

Dariusz Biskup

BAYESOWSKA PROCEDURA WYBORU LICZBY SKŁADNIKÓW MIESZANKI ROZKŁADÓW

1. Wstęp

W opracowaniu opisana zostanie bayesowska procedura wyboru liczby składników mieszanek. Opiera się ona na procedurach wyboru modelu, których numeryczna implementacja wykorzystuje algorytm *Reversible Jump Markov Chain Monte Carlo* (RJMCMC) (por. [2; 4]).

Szczególną trudnością związaną z estymacją liczby składników mieszanki rozkładów jest nieustalona liczba parametrów modelu. Do wyznaczenia liczby składników nie jest możliwe stosowanie np. metody największej wiarygodności, ponieważ wyznacza ona jako optymalny model o największej liczbie składników.

Bayesowskie procedury wyboru modelu zakładają, że model jest traktowany jako nieznany parametr, dla którego można określić rozkład prawdopodobieństwa. Model, a w opisywanym zagadnieniu liczba składników mieszanki rozkładów, wybierany jest według kryterium największego prawdopodobieństwa *a posteriori*.

2. Wybór liczby składników mieszanki rozkładów zero-jedynkowych

Założmy, że dane są 3 urny. W każdej z urn znajduje się odpowiednio $\tilde{\theta}_i\%$ kul białych i $(1 - \tilde{\theta}_i)\%$ ($i = 1, 2, 3$) kul czarnych. Zdefiniujmy zmienną losową X , która zdarzeniu polegającemu na wyciągnięciu kuli białej przypisuje liczbę 1, natomiast kuli czarnej liczbę 0. Rozpatrzmy dwa modele losowania:

1. Z prawdopodobieństwem $\tilde{\omega}$ losujemy urnę pierwszą, z prawdopodobieństwem $1 - \tilde{\omega}$ urnę drugą. Następnie z wylosowanej urny wyciągamy jedną kulę.

2. Losujemy jedną kulę z urny trzeciej.

Obserwator otrzymuje informację, że wykonano doświadczenie polegające na wylosowaniu kuli albo według schematu 1, albo według schematu 2. Na podstawie wartości zmiennej losowej X (znajomości koloru wylosowanej kuli) należy obliczyć prawdopodobieństwa modelu 1 i modelu 2.

Model 1 odpowiada losowaniu z mieszanki dwóch rozkładów zero-jedynkowych o prawdopodobieństwach sukcesu $\tilde{\theta}_1$ i $\tilde{\theta}_2$ i wagach $\tilde{w}$ i $1 - \tilde{w}$. Model 2 polega na losowaniu z rozkładu zero-jedynkowego z prawdopodobieństwem sukcesu $\tilde{\theta}_3$.

Niech zmienna losowa M określa numer modelu (schematu losowania). Załóżmy *a priori*, że $p(M = 1) = p(M = 2) = 0,5$. Niepewność odnośnie wartości parametrów $\tilde{\theta}_1$, $\tilde{\theta}_2$, $\tilde{\theta}_3$, $\tilde{w}$ opisana zostanie przez zadane arbitralnie rozkłady *a priori* (przyjęta zostanie również *a priori* niezależność zmiennych $\tilde{\theta}_1$, $\tilde{\theta}_2$, $\tilde{\theta}_3$, $\tilde{w}$):

$$\begin{aligned} p(\tilde{\theta}_1 = 0,1 | M = 1) &= p(\tilde{\theta}_1 = 0,9 | M = 1) = 0,5, \\ p(\tilde{\theta}_2 = 0,1 | M = 1) &= p(\tilde{\theta}_2 = 0,9 | M = 1) = 0,5, \\ p(\tilde{\theta}_3 = 0,1 | M = 2) &= 0,1, \quad p(\tilde{\theta}_3 = 0,9 | M = 3) = 0,9, \\ p(\tilde{w} = 0,1 | M = 1) &= p(\tilde{w} = 0,9 | M = 1) = 0,5. \end{aligned}$$

