COMPUTATIONAL SCIENCE AND TECHNIQUES
Volume 1, Number 2, 2013, 222-228
© Klaipėda University, 2013
Publisher: Klaipėda University
http://journals.ku.lt/index.php/CST
Online ISSN: 2029-9966

DAUGIAMATIS RETŲ ĮVYKIŲ TIKIMYBIŲ VERTINIMO ALGORITMAS

Leonidas Sakalauskas, IngridaVaičiulytė
Vilniaus universitetas, Matematikos ir informatikos institutas,
Akademijos g. 4, LT-08663, Vilnius
sakal@ktl.mii.lt, ingrida_vaiciulyte@yahoo.com

Anotacija. Šiame darbe sudarytas Monte–Karlo Markovo grandinių algoritmas daugiamačių retų įvykių dažniams vertinti. Retų įvykių tikimybės logitai modeliuojami Puasono skirstiniu, kurio parametrai pasiskirstę pagal daugiamatį normalųjį dėsnį su nežinomais vidurkių vektoriumi ir kovariacijų matrica. Nežinomų parametrų įverčiai gaunami didžiausio tikėtinumo metodu. Išvestos lygtys, kurias turi tenkinti modelio didžiausio tikėtinumo parametrų įverčiai. Vertinamos kovariacijų matricos teigiamas apibrėžtumas kontroliuojamas apskaičiuojant matricos didžiausios ir mažiausios tikrinių reikšmių santykį.

Pagrindiniai žodžiai: Monte–Karlo Markovo grandinė, Bajeso metodas, tikėtinumo metodas, Puasono–Gauso modelis.

Įvadas

Bajeso metodas plačiai taikomas įvairiems sprendimo priėmimo uždaviniams spręsti bei verslo ar finansinių rodiklių vertinimui (Bradley ir Thomas, 2000, Clayton ir Kaldor, 1987, Chen, 2009; Liseo ir Loperfido, 2003; Tsutakawa ir kt. 1985). Monte–Karlo Markovo grandinės (Markov Chain Monte Carlo, MCMC) yra kompiuterinio imitavimo būdas dažnai taikomas nežinomiems parametrams įvertinti. Šiame darbe pasinaudojant Puasono–Gauso modeliu konstruojamas Monte–Karlo Markovo sekų algoritmas, skirtas kelių retų įvykių tikimybėms vertinti. Pavyzdžiui, gali būti nagrinėjami draudiminiai įvykiai (skirtingų automobilių rūšių avarijų skaičius), susirgimų ar mirties atvejai populiacijoje. Aposteriorinis įvykių dažnių skirstinys yra sudaromas Bajeso metodu. Darbe nagrinėjamas logit modelis, kuriame nepriklausomas kintamasis α išreiškiamas per nagrinėjamo įvykio tikimybę P:

$$\alpha = \ln \frac{P}{1-P}, \qquad (1)$$

čia P yra tikimybė, kad priklausomas kintamasis įgys reikšmę 1, o $(1-P)$ – tikimybė, kad priklausomas kintamasis įgys reikšmę 0 (Altaleb ir Chauveau, 2002; Pearce, 2006). Vieno įvykio vertinimo algoritmas panaudojant Bajeso metodą buvo sudarytas Bradley ir Thomas (2000), Clayton ir Kaldor (1987), Tsutakawa ir kt. (1985).

1. Puasono–Gauso modelis

Tegul turime generalinę imtį $\Theta = (\theta_1, \theta_2, \ldots, \theta_K)$, sudarytą iš K populiacijų ir kiekviena populiacija θ_j turi N_j individų, $j = \overline{1,K}$. Tarkime, kad stebėjimo metu populiacijoje gali įvykti tam tikri nepalankūs įvykiai (draudiminio įvykio, susirgimo, mirties,...). Tiek

statistinio modelio parametrai, tiek stebėjimai iš esmės nesiskiria – abu yra atsitiktiniai kintamieji. Tikslas yra įvertinti nepalankių įvykių tikimybes P_j^m, kai Y_j^m yra stebėtas m-tojo įvykio pasirodymų skaičius, $j=\overline{1,K}$ ir $m=\overline{1,M}$. Dėl didelių populiacijos apimties N_j skirtumų paprastas santykinis įvertis $\frac{Y_j^m}{N_j}$ ne visada tinka. Todėl nagrinėjamas empirinis Bajeso metodas, taikant retų įvykių tikimybėms modeliuoti empirinį Bajeso Puasono–Gauso modelį.

