
> n1B0<-((nB*2*p1B*p10)/((p1B)^2+(2*p1B*p10)))
> p20<-((n1A0+n1B0+(2*n0))/(2*n))
> p2B<-(((2*n1BB)+n1B0+nAB)/(2*n))
> p2A<-(((2*n1AA)+n1A0+nAB)/(2*n))
> p20;p2B;p2A
[1] 0.5286676
[1] 0.1621332
[1] 0.3091991
> n2AA<-((nA*(p2A)^2)/((p2A)^2+(2*p2A*p20)))
```

```

> n2A0<-((nA*2*p2A*p20)/((p2A)^2+(2*p2A*p20)))
> n2BB<-((nB*(p2B)^2)/((p2B)^2+(2*p2B*p20)))
> n2B0<-((nB*2*p2B*p20)/((p2B)^2+(2*p2B*p20)))
> p30<-((n2A0+n2B0+(2*n0))/(2*n))
> p3B<-(((2*n2BB)+n2B0+nAB)/(2*n))
> p3A<-(((2*n2AA)+n2A0+nAB)/(2*n))
> p30;p3B;p3A
[1] 0.5383631
[1] 0.1600167
[1] 0.3016202
> n3AA<-((nA*(p3A)^2)/((p3A)^2+(2*p3A*p30)))
> n3A0<-((nA*2*p3A*p30)/((p3A)^2+(2*p3A*p30)))
> n3BB<-((nB*(p3B)^2)/((p3B)^2+(2*p3B*p30)))
> n3B0<-((nB*2*p3B*p30)/((p3B)^2+(2*p3B*p30)))
> p40<-((n3A0+n3B0+(2*n0))/(2*n))
> p4A<-(((2*n3AA)+n3A0+nAB)/(2*n))
> p4B<-(((2*n3BB)+n3B0+nAB)/(2*n))
> p40;p4B;p4A
[1] 0.5402452
[1] 0.1596669
[1] 0.3000878
> n4AA<-((nA*(p4A)^2)/((p4A)^2+(2*p4A*p40)))
> n4A0<-((nA*2*p4A*p40)/((p4A)^2+(2*p4A*p40)))
> n4BB<-((nB*(p4B)^2)/((p4B)^2+(2*p4B*p40)))
> n4B0<-((nB*2*p4B*p40)/((p4B)^2+(2*p4B*p40)))
> p50<-((n4A0+n4B0+(2*n0))/(2*n))
> p5A<-(((2*n4AA)+n4A0+nAB)/(2*n))
> p5B<-(((2*n4BB)+n4B0+nAB)/(2*n))
> p50;p5B;p5A
[1] 0.5406093

```

```

[1] 0.1596044
[1] 0.2997863
> n5AA<-((nA*(p5A)^2)/((p5A)^2+(2*p5A*p50)))
> n5A0<-((nA*2*p5A*p50)/((p5A)^2+(2*p5A*p50)))
> n5BB<-((nB*(p5B)^2)/((p5B)^2+(2*p5B*p50)))
> n5B0<-((nB*2*p5B*p50)/((p5B)^2+(2*p5B*p50)))
> p60<-((n5A0+n5B0+(2*n0))/(2*n))
> p6A<-(((2*n5AA)+n5A0+nAB)/(2*n))
> p6B<-(((2*n5BB)+n5B0+nAB)/(2*n))
> p60;p6B;p6A
[1] 0.5406799
[1] 0.1595927
[1] 0.2997275
> n6AA<-((nA*(p6A)^2)/((p6A)^2+(2*p6A*p60)))
> n6A0<-((nA*2*p6A*p60)/((p6A)^2+(2*p6A*p60)))
> n6BB<-((nB*(p6B)^2)/((p6B)^2+(2*p6B*p60)))
> n6B0<-((nB*2*p6B*p60)/((p6B)^2+(2*p6B*p60)))
> p70<-((n6A0+n6B0+(2*n0))/(2*n))
> p7A<-(((2*n6AA)+n6A0+nAB)/(2*n))
> p7B<-(((2*n6BB)+n6B0+nAB)/(2*n))
> p70;p7B;p7A
[1] 0.5406936
[1] 0.1595904
[1] 0.299716
> n7AA<-((nA*(p7A)^2)/((p7A)^2+(2*p7A*p70)))
> n7A0<-((nA*2*p7A*p70)/((p7A)^2+(2*p7A*p70)))
> n7BB<-((nB*(p7B)^2)/((p7B)^2+(2*p7B*p70)))
> n7B0<-((nB*2*p7B*p70)/((p7B)^2+(2*p7B*p70)))
> p80<-((n7A0+n7B0+(2*n0))/(2*n))
> p8A<-(((2*n7AA)+n7A0+nAB)/(2*n))

```

```

> p8B<-(((2*n7BB)+n7B0+nAB)/(2*n))
> p80;p8B;p8A
[1] 0.5406962
[1] 0.15959
[1] 0.2997138
> n8AA<-((nA*(p8A)^2)/((p8A)^2+(2*p8A*p80)))
> n8A0<-((nA*2*p8A*p80)/((p8A)^2+(2*p8A*p80)))
> n8BB<-((nB*(p8B)^2)/((p8B)^2+(2*p8B*p80)))
> n8B0<-((nB*2*p8B*p80)/((p8B)^2+(2*p8B*p80)))
> p90<-((n8A0+n8B0+(2*n0))/(2*n))
> p9A<-(((2*n8AA)+n8A0+nAB)/(2*n))
> p9B<-(((2*n8BB)+n8B0+nAB)/(2*n))
> p90;p9B;p9A
[1] 0.5406967
[1] 0.1595899
[1] 0.2997134
> n9AA<-((nA*(p9A)^2)/((p9A)^2+(2*p9A*p90)))
> n9A0<-((nA*2*p9A*p90)/((p9A)^2+(2*p9A*p90)))
> n9BB<-((nB*(p9B)^2)/((p9B)^2+(2*p9B*p90)))
> n9B0<-((nB*2*p9B*p90)/((p9B)^2+(2*p9B*p90)))
> p100<-((n9A0+n9B0+(2*n0))/(2*n))
> p10A<-(((2*n9AA)+n9A0+nAB)/(2*n))
> p10B<-(((2*n9BB)+n9B0+nAB)/(2*n))
> p100;p10B;p10A
[1] 0.5406968
[1] 0.1595899
[1] 0.2997133
> n10AA<-((nA*(p10A)^2)/((p10A)^2+(2*p10A*p100)))
> n10A0<-((nA*2*p10A*p100)/((p10A)^2+(2*p10A*p100)))
> n10BB<-((nB*(p10B)^2)/((p10B)^2+(2*p10B*p100)))

```

```
> n10B0<-((nB*2*p10B*p100)/((p10B)^2+(2*p10B*p100)))
> p110<-((n10A0+n10B0+(2*n0))/(2*n))
> p11A<-(((2*n10AA)+n10A0+nAB)/(2*n))
> p11B<-(((2*n10BB)+n10B0+nAB)/(2*n))
> p110;p11B;p11A
[1] 0.5406968
[1] 0.1595899
[1] 0.2997133
```