Bayesi módszerek a térbeli statisztikában

Diplomamunka

Írta: Torma Róbert
matematikus szak

Témavezető: Arató Miklós egyetemi docens
Valószínűségelméleti és Statisztika Tanszék

Eötvös Loránd Tudományegyetem
Természettudományi Kar
2010.

Tartalomjegyzék

Bevezetés

A bayesi statisztika kiindulópontja, hogy a megbecsülni kívánt paramétereket is valószínűségi változónak tekintjük valamilyen általunk választott apriori eloszlással, és a mintaelemek eloszlását feltételes eloszlás formájában adjuk meg. A paraméterek feltételes eloszlása a megfigyelt értékekre az aposteriori eloszlás, amelynek valamilyen funkcionálját tekintjük a paraméter becslésének. Sokszor a megfigyelések, paraméterek és segédváltozók egy olyan összetett rendszerét állítjuk fel, amelyben az aposteriori eloszlás meghatározásához szükséges számítások nem végezhetők el egzakt módon, hanem valamilyen közelítő eljárásra van szükség. Erre szolgál a bayesi statisztika egyik legfontosabb módszere, az MCMC (Markov Chain Monte Carlo) algoritmus, amelynek során egy megfelelően konstruált Markov-lánc segítségével mintát veszünk az aposteriori eloszlásból, és az így kapott tapasztalati eloszlással közelítjük azt. Az MCMC algoritmust az első részben tekintjük át.

Térbeli statisztikai problémák olyankor merülnek fel, amikor a megfigyelések valamilyen térbeli elhelyezkedéssel rendelkeznek, például egy bizonyos statisztikai adat egy ország összes megyéjében rendelkezésre áll. Ilyenkor szeretnénk a becslés elvégzésénél figyelembe venni az egyes megfigyelések térbeli közelségét, például feltételezzük, hogy két szomszédos megye adottságai egymáshoz hasonlóak, így a becslés során a két megyében a mért adatok közti különbséget inkább hajlandóak vagyunk a véletlen számlájára írni, mint két, egymástól távol elhelyezkedő megye esetében. Az ilyen térbeli összefüggések bayesi megközelítéséhez jó segédeszköz a statisztikus mechanikából származó ún. Potts-modell. Ebben a modellben a területek egy bizonyos számú színosztályba sorolásait tekintjük, és a színezések halmazán definiálunk egy valószínűségeloszlást úgy, hogy azon színezéseknek legyen nagyobb a valószínűsége, amelyben több szomszédos területet színeztünk azonos színűre. Erre az eloszlásra a bayesi statisztikai módszerek során apriori eloszlásként fogunk tekinteni. A Potts-modellt a második részben nézzük meg.

Egy lényeges térbeli statisztikai probléma a betegségtérképezés (disease mapping). Ilyenkor a területeinken adott egy bizonyos betegség által bekövetkezett halálesetek száma, és szeretnénk olyan térképet készíteni, amely hitelesen ábrázolja az egyes területeken a halálozás kockázatát. A becslést nehezíti, hogy az egyes területek lakosságszáma nagyon eltérő lehet, így bizonyos helyeken megbízhatóbb, más helyeken jóval kevésbé megbízható becsléseket tehetünk. A harmadik részben háromféle modellt nézünk meg, amelyek a probléma kezelésére szolgálnak.

Dolgozatunk alapját képezi Green és Richardson [1] cikke, amely szintén a betegségtérképezés problémájával foglalkozik. A cikkben leírt módszer a területeket a halálozás kockázata szerint osztályokba sorolja, és a lehetséges osztályokba sorolások apriori eloszlásaként a Potts-modellt használja. A becslés ezután MCMC módszer használatával történik. A cikket a negyedik részben ismertetjük.

Az ötödik részben megnézzük, hogyan lehet eljárni abban az esetben, ha a betegségtérképezéstől eltérően a területeken nem gyakorisági, hanem más jellegű adatokkal rendelkezünk. Például meg szeretnénk becsülni az autóbalesetekből származó kár várható nagyságát, ha rendelkezésünkre állnak az egyes területeken a bekövetkezett károk nagyságai. Green és Richardson modelljét fogjuk adaptálni három különböző problémához, és részletesen ismertetjük a modellek felépítését és az MCMC módszer alkalmazását. Az ötödik rész tartalmazza a saját eredményeket, az első négy rész témáit pedig a szakirodalom alapján mutatjuk be.

1. Az MCMC módszer

Brooks [2] cikkét felhasználva megnézzük a bayesi statisztika fontos módszerét, az MCMC eljárást. A bayesi statisztikai vizsgálatok során gyakran van szükség bonyolult, többdimenziós integrálok kiszámítására. Legyen például az y minta x paraméterre vett feltételes eloszlása $L(y|x)$. Ekkor a paraméter

$\pi(x|y)$ aposteriori eloszlását konstans szorzótól eltekintve megkaphatjuk

$$\pi(x|y) \propto L(y|x)p(x)$$

alakban, ahol $p(x)$ az x paraméter apriori eloszlása, és a normalizáló konstans az

$$\int L(y|x)p(x)dx$$

képlettel kapható meg. Ha szeretnénk meghatározni a paraméter valamely $\theta(x)$ függvényének aposteriori várható értékét, akkor az

$$E(\theta(x)|y) = \int \theta(x)\pi(x|y)dx$$

integrált kellene kiszámolnunk. Hasonlóképpen, ha az x többdimenziós paraméter valamelyik komponensének marginális aposteriori eloszlását szeretnénk megkapni, akkor a $\pi(x|y)$ aposteriori eloszlást integrálni kellene a többi változó szerint. Sok esetben sem az aposteriori eloszlás normalizáló konstansát nem tudjuk pontosan meghatározni, sem az említett két feladatban fellépő integrálokat nem tudjuk kiszámolni, ezért közelítő módszerre van szükség. Az egyik ilyen módszer az MCMC (Markov Chain Monte Carlo) eljárás.

Az MCMC eljárás során az aposteriori eloszlásból mintát veszünk, és ebből a mintából becsüljük a szóban forgó mennyiségeket, vagyis lényegében a fenti integrálok értékét becsüljük meg közvetett úton. Az eljárás lényege, hogy egy olyan Markov-láncot készítünk, amelynek stacionárius eloszlása éppen a $\pi(x|y)$ aposteriori eloszlás, amelyet ezután csak $\pi(x)$-szel jelölünk, és céleloszlásnak nevezünk. Ezt az a tétel alapozza meg, hogy minden aperiodikus, irreducibilis Markov-láncnak egyértelműen létezik stacionárius eloszlása, és a lánc ehhez az eloszláshoz konvergál. Ha tehát egy adott kiinduló állapotból elég sokáig futtatjuk a láncot, akkor a kapott értékek a π eloszlásból vett mintának tekinthetők. Jelöljük innentől a Markov-lánc magfüggvényét $K(x, y)$-nal, ez tehát a lánc következő tagjának a feltételes

sűrűségfüggvénye az y helyen, arra a feltételre nézve, hogy a lánc most az x állapotban van. Ekkor tehát π stacionaritása azt jelenti, hogy ha $x \sim \pi(x)$ és $y \sim K(x, y)$, akkor egyben $y \sim \pi(y)$ is teljesül. Az MCMC eljárás az állapotvektor módosítási szabálya, amely lehet egy magfüggvény, de lehet több magfüggvény egymás utáni vagy kevert végrehajtása is. Most áttekintjük az MCMC két alapvető módszerét, a Gibbs-mintavételt és a Metropolis-Hastings-algoritmust.

Gibbs-mintavétel

A Gibbs-mintavétel esetén az x állapotvektor komponensekre osztott: $x = (x_1, \ldots, x_k) \in \mathbb{R}^p$, a komponenseket pedig egymás után módosítjuk. Az x_i komponens módosítását úgy végezzük, hogy az összes többi komponensre vett feltételes eloszlása szerint kisorsolunk egy új értéket, és x_i-t ezzel helyettesítjük. Vagyis az új x_i' értékre $x_i' \sim x_i|\pi(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$. Ezt elvégezzük sorban $i = 1$-től k-ig, az i-edik lépésnél a $j < i$ indexű komponenseknél már a módosított értéket használva a feltételes eloszlás kiszámításánál. Ebből a k lépésből áll össze a Gibbs-mintavétel, amelynek magfüggvényét tehát a következő alakban írhatjuk fel:

$$K(x^t, x^{t+1}) = \prod_{i=1}^{k} \pi(x_i^{t+1}|x_1^{t+1}, \ldots, x_{i-1}^{t+1}, x_{i+1}^{t}, \ldots, x_k^{t})$$

A Gibbs-mintavétel során tehát a feltételes eloszlás szerinti véletlen változókat kell kisorsolni. Gyakran a feltételes eloszlás valamilyen jól ismert eloszlás, ekkor ez nem okoz nehézséget, de léteznek hatékony módszerek nem standard eloszlásból való mintavételre is.

Metropolis-Hastings-algoritmus

A Metropolis-Hastings-algoritmus esetén is az állapotvektor új értékét egy eloszlásból sorosoljuk ki, de ezt az új értéket csak egy bizonyos valószínűség-

gel fogadjuk el, ha pedig elutasítjuk, akkor az új állapotvektor megegyezik a régivel. Legyen az x vektor az előzőhöz hasonlóan komponensekre osztva, és tegyük fel, hogy az i-edik komponenst akarjuk módosítani. Először megadunk egy $q(x, y)$ függvényt, ami minden x esetén sűrűségfüggvény y-ban, ebből az eloszlásból generáljuk a javasolt új állapotot. Itt az x és y vektoroknak az i-ediktől különböző komponensei rögzítettek az $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$ értékekkel, és csak az i-edik komponens változik. Az x állapotvektorhoz tehát kisorsolunk a $q(x, y)$ sűrűségfüggvény szerint egy új y vektort, ennek elfogadási valószínűsége pedig legyen

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\} \tag{1}$$

Az y vektor kisorsolása után tehát $\alpha(x, y)$ valószínűséggel lecseréljük x-et y-ra, különben pedig megtartjuk x-et. Írjuk fel az algoritmus magfüggvényét, legyen $P(x, A)$ annak valószínűsége, hogy a lánc az x állapotból az A halmazba kerül. Ekkor

$$P(x, A) = \int_A K(x, y)dy + r(x)I_A(x)$$

teljesül, ahol

$$K(x, y) = q(x, y)\alpha(x, y)$$

az eloszlásnak az a része, amelyik az y vektor elfogadásához tartozik,

$$r(x) = \int q(x, y)(1 - \alpha(x, y))dy$$

pedig az elutasításhoz tartozó pontmérték. Látható, hogy az algoritmus során elég a π céleloszlást csak normalizáló tényezőtől eltekintve ismerni, mert az elfogadási hányadosban ez a tényező kiesik. Ha a $q(x, y)$ sűrűségfüggvényt szimmetrikusan választjuk, vagyis $q(x, y) = q(y, x)$ teljesül, akkor

az (1) elfogadási valószínűség az

$$\alpha(x,y) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\} \tag{2}$$

hányadosra egyszerűsödik.