Empiriniame Bajeso metode priimama, kad m-tojo įvykio j-tojoje populiacijoje skaičius Y_j^m pasiskirstęs pagal Puasono dėsnį su parametru $\lambda_j^m = N_j \cdot P_j^m$, $1 \le m \le M$, t.y., su tankiu (Bradley ir Thomas, 2000; Clayton ir Kaldor, 1987; Tsutakawa ir kt., 1985):

$$f\left(Y_j^m, \lambda_j^m\right) = e^{-\lambda_j^m} \frac{\left(\lambda_j^m\right)^{Y_j^m}}{\left(Y_j^m\right)!} \quad j=1,\ldots,K. \tag{2}$$

Puasono–Gauso modelyje priimama, kad įvykių tikimybių logitai populiacijose yra pasiskirstę pagal daugiamatį Gauso dėsnį su parametrais μ, Ω (Bradley ir Thomas, 2000), t.y., logit skirstinio (1) tankis

$$g(\alpha, \mu, \Omega) = \frac{\exp\left(-(\alpha-\mu)^T \Omega^{-1}(\alpha-\mu)\right)}{\sqrt{|\Omega|}\cdot(2\pi)^{\frac{M}{2}}}. \tag{3}$$

Tikimybės P_j^m įvertis yra apskaičiuojamas aposterioriniu vidurkiu

$$P_j^m = \frac{\int_{-\infty}^{+\infty} \frac{1}{1+e^{-\alpha_m}} \prod_{m=1}^{M} f\left(Y_j^m, \frac{N_j}{1+e^{-\alpha_m}}\right) g(\alpha, \mu, \Omega) d\alpha}{D_j(\mu, \Omega)}, \tag{4}$$

čia

$$D_j(\mu, \Omega) = \int_{-\infty}^{+\infty} \prod_{m=1}^{M} f\left(Y_j^m, \frac{N_j}{1+e^{-\alpha_m}}\right) g(\alpha, \mu, \Omega) d\alpha, \quad j=\overline{1,K}, \quad m=\overline{1,M} \tag{5}$$

Taikant Bajeso metodą statistikoje dažnai tenka minimizuoti tam tikras funkcijas, išreikštas per aposteriorinio tankio integralą. Taigi, tegul nežinomi parametrai μ, Ω yra vertinami didžiausio tikėtinumo metodu (Bradley ir Thomas, 2000; Tsutakawa ir kt., 1985). Gaunama, kad logaritminė tikėtinumo funkcija yra:

$$L(\mu, \Omega) = -\sum_{j=1}^{K} \ln\left(\int_{-\infty}^{+\infty} \prod_{m=1}^{M} f\left(Y_j^m, \frac{N_j}{1+e^{-\alpha_m}}\right) g(\alpha, \mu, \Omega) d\alpha\right) = -\sum_{j=1}^{K} \ln\left(D_j(\mu, \Omega)\right), \tag{6}$$

kurią minimizavus gaunami parametrų μ, Ω įverčiai.

2. Didžiausio tikėtinumo funkcijos išvestinės

Tikėtinumo funkcija (6) gali būti diferencijuojama daugelį kartų pagal parametrus μ, Ω. Nesunku įsitikinti, kad šios funkcijos pirmos eilės išvestinės yra:

$$\frac{\partial L(\mu,\Omega)}{\partial \mu}=\sum_{j=1}^{K}\frac{\int_{-\infty}^{+\infty}\Sigma^{-1}(\alpha-\mu)\prod_{m=1}^{M}f\left(Y_j^m,\frac{N_j}{1+e^{-\alpha_m}}\right)g(\alpha,\mu,\Omega)d\alpha}{D_j(\mu,\Omega)}, \quad (7)$$

$$\frac{\partial L(\mu,\Omega)}{\partial \Omega}=-\sum_{j=1}^{K}\frac{\int_{-\infty}^{+\infty}\left(\Omega^{-1}-\Omega^{-1}(\alpha-\mu)(\alpha-\mu)^T\Omega^{-1}\right)\prod_{m=1}^{M}f\left(Y_j^m,\frac{N_j}{1+e^{-\alpha_m}}\right)g(\alpha,\mu,\Omega)d\alpha}{D_j(\mu,\Omega)} \quad (8)$$

Išvestines (7), (8) prilyginus nuliui, gaunamos lygtys, kurias tenkina Puasono–Gauso modelio parametrų įverčiai:

$$\overline{\mu}=\frac{1}{K}\sum_{j=1}^{K}\frac{\int_{-\infty}^{+\infty}\alpha\cdot\prod_{m=1}^{M}f\left(Y_j^m,\frac{N_j}{1+e^{-\alpha_m}}\right)g(\alpha,\mu,\Omega)d\alpha}{D_{j,k}^m(\mu,\Omega)}, \quad (9)$$

$$\overline{\Omega}=\frac{1}{K}\sum_{j=1}^{K}\frac{\int_{-\infty}^{+\infty}(\alpha-\mu)(\alpha-\mu)^T\prod_{m=1}^{M}f\left(Y_{j,k}^m,\frac{N_j}{1+e^{-\alpha}}\right)g(\alpha,\mu,\Omega)d\alpha}{D_{j,k}^m(\mu,\Omega)} \quad (10)$$

3. Puasono–Gauso modelio parametrų įverčiai

Didžiausio tikėtinumo įverčius $\overline{\mu},\overline{\Omega}$ galima apskaičiuoti kvazi-Niutono metodu (Dennis ir Schnabel, 1996), pasinaudojus (6), (7), (8) išraiškomis. Be to, didžiausio tikėtinumo funkcijos minimizavimą ir integravimą galima atlikti pasinaudojus matematine programine įranga MATHCAD, MAPLE ir pan. Panagrinėsime „fiksuoto taško iteracijų“ metodą parametrų μ,Ω didžiausio tikėtinumo įverčiams rasti, pasinaudojus lygybėmis (9), (10) (Kantorovich ir Akilov, 1982):

$$\mu_{t+1}=\frac{1}{K}\sum_{j=1}^{K}\frac{\int_{-\infty}^{+\infty}\alpha\cdot f\left(Y_j,\frac{N_j}{1+e^{-\alpha}}\right)g(\alpha,\mu_t,\Omega_t)d\alpha}{D_j(\Omega_t,\Sigma_t)}, \quad (11)$$

$$\Omega_{t+1}=\frac{1}{K}\sum_{j=1}^{K}\frac{\int_{-\infty}^{+\infty}(\alpha-\mu_t)(\alpha-\mu_t)^T f\left(Y_j,\frac{N_j}{1+e^{-\alpha}}\right)g(\alpha,\mu_t,\Omega_t)d\alpha}{D_j(\mu_t,\Omega_t)}. \quad (12)$$

Integralus (4)–(6), (11)–(12) galima apskaičiuoti pasinaudojus Ermito–Gauso kvadratūrinėmis formulėmis (Abramowitz ir Stegun, 1964). Lygybėse (11), (12) galima paimti tokį pradinį tašką (μ_0,Ω_0):

$$\mu_0=P, \quad (13)$$

$$\Omega_0=\frac{\sum_{j=1}^{K}(Y_j-N_jP)^T\cdot(Y_j-N_jP)}{\sum_{j=1}^{K}N_j}, \quad (14)$$