A $q(x,y)$ függvény megválasztására számos lehetőség adódik. Egy speciális eset a véletlen bolyongású Metropolis-algoritmus, ekkor $q(x,y) = f(x-y)$ valamely f sűrűségfüggvényre, ilyenkor a javasolt y állapotra $y = x + z$ teljesül, ahol z-t az f sűrűségfüggvény szerint sorsoljuk. Egy másik lehetőség, hogy a $q(x,y)$ függvény nem függ x-től: $q(x,y) = f(y)$ egy f sűrűségfüggvényre. A Gibbs-mintavétel is felírható a Metropolis-Hastings-algoritmus speciális eseteként. Belátható, hogy az i-edik komponens Gibbs-mintavétel szerinti módosítása megfelel a Metropolis-Hastings-algoritmusnak a $q(x,y) = \pi(y_i|x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_k)$ választással.

Dimenzióváltó algoritmus

Előfordulhat, hogy az MCMC eljárásban a paraméterek száma nem állandó, hanem maga is egy paraméter. Ilyenek például az általunk vizsgált keverék modellek, amikor a keverék komponenseinek száma ismeretlen. Ilyenkor a Markov-lánc nem marad ugyanabban a p dimenziós térben, hanem különböző dimenziójú terekbe léphet. Az ilyen eljárásokat nevezzük dimenzióváltó MCMC algoritmusoknak, aminek általunk használt fajtáját Green [3] vezette be Reversible Jump MCMC (RJMCMC) néven, ami a Metropolis-Hastings mintavétel kiterjesztése. Legyen a C állapottér $C_1, C_2, \cdots \subseteq C$ részekre felosztva, ahol a $C_i \subseteq \mathbb{R}^{n_i}$ részek különböző dimenziós valós terek részhalmazai. A $\pi(x)$ céleloszlás helyett ekkor egy $\pi(dx)$ célmértéket kell tekintenünk. A különböző dimenziós terek közti átmenethez meg kell adnunk a módosítási lépéseknek egy $m = 1, 2, \ldots$-vel paraméterezett családját. Legyen az előzőekhez hasonlóan $q_m(x, dy)$ egy x-hez tartozó mérték, amelyből a javasolt új y állapotot kisorsoljuk. Ha az x állapotvektor például a C_1 tar-

tományban van, akkor először kisorsolunk egy m típusú lépést adott x-től függő valószínűségekkel, ezután a $q_m(x, dy)$ mérték alapján egy új y pontot, amely valamely C_2 tartományba esik. Gyakran nem alkalmazható az összes módosító lépés a C_1 tartománybeli állapotvektorokra, az általunk később vizsgált esetben például a dimenzió csak eggyel nőhet, illetve csökkenhet. Ilyenkor az x állapotban csak azokat a lépéseket sorsoljuk ki pozitív valószínűséggel, amelyek alkalmazhatók x-re. Miután a javasolt y állapotot meghatároztuk, ennek elfogadási valószínűsége

$$\min\left\{1, \frac{\pi(dy)q_m(y, dx)}{\pi(dx)q_m(x, dy)}\right\} \tag{3}$$

Ez a valószínűség jól definiált, hogyha a $q_m(x, dy)$ függvények teljesítenek egy dimenziókra vonatkozó feltételt, ami lényegében azt biztosítja, hogy az x-ről y-ra és az y-ról x-re való áttérés szabadságfokai megegyezzenek.

A dimenzióváltó lépést a gyakorlatban úgy végezhetjük el, hogy ha adott az x állapotvektor, és azt az m típusú lépést sorsoltuk ki, ami egy magasabb dimenziós y állapotvektort határoz meg, akkor sorsoljunk ki egy folytonos eloszlású véletlen u vektort és legyen y az x és u vektorok valamilyen invertálható függvénye, $y(x, u)$. A fordított irányú lépés pedig ennek az inverze, vagyis az y magasabb dimenziós vektorhoz meghatározzuk az (x, u) vektort, és x lesz a javasolt új állapotvektor. Ekkor a (3) elfogadási valószínűség a következőképpen írható fel:

$$\min\left\{1, \frac{p(y)r_m(y)}{p(x)r_m(x)q(u)}\left|\frac{\partial y}{\partial(x, u)}\right|\right\} \tag{4}$$

ahol $r_m(x)$ annak a valószínűsége, hogy az x állapotban az m típusú lépést választjuk, p az aposteriori sűrűségfüggvény, $\frac{\partial y}{\partial(x,u)}$ pedig az (x, u) változókról az y változóra való áttérés Jacobi-mátrixa. A dimenziókra vonatkozó feltétel ekkor azt követeli meg, hogy x és u dimenzióinak összege egyezzen meg y dimenziójával, így lehetséges, hogy az (x, u) és y közti áttérés bijektív és

deriválható.

2. A Potts-modell

A következőkben Francois és mások [4] cikke alapján áttekintjük az ún. Potts-modellt. Ha a statisztikai adataink térbeli területekhez vagy koordinátákhoz kötődnek, és a mögöttes gyakoriságok, kockázatok vagy más statisztikai paraméterek térbeli összefüggéseit szeretnénk vizsgálni, akkor célszerű először elkészíteni a megfigyelések szomszédsági gráfját. Ez természetes módon megtehető, ha már adott valamilyen területi beosztás, például ha egy ország egyes megyéire vonatkoznak az adataink. Ilyenkor két megfigyelést akkor tekintünk szomszédosnak, ha a területeik a valóságban szomszédosak. Ha a megfigyelések csak a koordinátáikkal vannak megadva a síkon, akkor nincs ilyen kézenfekvő megoldás. Megtehetjük, hogy elkészítjük a pontok Voronoj-diagramját. Ez minden egyes ponthoz hozzárendeli azt a konvex sokszöget a síkon (a pont Voronoj-celláját), amelynek pontjai ehhez az adott ponthoz vannak a legközelebb a megadott pontok közül. Ezután két megfigyelést akkor tekintünk szomszédosnak, ha a Voronoj-celláik szomszédosak. Ennél egyszerűbb eljárás, ha kijelölünk egy adott távolságot, és akkor tekintünk két pontot szomszédosnak, ha a távolságuk a megadottnál kisebb.

Ezután feltételezzük, hogy a gráf minden csúcsa k osztály valamelyikébe sorolódik. Legyen n csúcsa a gráfnak, és $z_i \in \{1, \ldots, k\}$ az i-edik csúcs osztálya. Jelölje $U(z)$ a gráf olyan éleinek számát, amelyek azonos osztályba sorolt pontokat kötnek össze, képlettel

$$U(z) = \sum_{i \sim j} \delta_{z_i, z_j}$$

ahol δ a Kronecker-delta. Ekkor azon színezésekhez, amelyekben nagy kiterjedésű azonos színű területek vannak, nagyobb $U(z)$ érték tartozik, a kevés térbeli összefüggést mutató színezésekhez pedig kisebb. A Potts-modell egy

valószínűségeloszlás a gráf lehetséges színezésein. Egy $\underline{z}$ színezéshez tartozó valószínűség:

$$p(z) \propto e^{\psi U(z)} \quad \underline{z} \in \{1, \ldots, k\}^n$$

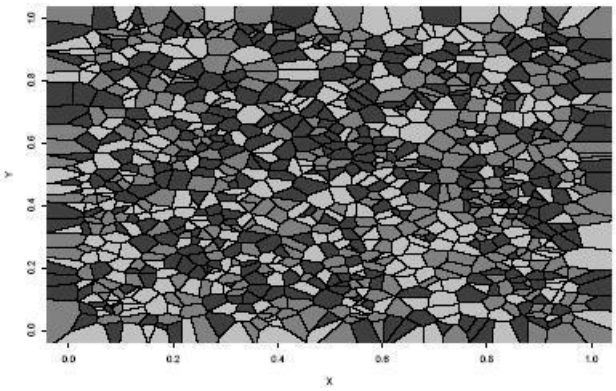
ahol a $\propto$ szimbólum jelentése: multiplikatív konstanstól eltekintve. Itt $\psi \geq 0$ a kölcsönhatási paraméter. $\psi = 0$ annak az esetnek felel meg, amikor nincs térbeli összefüggés, vagyis a csúcsok színe egymástól független. Minél nagyobb ψ értéke, annál valószínűbb, hogy kiterjedt területeket színezünk ki azonos színnel. A gyakorlat azt mutatja, hogy viszonylag kevés számú osztály esetén $\psi \leq 0,4$ gyenge, alig észrevehető térbeli összefüggést eredményez, $\psi \geq 1$ -nél pedig már annak a valószínűsége, hogy két szomszédos csúcs azonos színnel van színezve, közel 1, ilyenkor gyakran az egész térkép azonos színű. Az 1. ábrán látható két minta a Potts-modellből különböző ψ értékek esetén, ahol megfigyelhetjük, hogy magasabb ψ érték nagyobb térbeli összefüggést jelent a színezésben. Megjegyzendő, hogy a Potts-modell az ismertebb ún. Ising-modell általánosítása, ami tetszőleges számú színosztály helyett kettőt használ.