čia $$P=\left(\frac{\sum_{j=1}^{K}Y_j^1}{\sum_{j=1}^{K}N_j},\frac{\sum_{j=1}^{K}Y_j^2}{\sum_{j=1}^{K}N_j},\ldots,\frac{\sum_{j=1}^{K}Y_j^M}{\sum_{j=1}^{K}N_j}\right),\ Y_j=\left(Y_j^1,Y_j^2,\ldots,Y_j^M\right).$$

4. MCMC algoritmas

„Fiksuoto taško“ algoritmą (11), (12) galima realizuoti Monte–Karlo Markovo grandinių metodu. Tegul, sugeneruota t grandžių ir kiekvienoje apskaičiuoti įverčiai μ_t, Ω_t, paėmus pradines reikšmes (13), (14). Kiekvienoje grandyje generuojami daugiamačiai Gauso vektoriai $\alpha_{j,k} \sim N(\mu^t,\Omega^t),\quad k=1,\ldots,N^t$, čia N^t yra imties tūris t-toje grandyje. Siekiant išvengti skaičiuojamųjų problemų, galinčių atsirasti dėl labai mažų tarpinių rezultatų reikšmių, skaičiuojant integralus tiesiog pagal formules (4)–(12) įvedama pagalbinė funkcija

$$r_j(\alpha)=\ln\left(\prod_{m=1}^{M}f_j(Y_j^m,\frac{N_j}{1+e^{-\alpha_m}})/\prod_{m=1}^{M}f_j(Y_j^m,\frac{N_j}{1+e^{-\mu_m}})\right)=\sum_{m=1}^{M}\frac{N_j\left(e^{-\alpha_m}-e^{-\mu_m}\right)}{\left(1+e^{-\mu_m}\right)\left(1+e^{-\alpha_m}\right)}+Y_j^m\cdot\ln\left(\frac{1+e^{-\mu_m}}{1+e^{-\alpha_m}}\right). \tag{15}$$

Toliau apskaičiuojamos sumos:

$$\tilde{D}_j^t=\sum_{k=1}^{N^t}r_j(\alpha_{j,k}), \tag{16}$$

$$\tilde{D}2_j^t=\sum_{k=1}^{N^t}\left(r_j(\alpha_{j,k})-\frac{\tilde{D}_j^t}{N^t}\right)^2, \tag{17}$$

$$\tilde{m}_j^t=\sum_{k=1}^{N^t}\alpha_{j,k}\cdot r(\alpha_{j,k}), \tag{18}$$

$$\tilde{S}_j^t=\sum_{k=1}^{N^t}\left(\alpha_{j,k}-\tilde{m}_j^t\right)\cdot\left(\alpha_{j,k}-\tilde{m}_j^t\right)^T\cdot r(\alpha_{j,k}), \tag{19}$$

$$p_{j,m}^t=\sum_{k=1}^{N^t}\frac{r(\alpha_{j,k})}{1+e^{-\alpha_{j,k,m}}}, \tag{20}$$

pagal kurias gaunami sekančios iteracijos įverčiai:

$$\mu^{t+1}=\frac{1}{K}\sum_{j=1}^{K}\frac{\tilde{m}_j^t}{\tilde{D}_j^t}, \tag{21}$$

$$\Sigma^{t+1}=\frac{1}{K}\sum_{j=1}^{K}\frac{\tilde{S}_j^t}{\tilde{D}_j^t}, \tag{22}$$

tikėtinumo funkcijos įvertis:

$$L^t=-\sum_{j=1}^{K}\ln\tilde{D}_j^t, \tag{23}$$

jos imties dispersijos įvertis:

$$d^t = \sum_{j=1}^{K} \left(\frac{\widetilde{D}2_j^t \cdot N^t}{\left(\widetilde{D}_j^t\right)^2} - 1 \right), \tag{24}$$