A Potts-modell egyik előnye, hogy rendelkezik a térbeli Markov-tulajdonsággal, vagyis egy csúcs színének feltételes eloszlása a többi csúcsra megegyezik a szomszédos csúcsaira vett feltételes eloszlással:

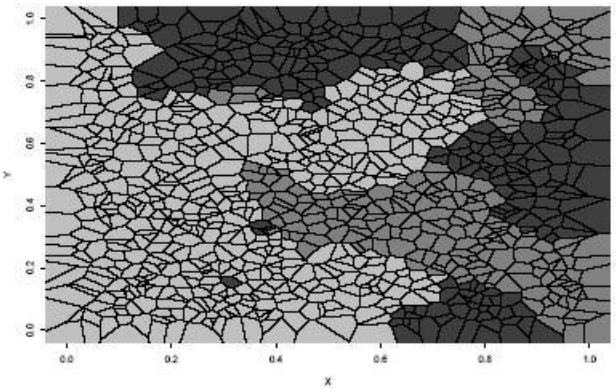
$$p(z_i|z_{(i)}) = p(z_i|z_{\partial z_i}) \quad (i = 1, \ldots, n)$$

ahol $z_{(i)}$ az i-től különböző csúcsok halmaza, $z_{\partial z_i}$ pedig a z_i szomszédainak halmaza. Másképpen mondva egy csúcs színezése feltételesen független a többi csúcs színezésétől a szomszédaira nézve. Az ilyen tulajdonsággal rendelkező eloszlásokat Markov-mezőknek nevezzük, a Markov-láncok analógiájára, ahol egy változó eloszlását az időbeli "szomszédja" határozza meg. A térbeli Markov-tulajdonság miatt az egyes csúcsokra vonatkozó számításoknál nem kell az egész színezést figyelembe vennünk, hanem elég a csúcs környezetét tekinteni. Ezért ez a tulajdonság lehetővé teszi, hogy gyors és

egyszerűen kiszámítható algoritmusokat használjunk.



(a) $\psi = 0.1$



(b) $\psi = 0.9$

1. ábra. *A Potts-modell egy-egy előfordulása* $k = 3$ *és (a)* $\psi = 0.1$ *(b)* $\psi = 0.9$ *esetén. Az első esetben nem látható térbeli összefüggés, a másodikban nagy összefüggő területek vannak azonos színnel színezve. ([4]-ből)*

A normalizáló konstans kiszámítása

A Potts-modell fenti megadásánál az egyes színezések valószínűségeit csak egymáshoz viszonyítva, normalizáló tényezőtől eltekintve definiáltuk. Ez bizonyos esetekben, például a Metropolis-Hastings módszernél elegendő, de máskor ismernünk kell a pontos valószínűséget. A k, ψ értékekhez tartozó

normalizáló konstans (fizikus szóhasználattal partíciófüggvény) logaritmusát jelöljük $\theta_k(\psi)$-vel:

$$\theta_k(\psi) = \log \sum_{z \in \{1,\dots,k\}^n} e^{\psi U(z)} \tag{5}$$

Ekkor egy z színezés valószínűsége:

$$p(z) = e^{\psi U(z) - \theta_k(\psi)}$$

A $\theta_k(\psi)$ konstans fenti felírásából látható, hogy még viszonylag kis méretű és kevés színosztállyal rendelkező modellek esetén is olyan sok tagból áll, hogy pontos értékének nyilvántartása egy algoritmus folyamán nem kivitelezhető, ezért közelítő eljárásra van szükség. Számos ilyen eljárás létezik, ezek közül az egyik legegyszerűbbet tekintjük most át.

Deriváljuk ψ szerint (5)-öt:

$$\frac{\partial}{\partial \psi}\theta_k(\psi) = \frac{\partial}{\partial \psi} \log \sum_z e^{\psi U(z)} = \sum_z U(z) e^{\psi U(z) - \theta_k(\psi)} =$$

$$= \sum_z U(z) p_{k,\psi}(z) = E_{k,\psi}(U(z))$$

ahol az összegzés az összes lehetséges színezésen fut végig, az indexben lévő k, ψ pedig azt jelenti, hogy z eloszlását a megadott k és ψ értékek mellett tekintjük. Láttuk, hogy ha $\psi = 0$, akkor az egyes z_i értékek egymástól függetlenek és azonos eloszlásúak az $\{1, \dots, k\}$ halmazon, így minden színezés egyforma valószínűségű, tehát $\theta_k(0) = \log k^n = n \log k$. Ebből integrálással

$$\theta_k(\psi) = n \log k + \int_0^\psi E_{k,\psi'}(U(z)) d\psi'$$

Ezután az $E_{k,\psi}(U(z))$ várható értékeket megbecsülhetjük MCMC mintavétellel egyes diszkrét ψ értékekre, például a $\{0, \delta, 2\delta, \dots, \psi_{max}\}$ halmaz elemeire, vagyis a Potts-modell eloszlása szerinti színezéseket generálunk MCMC mód-

szerrel, mindegyik színezésre kiszámítjuk $U(z)$ értékét, és ezeket átlagoljuk. Ezen diszkrét ψ-kre vonatkozó becslésekből a fenti integrál értékét például harmadfokú spline illesztésével becsülhetjük.

3. Betegségtérképezés

Clayton és Kaldor [5] cikke alapján megvizsgáljuk a betegségtérképezés problémáját és az általuk leírt két modellt, majd egy harmadik, népszerű modellt is megnézünk. A betegségek térképezésének célja olyan térképek készítése, amelyek egy betegség helyi kockázatát jelenítik meg egy adott országon vagy régión belüli kisebb területekre. Ehhez minden területen rendelkezésre áll a betegségben való elhalálozások megfigyelt száma és várható száma, ami az adott terület lakosságszámával arányos. Ezek hányadosát nevezzük standardizált halandósági hányadosnak (SMR), ami a kockázat tapasztalati becslése. A térképezés egyszerű megközelítése a területenkénti SMR ábrázolása, ez azonban általában félrevezető. A területek lakosságszámában ugyanis jelentős eltérések lehetnek, és az alacsony lélekszámú területeken a becslés esetleg csupán néhány előforduláson alapszik, ami nagyon pontatlan lehet. Így a térképet a ritkán lakott területeken tapasztalt extrém értékek uralják. Egy másik, szintén nem kielégítő lehetőség, hogy az egyes területeken a teljes átlagtól való eltérés statisztikai szignifikanciaszintjét ábrázoljuk. Ebben az esetben az azonos SMR-rel, de más lakosságszámmal rendelkező területeket egészen másképpen jelöljük meg a térképen, az extrém értékeket pedig általában a nagy lélekszámú területek adják, ami szintén félrevezető lehet.

Legyen N darab területünk, az i-edik területen O_i a megfigyelt, E_i a várt halálozások száma, ahol E_i ismert érték. Tegyük fel, hogy O_i Poisson-eloszlású $\theta_i E_i$ paraméterrel, θ_i a becsülendő paraméter. A következőkben a θ vektornak valamilyen $f(\theta)$ apriori eloszlást választunk, és θ_i-t az aposteriori várható értékével becsüljük.

A gamma modell

A gamma modellben a θ_i-knek független $Gamma(\alpha, \nu)$ apriori eloszlást választunk, az α és ν paramétereket később megbecsüljük. Ilyenkor az O_i megfigyelések feltétel nélküli eloszlása negatív binomiális, és könnyen láthatóan

$$E(O_i) = \frac{E_i \nu}{\alpha}$$
$$D^2(O_i) = \frac{E_i \nu}{\alpha} + \frac{E_i^2 \nu}{\alpha^2}$$

A θ_i aposteriori eloszlása $Gamma(E_i + \alpha, O_i + \nu)$, amiből θ_i aposteriori várható értéke:

$$E(\theta_i | O_i) = \frac{O_i + \nu}{E_i + \alpha}$$

Ebbe behelyettesítve α és ν becslését kapjuk a θ_i-re vonatkozó becslést:

$$\hat{\theta}_i = \frac{O_i + \hat{\nu}}{E_i + \hat{\alpha}} \tag{6}$$

Az α és ν paraméterek becslését maximum-likelihood módszerrel végezzük. Az O_i megfigyelések feltétel nélküli eloszlásából kapjuk, hogy a loglikelihood-függvény a következő:

$$L(\alpha, \nu) = \sum_{i=1}^{N} \left(\log \frac{\Gamma(O_i + \nu)}{\Gamma(\nu)} + \nu \log(\alpha) - (O_i + \nu) \log(E_i + \alpha) \right)$$

A paraméterek becslését ezután deriválással, majd a keletkező egyenletrendszer közelítő megoldásával kapjuk.

Látható, hogy θ_i (6) becslése kompromisszum az O_i/E_i SMR és a ν/α között, ami a θ_i eloszlásának várható értéke. A becslés így nagy megfigyelésszámmal rendelkező területeken az SMR-hez van közel, kis megfigyelésszám esetén pedig az összesített kockázathoz.

A lognormális modell

A lognormális modell esetén azt feltételezzük, hogy a θ_i kockázatok apriori eloszlása lognormális, vagyis a $\beta_i = \log(\theta_i)$ változók eloszlása valamilyen többdimenziós normális eloszlás μ várható értékkel és Σ szórásmátrixszal. Ilyenkor a θ_i-k aposteriori várható értéke nem kapható meg zárt alakban, ezért közelítésre van szükség. Megmutatható, hogy ha a β vektor O_i mintaelemekre vonatkozó $\psi(\beta)$ likelihood-függvényét négyzetes függvénnyel közelítjük, akkor a β aposteriori eloszlása szintén többdimenziós normális eloszlás. Ehhez a $\psi(\beta)$ függvényt sorbafejtjük a β^* vektor körül, ahol

$$\beta_i^* = \log\left(\frac{O_i + \frac{1}{2}}{E_i}\right) \tag{7}$$

ami lényegében a koordinátánkénti ML-becslés azzal a módosítással, hogy az $O_i = 0$ eset kezelésére a megfigyelésszámhoz hozzáadtunk 1/2-et. A μ és Σ paraméterek ML-becslésére az ún. EM algoritmust használják a szerzők, ami egy rejtett változók becslésére szolgáló közelítő algoritmus.

A β vektor apriori eloszlásaként szolgáló többdimenziós normális eloszlás megválasztására több lehetőség is van. A legegyszerűbb esetben a koordinátákat független azonos $N(m, \sigma^2)$ eloszlásúra választjuk. Ilyenkor az EM algoritmus a következő becslést adja β_i-re:

$$\hat{\beta}_i = \frac{\hat{m} + \left(O_i + \frac{1}{2}\right)\hat{\sigma}^2\beta_i^* - \frac{\hat{\sigma}^2}{2}}{1 + \left(O_i + \frac{1}{2}\right)\hat{\sigma}^2}$$

ahol $\hat{m}$ és $\hat{\sigma}^2$ az algoritmusból származó közelítő becslések. Láthatóan ez a becslés is kompromisszum a β_i^* lokális becslés (7) és az $\hat{m}$ teljes mintára vonatkozó becslés között. Az O_i megfigyelés nagy értékeire az előbbihez, kis értékeire az utóbbihoz van közel β_i becslése.

Az apriori eloszlást úgy is megválaszthatjuk, hogy a földrajzi elhelyezkedést is figyelembe vesszük. Legyen W a területek szomszédsági mátrixa, és

vegyük a következő feltételes autoregressziós modellt:

$$E(\beta_i|\beta_j, j \neq i) = \mu_i + \rho \sum_{j=1}^{N} W_{ij}(\beta_j - \mu_j)$$

$$D^2(\beta_i|\beta_j, j \neq i) = \sigma^2$$

Megmutatható, hogy ekkor $E(\beta) = \mu$ és $\Sigma(\beta) = \sigma^2(I - \rho W)^{-1}$. A paraméterek becslését nem részletezzük.

A BYM modell

Gyakran használt módszer a betegségtérképezésben az ún. BYM modell, amelyet Green és Richardson [1] cikke alapján mutatunk be. A BYM modell is lognormális modell, vagyis a λ_i kockázatok logaritmusainak együttes apriori eloszlását többdimenziós normális eloszlásnak választjuk, amit a szomszédsági gráf határoz meg. Legyen az i-edik terület relativ kockázata $\lambda_0 e^{u_i+v_i}$, ahol a λ_0, u_i, v_i változók feltételesen függetlenek a $\alpha, \beta, \tau_u, \tau_v$ változókra nézve, és apriori eloszlásuk

$$\lambda_0 \sim Gamma(\alpha, \beta)$$

$$p(u|\tau_u) \propto exp(-\frac{1}{2}\tau_u \sum_{i \sim i'} (u_i - u_{i'})^2) \tag{8}$$

$$p(v|\tau_v) \propto exp(-\frac{1}{2}\tau_v \sum_{i=1}^{n} v_i^2) \tag{9}$$

ahol a szomszédsági gráf minden C összefüggőségi komponensére

$$\sum_{i \in C} u_i = \sum_{i \in C} v_i = 0$$

teljesül. (8)-ból és (9)-ből könnyen ellenőrizhető, hogy a kockázatok logaritmusainak feltételes eloszlása valóban normális. Látható, hogy ha τ_u és τ_v végtelenhez tartanak, akkor a becslésünk minden területen ugyanaz, ez

az összeöntött mintából való becslésnek felel meg, ha pedig 0-hoz tartanak, akkor az egyes területekre függetlenül becslünk. Itt feltételezhetjük a $\tau_u \sim Gamma(\alpha_u, \beta_u)$ és $\tau_v \sim Gamma(\alpha_v, \beta_v)$ apriori eloszlásokat és az $\alpha_u = \alpha_v = \beta_u = \beta_v = 0.1$ választást. Ezután a becslést MCMC módszer alkalmazásával végezzük.

4. Rejtett Markov-modell

Green és Richardson cikkében ritka betegségek előfordulásait vizsgálja Franciaországban. Az ország összes megyéjében rendelkezésre áll a megbetegedések száma egy adott évben, és ebből szeretnénk a mögöttes kockázatokra következtetni. Az előzőekhez hasonlóan itt is Poisson-eloszlást tételezünk fel a megfigyelt esetek számára:

$$y_i \sim Poisson(\lambda_i E_i) \quad (i = 1, \ldots, n)$$

ahol y_i a megfigyelések száma, E_i a várt esetek száma, és λ_i a becsülendő kockázat. Az előző cikktől eltérően ebben az esetben teljesen bayesi megközelítést fogunk használni, ahol minden paraméternek apriori eloszlást adunk meg, és a becslést MCMC módszerrel végezzük. A cikk módszerének kiindulópontja az a feltételezés, hogy a λ_i kockázatok egy rejtett Markov-mező, esetünkben a Potts-modell előfordulását alkotják, innen a modell neve. Ahhoz, hogy a Potts-modellt alkalmazhassuk, a kockázatokat diszkretizálnunk kell, vagyis meg kell adnunk k darab lehetséges $\lambda_1, \ldots, \lambda_k$ értéket, amelyből az egyes kockázatok az értéküket felveszik. Ekkor tehát az i-edik területen lévő kockázat λ_{z_i} alakban írható, ahol z_i a k lehetséges osztály valamelyikében van: $z_i \in \{1, \ldots, k\}$. A z_i allokációs változókra k színosztályú Potts-modellt tételezünk fel apriori eloszlásként, ahol a ψ kölcsönhatási paramétert is változónak tekintjük, egyenletes apriori eloszlással a $\{0, 0.1, \ldots, \psi_{max}\}$ halmazon. Az osztályok k száma szintén változó, a $\{1, \ldots, k_{max}\}$ halmazon egyenletes eloszlással.

A $\lambda_1, \ldots, \lambda_k$ változók apriori eloszlásának független $Gamma(\alpha, \beta)$ eloszlást választunk. Az α, β paramétereket itt rögzítettnek tekintjük az $\alpha = 1$ és

$$\beta = \frac{\sum_{i=1}^{n} E_i}{\sum_{i=1}^{n} y_i} \tag{10}$$

értékekkel, de más választás is lehetséges. A szerzők megvizsgálták az α paraméter értékének más megválasztását (az $\alpha/\beta = \sum_i y_i / \sum_i E_i$ egyenlőség fenntartásával), és azt is, hogy a β paraméternek is apriori eloszlást adnak, de ezek nem változtattak lényegesen az eredményeken. Föltesszük még, hogy a $\lambda_1, \ldots, \lambda_k$ paraméterek növekvő sorrendbe vannak állítva, így tehát a λ vektor apriori sűrűségfüggvénye:

$$p(\lambda|k) = k! I(\lambda_1 < \cdots < \lambda_k) \prod_{j=1}^{k} \frac{\beta^\alpha \lambda_j^{\alpha-1} e^{-\beta\lambda_j}}{\Gamma(\alpha)}$$

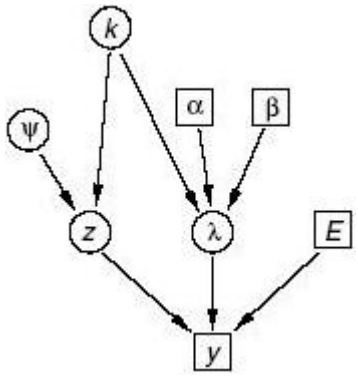
Ezzel megadtuk a modellt, amelyben az összes változó együttes sűrűségfüggvénye tehát a következő alakban írható fel:

$$p(k, \psi, \lambda, z, y) = p(k)p(\psi)p(z|k, \psi)p(\lambda|k)p(y|\lambda, z)$$

A változók apriori eloszlásai közti összefüggéseket a 2. ábra mutatja.

Az MCMC algoritmus lépései

A kockázatok becsléséhez tehát MCMC módszert használunk, amelynek során a k, ψ, z, λ paraméterek értékeit módosítjuk. Ez három fix dimenziós és egy dimenzióváltó (RJMCMC) lépéssel történik meg. Először a z allokációs változók értékeit módosítjuk. Ez Gibbs-mintavétellel történik, vagyis a változókat egymás után módosítjuk, és z_i módosításakor az új értéket a többi változó rögzített értéke melletti feltételes eloszlásából sorsoljuk ki. Látni fogjuk, hogy a térbeli Markov-tulajdonság miatt a módosításnál elegendő z_i szomszédait figyelembe venni.



2. ábra. *A változók összefüggéseit ábrázoló irányított körmentes gráf. Az X változóból az Y-ba nyíl mutat, ha X szerepel az Y változó apriori eloszlásának felírásában. A négyzettel jelölt változók értéke rögzített, a körrel jelölteknél MCMC becslésre van szükség. ([1]-ből)*

A ψ kölcsönhatási együtthatót véletlen bolyongású Metropolis-algoritmussal módosítjuk, ± 0.1-es ugrásokkal pozitív és negatív irányba egyforma valószínűséggel. Ha a javasolt új ψ' így kilép az értékkészletéből, akkor ott a sűrűségfüggvényt 0-nak tekintjük és a módosítást elutasítjuk.

A λ vektort Metropolis-Hastings-algoritmussal módosítjuk úgy, hogy a λ_j-k logaritmusaihoz független normális eloszlású változókat adunk hozzá, és az így keletkezett vektort újrarendezzük, ez lesz a javasolt új vektor.