įvykių tikimybių populiacijose įverčiai:

$$\widetilde{P}_{j,m}^t = \frac{p_{j,m}^t}{\widetilde{D}_j^t}. \tag{25}$$

Algoritmo stabdymui įvestas kriterijus, skirtas patikrinti hipotezei apie vidurkių ir kovariacijų matricų sutapimą dviejose gretimose iteracijose:

$$H^t = \frac{K}{\frac{1}{K}\sum_{j=1}^{K} \frac{\widetilde{D}2_j^t}{\left(\widetilde{D}_j^t\right)^2}} \cdot \left(\ln \frac{\left|\Omega^{k+1}\right|}{\left|\Omega^k\right|} + SP\left(\Omega^{k+1} \cdot \left(\Omega^k\right)^{-1}\right) + \left(\mu^{k+1} - \mu^k\right)^T \cdot \left(\Omega^k\right)^{-1} \cdot \left(\mu^{k+1} - \mu^k\right) - M \right) \tag{26}$$

Eksperimentiniai tyrimai ir teorinė analizė parodė, kad esant pakankamai dideliam imties tūriui tikėtinumo funkcijos maksimumo taške šis kriterijus yra pasiskirstęs pagal Fišerio dėsnį su $\nu = \frac{M(M+3)}{2}$ laisvės laipsnių.

Algoritmas stabdomas, jei stabdymo kriterijus neprieštarauja hipotezei apie vidurkių ir kovariacijų matricų sutapimą dviejose gretimose iteracijose:

$$H^t \le F_{\alpha,\nu} \tag{27}$$

tikėtinumo funkcijos pasikliovimo intervalas yra mažesnis už pasirinktą mažą reikšmę ε:

$$2 \cdot \eta_\gamma \cdot \sqrt{\frac{d^t}{N^t}} \le \varepsilon, \tag{28}$$

α, γ yra atitinkami Fišerio ir normaliojo skirstinio kvantiliai, ir kovariacijų matricų įverčių didžiausios ir mažiausios tikrinių reikšmių santykis ne daugiau 10.

Jei nors vienas stabdymo sąlyga nėra tenkinama, generuojama nauja Monte–Karlo imtis. Imties tūrį galima reguliuoti pagal taisyklę (Sakalauskas, 2000):

$$N^{t+1} \ge \frac{N^t \cdot \nu}{H^t} \cdot F_{\beta,\nu}, \tag{29}$$

čia $F_{\beta,\nu}$ – Fišerio skirstinio kvantilis, β – pasikliovimo lygmuo. Šitokios taisyklės taikymas leidžia racionaliai parinkti imčių tūrį Monte–Karlo Markovo grandinėje ir kartu užtikrina konvergavimą į tikėtinumo funkcijos maksimumą.

5. Kompiuterinis modeliavimas

Sudarytam algoritmui patikrinti buvo atliktas kompiuterinis eksperimentas. Sugeneruota imtis $\Theta = (\theta_1, \theta_2, \ldots, \theta_K)$, sudaryta iš $K = 10$ populiacijų, kuriose galėjo įvykti $M = 3$ įvykių, kai tikimybių logitai pasiskirstę pagal daugiamatį Gauso dėsnį su parametrais:

$$\mu = \begin{pmatrix} -3 \\ -4 \\ -5 \end{pmatrix}; \quad \Sigma = \begin{pmatrix} 0{,}25 & 0 & 0 \\ 0 & 0{,}25 & 0 \\ 0 & 0 & 0{,}25 \end{pmatrix}. \tag{30}$$

Toliau buvo sugeneruota $t = 100$ Monte–Karlo Markovo grandžių pagal (15)–(20), (29) išraiškas. Siekiant išvengti labai mažų arba labai didelių imties tūrio reikšmių, buvo taikomos ribos: $500 \le N^k \le 17000$.

Sprendinys pradėjo konverguoti jau po $t = 6$ iteracijų. Gautos parametrų reikšmės pateiktos 1 lentelėje:

1 lentelė. Parametrų įverčiai.