Az osztályok k számának módosítását RJMCMC algritmussal végezzük. Először kiválasztjuk, hogy a dimenziót eggyel csökkentjük vagy növeljük, ezt b_k illetve d_k valószínűséggel tesszük. Ezután kisorsoljuk, hogy melyik két szomszédos λ_j-t vonjuk össze, illetve melyiket osztjuk ketté. Amikor a dimenziók számát növeljük, akkor a kettéosztandó λ_j-ből létrehozzuk az új $\lambda_- = \lambda_j u^- c$ és $\lambda_+ = \lambda_j u^c$ változókat, ahol $u \sim U(0,1)$ és $c = 0.1$. Ha az így kapott új λ vektor nem rendezett, akkor a módosítást elvetjük. Különben az eddig a j osztályba tartozó területeket kettéosztjuk az új $+$ és $-$ osztályokba. Ezt úgy végezzük, hogy a területeket egymás után soroljukbe az új osztályokba úgy, hogy annak a valószínűsége, hogy z_i-t a $-$ osztályba

soroljuk,

$$\frac{e^{\psi\nu_- - \lambda_- E_i}\lambda_-^{y_i}}{e^{\psi\nu_- - \lambda_- E_i}\lambda_-^{y_i} + e^{\psi\nu_+ - \lambda_+ E_i}\lambda_+^{y_i}} \tag{11}$$

ahol ν_- és ν_+ az i-edik terület azon szomszédainak száma, amelyeket már korábban a $-$ és $+$ osztályokba soroltunk. Ennek a besorolásnak a célja, hogy növelje az elfogadás valószínűségét, a besorolást függetlenül is végezhettük volna az egyes területekre. Ezzel tehát megkaptuk a javasolt új állapotot. A dimenzió csökkentése ennek a lépésnek a megfordítása: véletlenszerűen kiválasztjuk, hogy melyik két szomszédos λ_j, λ_{j+1}-et vonjuk össze, és az új λ'_j a két érték mértani közepe lesz. A két régi osztályba sorolt területek allokációs változóit pedig összevonjuk, így kapjuk a javasolt állapotot, aminek elfogadásáról a Metropolis-Hastings tört alapján döntünk.

A fenti MCMC lépéseket a későbbiekben részletesebben tárgyaljuk.

5. Változatok a rejtett Markov-modellre

5.1. Exponenciális megfigyelések

A most következőkben Green és Richardson modelljének alkalmazását mutatjuk be néhány változatban. Az előzőekben a megfigyeléseink bizonyos eseményeknek a számára vonatkoztak, ezért tételeztük fel, hogy Poisson-eloszlásúak. Elképzelhető azonban, hogy az adataink más jellegűek, például vonatkozhatnak a megfigyeléseink az egyes területek autóbaleseteiben fellépő kár nagyságára. Ebben az esetben feltételezhetjük például, hogy a kár nagysága exponenciális eloszlású. Természetesen ilyenkor egy területen több megfigyelésünk is lehet.

Tegyük fel, hogy a megfigyelések n területre esnek, mindegyik területen n_i darab. A megfigyelések a többi paraméter rögzített értékére nézve feltételesen függetlenek, és exponenciális eloszlásúak λ_{z_i} paraméterrel:

$$y_{il} \sim Exp(\lambda_{z_i}) \quad (i = 1, \ldots, n), (l = 1, \ldots, n_i)$$

Minden területen a λ_{z_i} paramétert szeretnénk megbecsülni.

A z_i allokációs változók apriori eloszlásaként továbbra is a Potts-modellt használjuk, a kockázatok lehetséges $\lambda_1, \ldots, \lambda_k$ értékeinek apriori eloszlása az előzőekhez hasonlóan egy független Gamma-eloszlású vektor koordinátáinak növekvő sorrendbe rendezésével áll elő. A Gamma-eloszlás paramétereinek megválasztásakor az eredeti cikkhez hasonlóan járunk el, vagyis az α paraméternek valamilyen rögzített értéket adunk, például $\alpha = 1$-et, ezután β-t úgy választjuk meg, hogy a Gamma-eloszlás várható értéke, α/β megegyezzen a λ paraméter valamilyen egyszerű becslésével. Tegyük fel tehát, hogy az egyes területekhez tartozó λ paraméterek megegyeznek, ekkor az y_{il} megfigyelések függetlenek és azonos $Exp(\lambda)$ eloszlásúak. Ebből λ ML-becslése az átlag reciproka, vagyis β-t úgy választjuk, hogy teljesüljön

$$\frac{\alpha}{\beta} = \frac{\sum_{i=1}^{n} n_i}{\sum_{i=1}^{n} \sum_{l=1}^{n_i} y_{il}} \tag{12}$$

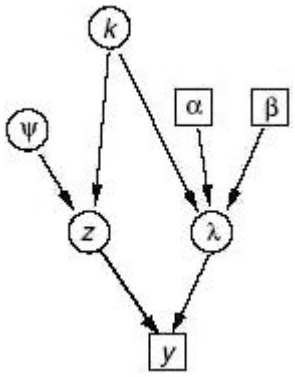
A k és ψ paraméterek apriori eloszlásán nem változtatunk. A változók együttes sűrűségfüggvénye tehát:

$$p(k)p(\psi)p(z|k,\psi)p(\lambda|k)p(y|z,\lambda) = p(k)p(\psi)e^{\psi U(z) - \theta_k(\psi)} \times$$
$$\times k! I(\lambda_1 < \cdots < \lambda_k) \left(\prod_{j=1}^{k} \frac{\beta^\alpha \lambda_j^{\alpha-1} e^{-\beta\lambda_j}}{\Gamma(\alpha)} \right) \times \prod_{i=1}^{n} \prod_{l=1}^{n_i} \lambda_{z_i} e^{-\lambda_{z_i} y_{il}} \tag{13}$$

A változók közti összefüggéseket a 3. ábrán tüntettük fel. Ezután nézzük az MCMC eljárás lépéseit.

Fix dimenziós lépések

A z_i allokációs változók értékeinek módosítását a Gibbs-módszerrel végezzük, vagyis a z_i értékeket egyesével egymás után módosítjuk úgy, hogy az új értéküket az összes többi változóra vett feltételes eloszlásuk szerint sorsoljuk ki. z_i aposteriori eloszlásának meghatározásakor a normalizáló tényezőtől



3. ábra. *Az 5.1 szakasz változóinak összefüggéseit ábrázoló gráf*

eltekintünk. Ekkor tehát a (13) együttes sűrűségfüggvényben a z_i-től különböző változók értékét rögzítettnek vesszük, és csak a z_i-től függő tagokat tartjuk meg. Az $U(z)$ összegben tehát csak a gráf azon éleit vesszük figyelembe, amelyek z_i-be vezetnek:

$$p(z_i = j|\dots) \propto e^{\psi U(z)} \prod_{l=1}^{n_i} (\lambda_j e^{-\lambda_j y_{il}}) \propto \lambda_j^{n_i} \exp\left(\psi \nu_{ij} - \lambda_j \sum_{l=1}^{n_i} y_{il}\right)$$

ahol ν_{ij} az i olyan szomszédainak száma, amelyek értéke j. Mivel z_i fenti feltételes eloszlása csak véges sok, k darab pontra koncentrált, ezért nem okoz problémát az, hogy a normalizáló tényezőt nem ismerjük, mert a fenti képletet j-re összegezve ez megkapható. Ezután z_i új értékének kisorsolása nem okoz nehézséget.

A ψ paramétert Metropolis-Hastings módszerrel módosítjuk. A q átmenetfüggvény legyen olyan, hogy az x pontból $\frac{1}{2} - \frac{1}{2}$ valószínűséggel lépünk az $x + 0,1$ és $x - 0,1$ pontokba, tehát $q(x, y) = \frac{1}{2}$, ha $|y - x| = 0,1$, és 0 különben. Látható, hogy ez egy véletlen bolyongásos Metropolis-módszer, ahol q szimmetrikus, tehát (2) alkalmazható. Ha $\psi' = -0,1$ vagy $\psi_{max} + 0,1$, akkor a változtatást elvetjük, mert ezeken a helyeken ψ apriori sűrűségfüggvénye

0. Különben az elfogadási valószínűség tehát:

$$\min\left\{1, e^{(\psi'-\psi)U(z)-(\theta_k(\psi')-\theta_k(\psi))}\right\}$$

A λ_j-k módosítását is Metropolis-Hastings módszerrel végezzük, egyszerre változtatva meg az egész λ vektort. Ehhez meg kell adni a javasolt λ' vektor sűrűségfüggvényét az eredeti λ függvényében, tehát a $Q(\lambda, \lambda')$ átmenetfüggvényt. A javasolt λ'-t úgy állítjuk elő, hogy minden $\log \lambda_j$-hez független, 0 várható értékű, adott σ^2 szórásnégyzetű normális változókat adunk hozzá, és az így kapott változókat újra növekvő sorrenbe rendezve kapjuk a λ'_j értékeket. Ennek a változtatásnak az elfogadásáról a Metropolis-Hastings tört alapján döntünk, vagyis (1) alapján az elfogadás valószínűsége

$$min\left\{1, \frac{P(\lambda')Q(\lambda', \lambda)}{P(\lambda)Q(\lambda, \lambda')}\right\}$$

ahol P az aposteriori sűrűségfüggvény. Láthatjuk, hogy a fenti eljárást az újrarendezéstől eltekintve a koordinátákon függetlenül végezzük. A Q függvény kiszámításához legyen $q(x, y)$ az egy koordinátára vonatkozó átmenetfüggvény, ezt a következő alakban írhatjuk fel:

$$q(x, y) = f_{xe^{N(0,\sigma^2)}}(y) = \frac{1}{y} f_{N(0,\sigma^2)}\left(\log \frac{y}{x}\right)$$

amiből következik

$$q(x, y) = \frac{x}{y} q(y, x) \tag{14}$$

Ha az újrarendezéstől eltekintünk, akkor egy adott rendezetlen λ' vektorban a $Q(\lambda, \lambda')$ függvény értéke $\prod_{j=1}^{k} q(\lambda'_j, \lambda_j)$ lenne. A rendezés miatt azonban nem különböztetjük meg, hogy a λ koordinátái a λ' mely koordinátáiba mentek át, vagyis $Q(\lambda, \lambda')$ valódi értéke rendezetlen λ'-kre 0, rendezett λ'-kre pedig egy $k!$ tagú összeg, amelynek tagjai a lehetséges permutációkhoz tartozó sűrűségfüggvény-értékek. Az alábbi számolásból látható, hogy a $\frac{Q(\lambda', \lambda)}{Q(\lambda, \lambda')}$

törtben szerencsére nem marad meg a $k!$ számú tag, mert lehet egyszerűsíteni:

$$Q(\lambda',\lambda) = \sum_{\pi \text{ permutáció}} \prod_{j=1}^{k} q(\lambda'_j, \lambda_{\pi(j)}) = \sum_{\pi} \prod_{j=1}^{k} \frac{\lambda'_j}{\lambda_{\pi(j)}} q(\lambda_{\pi(j)}, \lambda'_j) =$$
$$= \frac{\prod_{j=1}^{k} \lambda'_j}{\prod_{j=1}^{k} \lambda_j} \sum_{\pi} \prod_{j=1}^{k} q(\lambda_{\pi(j)}, \lambda'_j) = \left(\prod_{j=1}^{k} \frac{\lambda'_j}{\lambda_j} \right) \sum_{\pi} \prod_{j=1}^{k} q(\lambda_j, \lambda'_{\pi^{-1}(j)}) =$$
$$= \left(\prod_{j=1}^{k} \frac{\lambda'_j}{\lambda_j} \right) \sum_{\pi} \prod_{j=1}^{k} q(\lambda_j, \lambda'_{\pi(j)}) = \left(\prod_{j=1}^{k} \frac{\lambda'_j}{\lambda_j} \right) Q(\lambda,\lambda')$$

Tehát a tört értéke:

$$\frac{Q(\lambda',\lambda)}{Q(\lambda,\lambda')} = \prod_{j=1}^{k} \frac{\lambda'_j}{\lambda_j} \tag{15}$$

A számolás során először kihasználtuk (14)-et, majd azt, hogy a λ_j-k szorzata nem függ az összeszorzás sorrendjétől, végül a π permutációról az inverz permutációra tértünk át, és eszerint összegeztünk.

A $\frac{P(\lambda')}{P(\lambda)}$ tört értékét a (13) együttes sűrűségfüggvény λ' és λ helyen felvett értékeinek a hányadosából kapjuk meg. A hányadost a koordináták szerint bontjuk szorzatra, λ_j és λ'_j azon területekhez tartozó tényezőkben szerepel, amelyek a j-edik osztályba tartoznak.

$$\frac{P(\lambda')}{P(\lambda)} = \prod_{j=1}^{k} \left(\left(\frac{\lambda'_j}{\lambda_j} \right)^{\alpha - 1 + \sum_{i:z_i=j} n_i} \exp\left(-(\lambda'_j - \lambda_j)(\beta + \sum_{i:z_i=j} \sum_{l=1}^{n_i} y_{il} \right) \right) \tag{16}$$

A javasolt λ' vektor elfogadásának valószínűsége tehát $\min\{1, R\}$, ahol R

értéke (15) és (16) szerint:

$$R = \frac{P(\lambda')Q(\lambda',\lambda)}{P(\lambda)Q(\lambda,\lambda')} =$$
$$= \prod_{j=1}^{k} \left(\left(\frac{\lambda'_j}{\lambda_j} \right)^{\alpha-1+\sum_{i:z_i=j} n_i} \exp\left(-(\lambda'_j - \lambda_j)(\beta \sum_{i:z_i=j} \sum_{l=1}^{n_i} y_{il} \right) \right) \prod_{j=1}^{k} \frac{\lambda'_j}{\lambda_j} =$$
$$= \prod_{j=1}^{k} \left(\left(\frac{\lambda'_j}{\lambda_j} \right)^{\alpha+\sum_{i:z_i=j} n_i} \exp\left(-(\lambda'_j - \lambda_j)(\beta \sum_{i:z_i=j} \sum_{l=1}^{n_i} y_{il} \right) \right)$$

Dimenzióváltó lépés

A különböző összetevők k számának módosításához az előző részben ismertetett RJMCMC eljárást használjuk. Először tehát b_k és d_k valószínűségekkel kiválasztjuk, hogy dimenziócsökkentő vagy -növelő lépést végzünk, majd növelés esetén a kettéosztandó λ_j változó helyére a $\lambda_- = \lambda_j u^{-c}$ és $\lambda_+ = \lambda_j u^c$ új változókat vesszük fel, és ha a λ vektor rendezett maradt, akkor a j-edik osztályba tartozó területeket a fent leírt módon a két új osztályba soroljuk. Ekkor tehát annak a valószínűsége, hogy az i-edik területet a $-$ osztályba soroljuk:

$$\frac{\lambda_-^{n_i} \exp\left(\psi\nu_- - \lambda_- \sum_{l=1}^{n_i} y_{il}\right)}{\lambda_-^{n_i} \exp\left(\psi\nu_- - \lambda_- \sum_{l=1}^{n_i} y_{il}\right) + \lambda_+^{n_i} \exp\left(\psi\nu_+ - \lambda_+ \sum_{l=1}^{n_i} y_{il}\right)}$$

amit (11) mintájára készítettünk el, vagyis a z_i változó új értékét a többi változó rögzítése mellett a $\{+,-\}$ halmazra vett feltételes eloszlásából sorsoljuk ki.

A fordított lépésnél az összvonandó λ_j-ket a mértani közepükkel helyettesítjük, és a z_i változókat megfelelően módosítjuk.

A szétválasztás elfogadási valószínűségét az RJMCMC módszer leírásában szereplő (4) képletből számolhatjuk ki. Legyen az elfogadás valószínűsége

$min\{1, R\}$. Az R-ben szereplő tagok közül a Jacobi-determináns:

$$\left|\frac{\partial y}{\partial(x,u)}\right| = \left|\frac{\partial(\lambda_j u^c, \lambda_j u^{-c})}{\partial(\lambda_j, u)}\right| = \left|\det\begin{pmatrix} u^c & \lambda_j c u^{c-1} \\ u^{-c} & -\lambda_j c u^{-c-1} \end{pmatrix}\right| = \frac{2c\lambda_j}{u}$$

R meghatározásakor szükség van még a kisorsolt új allokáció valószínűségére, amit jelöljünk P_{allok}-vel. Ez tehát úgy áll elő, hogy összeszorozzuk a valószínűségeket, amelyekkel z_i-t a $+$ vagy $-$ osztályba soroljuk. R maradék részét az együttes sűrűségfüggvény régi és új helyen felvett értékének hányadosa adja, tehát R értéke a következő:

$$\begin{aligned} R = &\prod_{i:z_i'=-} \left(\frac{\lambda_-}{\lambda_j}\right)^{n_i} \exp\left(-(\lambda_- - \lambda_j)\sum_{l=1}^{n} y_{il}\right) \times \\ &\times \prod_{i:z_i'=+} \left(\frac{\lambda_+}{\lambda_j}\right)^{n_i} \exp\left(-(\lambda_+ - \lambda_j)\sum_{l=1}^{n} y_{il}\right) \frac{\beta^\alpha \left(\frac{\lambda_-\lambda_+}{\lambda_j}\right)^{\alpha-1} e^{-\beta(\lambda_- + \lambda_+ - \lambda_j)}}{\Gamma(\alpha)} \times \\ &\times (k+1)\frac{p(k+1)}{p(k)} e^{\psi(U(z')-U(z))+\theta_k(\psi)-\theta_{k+1}(\psi)} \frac{d_{k+1}}{b_k P_{allok}} \frac{2c\lambda_j}{u} \end{aligned} \tag{17}$$

Az összevonás elfogadási valószínűsége ebből már egyszerűen kiszámítható. Korábban láthattuk, hogy a Metropolis-Hastings tört értéke abban az esetben, ha az x pontban a javasolt y állapotra írjuk fel, reciprokára változik, ha az y pontban írjuk fel az x javasolt állapotra. Így tehát kiszámoljuk az összevonás során kapott javasolt állapotból kiindulva a fenti R értéket, a két állapot szerepét felcserélve, majd az elfogadás valószínűsége $min\{1, R^{-1}\}$ lesz.

5.2. A megfigyélésszám és -nagyság együttes becslése

Az előző részben csak a megfigyelések nagyságával foglalkoztunk, a számukat viszont figyelmen kívül hagytuk, Green és Richardson cikkében pedig csak a

megfigyelések számát vettük figyelembe. Elképzelhető, hogy a megfigyelések várható gyakoriságát és nagyságát egyszerre akarjuk megbecsülni. A megfigyelések számát továbbra is Poisson eloszlásúnak, nagyságát exponenciális eloszlásúnak tételezzük fel. Ilyenkor tehát minden területen két paramétert kell megbecsülnünk. Azt tételezzük fel, hogy a két paraméterhez közös színezés tartozik a Potts-modellben, tehát

$$n_i \sim Poisson(\mu_{z_i} E_i) \quad (i = 1, \ldots, n)$$

$$y_{il} \sim Exp(\lambda_{z_i}) \quad (i = 1, \ldots, n), (l = 1, \ldots, n_i)$$

ahol $z_i \in \{1, \ldots, k\}$ a közös allokációs változó, $\mu_1, \ldots, \mu_k$ és $\lambda_1, \ldots, \lambda_k$ pedig a lehetséges értékek, E_i az előfordulások várható száma. Legyen a λ_j paraméterek apriori eloszlása $Gamma(\alpha_1, \beta_1)$, a μ_j-k eloszlása $Gamma(\alpha_2, \beta_2)$, és tegyük fel, hogy a (λ_j, μ_j) párok az első változó szerint rendezettek, vagyis

$$p(\lambda|k) = k! I(\lambda_1 < \cdots < \lambda_k) \prod_{j=1}^{k} \frac{\beta_1^{\alpha_1} \lambda_j^{\alpha_1 - 1} e^{-\beta_1 \lambda_j}}{\Gamma(\alpha_1)}$$

$$p(\mu|k) = \prod_{j=1}^{k} \frac{\beta_2^{\alpha_2} \mu_j^{\alpha_2 - 1} e^{-\beta_2 \mu_j}}{\Gamma(\alpha_2)}$$

A β_1 és β_2 hiperparaméterek értékét ismét a λ és μ paraméterek egyszerű becslése alapján állítjuk be. Tegyük fel tehát, hogy λ és μ értéke területenként nem változik. Ekkor a két paraméter becslését függetlenül végezhetjük, tehát β_1 értékét (12)-nek megfelelően határozzuk meg:

$$\frac{\alpha_1}{\beta_1} = \frac{\sum_{i=1}^{n} n_i}{\sum_{i=1}^{n} \sum_{l=1}^{n_i} y_{il}}$$

A μ becslésénél adottak az $n_i \sim Poisson(\mu E_i)$ független mintaelemek. A likelihood-függvény:

$$L(\mu) = \prod_{i=1}^{n} \frac{(\mu E_i)^{n_i}}{n_i!} e^{-\mu E_i} = \left(\prod_{i=1}^{n} \frac{E_i^{n_i}}{n_i!} \right) \mu^{\sum_i n_i} e^{-\mu \sum_i E_i}$$

amiből μ ML-becslése és így β_2 értéke a következő:

$$\frac{\alpha_2}{\beta_2} = \hat{\mu} = \frac{\sum_i n_i}{\sum_i E_i}$$

ami megegyezik a Green és Richardson cikkében látott (10) választással.