Iteracija	μ_1	μ_2	μ_3	Tikslo funkcija	Pasikliovimo intervalas	Imties tūris	Statistika
1	–2,96	–4,29	–5,52	–62,90	5,57	500	9,55
2	–2,89	–4,04	–5,27	–396,58	4,81	500	6,18
3	–2,91	–4,03	–5,19	–420,42	2,97	500	3,86
4	–2,90	–4,04	–5,16	–424,87	3,2	500	0,35
5	–2,91	–4,04	–5,13	–428,05	1,57	2 963	1,41
6	–2,90	–4,04	–5,14	–427,57	1,32	4 383	0,32
7	–2,91	–4,04	–5,13	–425,54	0,75	13 986	0,40
8	–2,91	–4,04	–5,14	–425,33	0,75	14 345	0,40
9	–2,91	–4,04	–5,13	–425,71	0,75	13 525	0,84
10	–2,91	–4,04	–5,13	–426,47	0,75	15 135	0,22

Išvados

Darbe sudarytas MCMC algoritmas, pasinaudojus statistiniu stabdymo kriterijumi ir imties tūrio reguliavimu. Atlikus skaičiavimus, gaunami nežinomų parametrų įverčiai ir tikimybės. Sudarytas algoritmas gali būti taikomas socialinių ir medicininių duomenų analizei.

Literatūra

Abramowitz, M.; Stegun, I. A. (1964). Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. New York: Dover Publications.

Altaleb, Anas; Chauveau, Didier (2002). Bayesian analysis of the Logit model and comparison of two. Metropolis–Hastings strategies. Computational Statistics and Data Analysis, 39: 137–152.

Bradley, P. C.; Thomas, A. L. (2000). Bayes and Empirical Bayes Methods for Data Analysis. New York: Chapman and Hall.

Chen, Fang (2009). Bayesian modeling using the MCMC procedure [interaktyvus]. [žiūrėta 2013 m. gegužės 4 d.]. Prieiga per internetą: <http://support.sas.com/resources/papers/proceedings09/257-2009.pdf>.

Clayton, David; Kaldor, John (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, Biometrics, 43(3): 671–681.

Dennis J. E.; Schnabel, R. B. (1996). Numerical methods for unconstrained optimization

and nonlinear equations. Philadelphia: Classics in Applied Mathematics.
Kantorovich, L. V.; Akilov, G. P. (1982). Functional Analysis. Oxford: Pergamon Press.
Liseo, Brunero; Loperfido, Nicola (2003). A Bayesian interpretation of the multivariate skew-normal distribution. Statistics & Probability Letters, 61(4): 395–401.
Pearce, D. W. (2006). Aiškinamasis ekonomikos anglų–lietuvių kalbų žodynas. Lithuania: Vilnius, TEV.
Sakalauskas, Leonidas (2000). Nonlinear stochastic optimization by Monte–Carlo estimators. Informatica, 11(4): 455–468.
Tsutakawa, Robert K.; Shoop, Gary L., Marienfeld, Carl J. (1985). Empirical Bayes estimation of cancer mortality rates. Statistics in Medicine, 4(2): 201–212.

MULTIDIMENSIONAL RARE EVENT PROBABILITY ESTIMATION ALGORITHM

Leonidas Sakalauskas, Ingrida Vaiciulyte

Summary

This work contains Monte–Carlo Markov Chain algorithm for estimation of multi-dimensional rare events frequencies. Logits of rare event likelihood we are modeling with Poisson distribution, which parameters are distributed by multivariate normal law with unknown parameters – mean vector and covariance matrix. The estimations of unknown parameters are calculated by the maximum likelihood method. There are equations derived, those must be satisfied with model's maximum likelihood parameters estimations. Positive definition of evaluated covariance matrixes are controlled by calculating ratio between matrix maximum and minimum eigenvalues.

Key words: Monte–Carlo Markov Chain, Bayesian method, likelihood method, Poisson–Gaussian model.