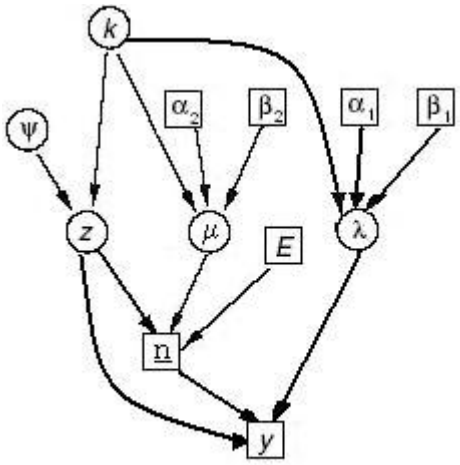
A z, ψ és k változók apriori eloszlásán az előzőekhez képest nem változtatunk. Ekkor az együttes sűrűségfüggvényt a következőképpen írhatjuk fel:

$$\begin{aligned} p(k, \psi, z, \lambda, \mu, \underline{n}, y) &= p(k)p(\psi)p(z|k,\psi)p(\lambda|k)p(\mu|k)p(\underline{n}|\mu, z)p(y|\lambda, z) = \\ &= p(k)p(\psi)e^{\psi U(z) - \theta_k(\psi)} \times k! I(\lambda_1 < \cdots < \lambda_k) \prod_{j=1}^{k} \frac{\beta_1^{\alpha_1} \lambda_j^{\alpha_1 - 1} e^{-\beta_1 \lambda_j}}{\Gamma(\alpha_1)} \times \\ &\times \prod_{j=1}^{k} \frac{\beta^{\alpha_2} \mu_j^{\alpha_2 - 1} e^{-\beta_2 \mu_j}}{\Gamma(\alpha_2)} \times \prod_{i=1}^{n} \frac{(\mu_{z_i} E_i)^{n_i}}{n_i!} e^{-\mu_{z_i} E_i} \times \prod_{i=1}^{n} \prod_{l=1}^{n_i} \lambda_{z_i} e^{-\lambda_{z_i} y_{il}} \end{aligned} \tag{18}$$

ahol az n_i megfigyelések vektorát $\underline{n}$-nel jelöltük, hogy a területek számától megkülönböztessük. A 4. ábra mutatja a változókhoz tartozó gráfot.

Fix dimenziós lépések

Az MCMC eljárás során négy fix dimenziós lépést hajtunk végre: a ψ, z, λ és μ változókat módosítjuk. A ψ paraméter módosítása az előző részben leírtakkal egyezik meg, tehát véletlen bolyongású Metropolis-módszert használunk, és az elfogadási valószínűség sem változik. A z_i allokációs változók módosításához ki kell számolnunk z_i többi változóra vett feltételes eloszlását.



4. ábra. *Az 5.2 szakasz változóinak összefüggéseit ábrázoló gráf*

A (18) együttes sűrűségfüggvényben z_i három helyen jelenik meg: a $p(z|k,\psi)$, a $p(\underline{n}|\mu,z)$ és a $p(y|\lambda,z)$ tagokban. Ebből a feltételes eloszlás a következő:

$$p(z_i = j|\dots) \propto e^{\psi U(z)} \frac{(\mu_j E_i)^{n_i}}{n_i!} e^{-\mu_j E_i} \prod_{l=1}^{n_i} (\lambda_j e^{-\lambda_j y_{il}}) \propto$$

$$\propto (\lambda_j \mu_j)^{n_i} \exp\left(\psi \nu_{ij} - \mu_j E_i - \lambda_j \sum_{l=1}^{n_i} y_{il}\right)$$

Ebből z_i új értékét már egyszerűen kisorsolhatjuk.

A λ vektor módosítása mindenben megegyezik az előző részben leírtakkal. Látható ugyanis, hogy az együttes sűrűségfüggvényben az előző részhez képest szereplő két új tagban λ nem szerepel, ezért az elfogadási valószínűség nem változik, és a rendezésre vonatkozó érvelés is átvihető.

A μ vektort a λ-hoz hasonlóan módosítjuk, vagyis független, 0 várható értékű, azonos szórásnégyzetű normális változókat adunk hozzá $\log \mu_j$-hez. Mivel ebben az esetben a rendezéssel nem kell foglalkoznunk, egyszerűen adódik a Metropolis-Hastings-tört átmenetfüggvényeinek hányadosára (15)

alapján

$$\frac{Q(\mu',\mu)}{Q(\mu,\mu')} = \prod_{j=1}^{k} \frac{\mu'_j}{\mu_j}$$

A sűrűségfüggvények hányadosa pedig:

$$\frac{P(\mu')}{P(\mu)} = \prod_{j=1}^{k} \left(\left(\frac{\mu'_j}{\mu_j} \right)^{\alpha_1 - 1 + \sum_{i:z_i=j} y_i} \exp\left(-(\mu'_j - \mu_j)(\beta_1 + \sum_{i:z_i=j} E_i) \right) \right)$$

Ebből tehát az elfogadási valószínűség:

$$min\left\{ 1, \prod_{j=1}^{k} \left(\left(\frac{\mu'_j}{\mu_j} \right)^{\alpha_1 + \sum_{i:z_i=j} y_i} \exp\left(-(\mu'_j - \mu_j)(\beta_1 + \sum_{i:z_i=j} E_i) \right) \right) \right\}$$

Dimenzióváltó lépés

A dimenzióváltó lépést az előzőek mintájára végezzük, a szétválasztó lépést nézzük meg részletesen. Itt egy területhez két paraméter is tartozik, ezért ha a j-edik osztályt akarjuk kettéosztani, akkor λ_j és μ_j helyére is két-két új értéket kell választani. Ehhez sorsoljuk ki a független $u, v \sim U(0,1)$ változókat, és legyen $\lambda_+ = \lambda_j u^c$, $\lambda_- = \lambda_j u^{-c}$, $\mu_+ = \mu_j v^c$ és $\mu_- = \mu_j v^{-c}$. A j-edik osztályhoz tartozó (λ_j, μ_j) pár helyett az új $-$ és $+$ osztályhoz a (λ_-, μ_-) és (λ_+, μ_+) párok tartoznak.

Ezután a j-edik osztályhoz tartozó változókat módosítjuk egymás után, az előzőekben leírtak szerint. A z_i változó feltételes eloszlásából a $-$ és $+$ osztályba sorolás p_- és p_+ relatív valószínűségeire a következőket kapjuk (11) mintájára:

$$p_- = (\lambda_- \mu_-)^{n_i} \exp\left(\psi\nu_- - \mu_- E_i - \lambda_- \sum_{l=1}^{n_i} y_{il} \right)$$

$$p_+ = (\lambda_+ \mu_+)^{n_i} \exp\left(\psi\nu_+ - \mu_+ E_i - \lambda_+ \sum_{l=1}^{n_i} y_{il} \right)$$

Ebből a tényleges valószínűségeket a két szám összegével való osztással kapjuk meg. Az új értékek beállításánál a valószínűségeket összeszorozva meghatározzuk az új allokáció P_{allok} valószínűségét.

Ezután az elfogadás valószínűségét hasonlóképpen meghatározhatjuk az együttes sűrűségfüggvény hányadosából az eredeti és a javasolt állapotban, a $(\lambda_j, \mu_j, u, v) \mapsto (\lambda_+, \lambda_-, \mu_-, \mu_+)$ leképezés Jacobi-determinánsából és az átmenetvalószínűségekből. A valószínűség $\min\{1, R\}$, ahol

$$
\begin{aligned}
R = & \prod_{i:z_i'=-} \left(\frac{\lambda_-\mu_-}{\lambda_j\mu_j}\right)^{n_i} \exp\left(-(\lambda_- - \lambda_j)\sum_{l=1}^{n} y_{il} - (\mu_- - \mu_j)E_i\right) \times \\
& \times \prod_{i:z_i'=+} \left(\frac{\lambda_+\mu_+}{\lambda_j\mu_j}\right)^{n_i} \exp\left(-(\lambda_+ - \lambda_j)\sum_{l=1}^{n} y_{il} - (\mu_+ - \mu_j)E_i\right) \times \\
& \times \frac{\beta_1^{\alpha_1}\left(\frac{\lambda_-\lambda_+}{\lambda_j}\right)^{\alpha_1-1} e^{-\beta_1(\lambda_- + \lambda_+ - \lambda_j)}}{\Gamma(\alpha_1)} \times \frac{\beta_2^{\alpha_2}\left(\frac{\mu_-\mu_+}{\mu_j}\right)^{\alpha_2-1} e^{-\beta_2(\mu_- + \mu_+ - \mu_j)}}{\Gamma(\alpha_2)} \times \\
& \times (k+1)\frac{p(k+1)}{p(k)} e^{\psi(U(z')-U(z))+\theta_k(\psi)-\theta_{k+1}(\psi)} \frac{d_{k+1}}{b_k P_{allok}} \frac{2c\lambda_j}{u} \frac{2c\mu_j}{v}
\end{aligned}
$$

5.3. Becslés ismeretlen megfigyelésszám esetén

Az előző részben láttuk, hogy ha az egyes területeken a megfigyelések száma és nagysága is ismert, akkor a λ és μ paraméterek becslését majdnem függetlenül végezhettük, hiszen a két paramétert csak a Potts-modell közös színezése kötötte össze. Az is látható, hogy nem volt szükséges egy területen az összes megfigyelés nagyságát ismerni, hanem elég volt az összegüket tudni, és ebből már minden lépés elvégezhető volt. Ennek oka, hogy exponenciális eloszlású megfigyelések esetén az összeg elégséges statisztika. Felmerülhet tehát a kérdés, hogy hogyan végezzük a becslést abban az esetben, ha az egyes területeken nem ismerjük a megfigyelések számát, hanem csak az összegüket. Látni fogjuk, hogy a probléma visszavezethető az előző részben tárgyalt módszerre.

Ekkor tehát minden területen egy y_i megfigyelésünk van, amely ismeretlen

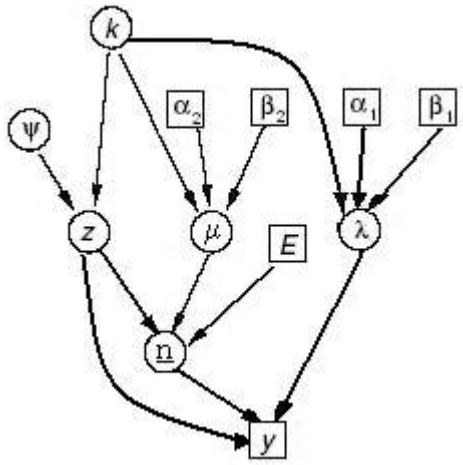
számú független, exponenciális eloszlású változó összege, tehát összetett Poisson-eloszlású $\mu_{z_i} E_i$ intenzitással és $Exp(\lambda_{z_i})$ összegzett eloszlással. Sajnos y_i sűrűségfüggvénye nem írható fel zárt alakban, ezért az előző részben felírt MCMC lépések nem végezhetők el közvetlenül. Ennek megoldása érdekében vezessük be az n_i megfigyelésszámokat új változóként, amelyre immár nem megfigyelésként, hanem paraméterként tekintünk, apriori eloszlásnak pedig $\mu_{z_i} E_i$ paraméterű Poisson-eloszlást választunk. Ekkor y_i feltételes eloszlása n_i darab független $Exp(\lambda_{z_i})$ eloszlású változó összege, ami Gamma-eloszlás abban az esetben, ha $n_i \geq 1$:

$$p(y_i|\lambda_{z_i}, n_i) = \frac{\lambda_{z_i}^{n_i} y_i^{n_i-1} e^{-\lambda_{z_i} y_i}}{(n_i - 1)!}$$

és $P(y_i = 0|n_i = 0) = 1$.

A többi változó apriori eloszlását úgy választjuk meg, mint az előző részben. Látható, hogy így az összes változó eloszlása ugyanaz maradt, mint az előző részben, csak annyi a különbség, hogy a sűrűségfüggvényt ezúttal a megfigyelések összegével fejeztük ki. Az 5. ábrán látható gráf csak annyiban tér el az előző részben, a 4. ábrán mutatott gráfhoz képest, hogy az $\underline{n}$ megfigyelésszámokat ezúttal körrel jeleztük, mert most az értéküket becsülnünk kell. Ezért az MCMC lépések változatlanok, csak az egy területre eső megfigyelések összegét kell helyettesíteni y_i-vel, ahol ez szükséges. Az előző részben leírtakhoz képest két különbség van. Az egyik, hogy a β_1 és β_2 hiperparaméterek értékét más módon kell megválasztanunk, mint az előbb, mert az előző felírásban az n_i értékek szerepeltek. A másik, hogy az n_i paramétereket is módosítanunk kell az MCMC eljárás során, mert az értékük most nem ismert.

A λ és μ változók apriori eloszlásában szereplő β_1 és β_2 hiperparamétereket tehát úgy választjuk, hogy a Gamma-eloszlás várható értéke megegyezzen λ és μ valamilyen egyszerű becslésével. Ehhez a becsléshez ismét feltesszük, hogy a λ és μ paraméterek értéke minden területen ugyanaz, ekkor tehát



5. ábra. *Az 5.3 szakasz változóinak összefüggéseit ábrázoló gráf*

$y_1, \ldots, y_n$ függetlenek, és y_i összetett Poisson-eloszlású μE_i várható értékkel és $Exp(\lambda)$ összegzett eloszlással. A ML-becslés ezúttal nem végezhető el közvetlenül, ezért egy egyszerűbb lehetőséget nézünk meg. Jelölje M az átlag, S_*^2 a korrigált tapasztalati szórásnégyzet függvényt. A becslést a momentum-módszerhez hasonlóan végezzük, a λ és μ becslését úgy választjuk, hogy az a következő két egyenlőséget teljesítse:

$$E_{\lambda,\mu}\left(M(Y_1,\ldots,Y_n)\right) = M(y_1,\ldots,y_n) \tag{19}$$

$$E_{\lambda,\mu}\left(S_*^2\left(\frac{Y_1}{E_1},\ldots,\frac{Y_n}{E_n}\right)\right) = S_*^2\left(\frac{y_1}{E_1},\ldots,\frac{y_n}{E_n}\right) \tag{20}$$

ahol az Y_i-k az y_i-vel megegyező eloszlású változókat jelölnek, y_i-k pedig maguk a megfigyelt értékek. (20)-ban azért osztottuk le a megfigyeléseket E_i-vel, hogy a tagok várható értéke megegyezzen. Jelöljük a fenti egyenletek jobb oldalát röviden M-mel és S_*^2-tel. Az összetett Poisson-eloszlás várható értékére és szórásnégyzetére vonatkozó formulákat felhasználva $E_{\lambda,\mu}(y_i) = \mu E_i \lambda^{-1}$ és $D^2_{\lambda,\mu}(y_i) = 2\mu E_i \lambda^{-2}$. Ebből (19) és (20) alapján a következő

egyenleteket kapjuk:

$$\frac{\left(\sum_i E_i\right)\mu}{n\lambda} = M$$
$$\frac{\left(\sum_i E_i^{-1}\right)2\mu}{n\lambda^2} = S_*^2$$

Ebből a becsléseket, és így β_1 és β_2 értékét így írhatjuk fel:

$$\frac{\alpha_1}{\beta_1} = \hat{\lambda} = \frac{\left(\sum_i E_i^{-1}\right)2M}{\left(\sum_i E_i\right)S_*^2}$$
$$\frac{\alpha_2}{\beta_2} = \hat{\mu} = \frac{n\left(\sum_i E_i^{-1}\right)2M^2}{\left(\sum_i E_i\right)^2 S_*^2}$$

Itt a nevező 1 valószínűséggel pozitív, ezért a becslés elvégezhető.

Az n_i paraméterek módosítását Gibbs-módszerrel végezzük. Az együttes sűrűségfüggvényben n_i az apriori eloszlásában és az y_i megfigyelések eloszlásában szerepel. Ha $y_i = 0$, akkor n_i aposteriori eloszlása a 0-ra koncentrált, ekkor $n_i = 0$-t választunk 1 valószínűséggel. Különben a feltételes eloszlás:

$$p(n_i = m|\dots) \propto \frac{(\lambda_{z_i}\mu_{z_i}E_i y_i)^m}{m!(m-1)!} \quad (m \geq 1)$$

Az n_i paraméter új értékét tehát ebből a nemstandard eloszlásból sorsoljuk ki.

Összefoglalás

A dolgozat során megismerkedtünk a bayesi statisztikai módszerek alkalmazásával bizonyos térbeli statisztikai problémák kezelésére. Először áttekintettük az MCMC módszer alapjait, ami egy közelítő algoritmus összetett rendszerekben történő becslések elvégzésére. Ezután bemutattunk egy fizikából származó modellt, a Potts-modellt, ami egy jól alkalmazható modell térbeli összefüggőségek kezelésére. Láttuk, hogy a Potts-modellel az MCMC

algoritmus futása során könnyen tudtunk számolni, elsősorban térbeli Markov-tulajdonsága miatt. A harmadik részben leírtunk egy jellegzetes térbeli statisztikai problémát, a betegségtérképezést, majd röviden megnéztünk néhány, a mi módszerünktől eltérő megoldást, köztük a népszerű BYM modellt. Ezután következett a dolgozatunk alapjául szolgáló modell, a Green és Richardson cikkében leírt rejtett Markov-modell, amelyben felhasználtuk a Potts-modellt, és a becsléseket MCMC módszerrel végeztük.

Az ötödik részben a rejtett Markov-modellt alkalmaztuk a betegségtérképezéstől eltérő problémák megoldására. Először megnéztük az exponenciális eloszlású megfigyelések esetét, amelyre káreloszlásként is tekinthetünk. Ezután azt vizsgáltuk meg, miként lehet a megfigyelések gyakoriságát és nagyságát együtt megbecsülni. Végül egy olyan esetet néztünk meg, amikor az egyes területeken csak a megfigyelések összege ismert, így kevesebb információból kellett a paramétereket megbecsülnünk. Eközben részletesen ismertettük az MCMC algoritmus lépéseinek meghatározását. A dolgozatban tehát bemutattuk, milyen széles körben alkalmazható az MCMC algoritmus a térbeli statisztikában felmerülő becslések elvégzésére.

Köszönetnyilvánítás

Köszönöm Arató Miklósnak az érdekes témajavaslatot és a dolgozat elkészítése során nyújtott értékes segítségét.

Irodalomjegyzék

[1] P. J. Green, S. Richardson, *Hidden Markov models and disease mapping.* Journal of the American Statistical Association, 2002.

[2] S. P. Brooks, *Markov chain Monte Carlo method and its application.* The Statistician, 1998.

[3] P. J. Green, *Reversible Jump MCMC computation and Bayesian model determination.* Biometrika, 1995.

[4] O. Francois, S. Ancelet, G. Guillot, *Bayesian clustering using hidden Markov random fields in spatial population genetics.* Genetics, 2006.

[5] D. G. Clayton, J. M. Kaldor, *Empirical Bayes estimates of age-standardized relative risks for use in disease mapping.* Biometrics, 1987.