Wyznamy następnie następujące prawdopodobieństwa warunkowe:

$$\begin{aligned} p(X = 1 | M = 1, \tilde{\theta}_1 = \theta_1, \tilde{\theta}_2 = \theta_2, \tilde{w} = w) &= w\theta_1 + (1 - w)\theta_2, \\ p(X = 0 | M = 1, \tilde{\theta}_1 = \theta_1, \tilde{\theta}_2 = \theta_2, \tilde{w} = w) &= w(1 - \theta_1) + (1 - w)(1 - \theta_2), \\ p(X = 1 | M = 2, \tilde{\theta}_3 = \theta_3) &= \theta_3, \\ p(X = 0 | M = 2, \tilde{\theta}_3 = \theta_3) &= 1 - \theta_3. \end{aligned}$$

Niech zbiór $D_1 = (0,1; 0,9) \times (0,1; 0,9) \times (0,1; 0,9)$ oraz $D_2 = (0,1; 0,9)$. Obliczenie prawdopodobieństwa *a posteriori* modelu przeprowadzone zostanie przy założeniu, że zaobserwowano wartość zmiennej losowej $X = 1$. Wówczas:

$$\begin{aligned} p(M = 1 | X = 1) &= \frac{p(M = 1)p(X = 1 | M = 1)}{p(X = 1)} = \\ &= \frac{0,5}{p(X = 1)} \sum_{(\theta_1, \theta_2, w) \in D_1} p(\tilde{\theta}_1 = \theta_1, \tilde{\theta}_2 = \theta_2, \tilde{w} = w, X = 1 | M = 1) = \\ &= \frac{0,5}{p(X = 1)} \sum_{(\theta_1, \theta_2, w) \in D_1} p(\tilde{\theta}_1 = \theta_1, \tilde{\theta}_2 = \theta_2, \tilde{w} = w | M = 1) \times \end{aligned}$$

$$\begin{aligned} & \times p(X=1|M=1, \tilde{\theta}_1=\theta_1, \tilde{\theta}_2=\theta_2, \tilde{\omega}=\omega) = \\ & = \frac{0,5}{p(X=1)} \times \frac{1}{8} \times 4 = \frac{1}{4p(X=1)} \end{aligned}$$

oraz

$$\begin{aligned} p(M=2|X=1) &= \frac{p(M=2)p(X=1|M=2)}{p(X=1)} = \\ &= \frac{0,5}{p(X=1)} \sum_{\theta_3 \in D_2} p(\tilde{\theta}_3=\theta_3, X=1|M=2) = \\ &= \frac{0,5}{p(X=1)} \sum_{\theta_3 \in D_2} p(\tilde{\theta}_3=\theta_3|M=2)p(X=1|M=2, \tilde{\theta}_3=\theta_3) = \\ &= \frac{0,5}{p(X=1)} \times (0,1 \times 0,1 + 0,9 \times 0,9) = \frac{0,41}{p(X=1)}. \end{aligned}$$

Ponieważ $p(M=1|X=1) + p(M=2|X=1) = 1$, więc $p(X=1) = 0,66$. Stąd ostatecznie:

$$\begin{aligned} p(M=1|X=1) &= 0,38, \\ p(M=2|X=1) &= 0,62. \end{aligned}$$

W praktycznych zagadnieniach przestrzeń parametrów modelu nie jest przestrzenią dyskretną, w związku z czym do obliczenia prawdopodobieństw modeli wymagane jest zastosowanie metod numerycznych.

Opisany przykład stanowi prostą ilustrację zagadnienia bayesowskiego wyboru modelu. W kolejnym paragrafie pokazane zostaną ogólne wzory służące do wyznaczania prawdopodobieństw *a posteriori* modelu dla dowolnych przestrzeni parametrów.

3. Bayesowski wybór modelu

Prawdopodobieństwo prawdziwości modelu k można zgodnie z wzorem Bayesa wyznaczyć następująco:

$$p(M=k|y) = \frac{p(M=k)p(y|M=k)}{p(y)}, \quad (1)$$

gdzie: $y = (y_1, y_2, \dots, y_n)$,

$p(M=k|y)$ – prawdopodobieństwo, że prawdziwy jest model k , jeśli zaobserwowano dane y ,

$p(M = k)$ – prawdopodobieństwo *a priori* prawdziwości modelu k ,
 $p(y|M = k)$ – prawdopodobieństwo zaobserwowania danych y , jeśli prawdziwy jest model k .

Oznaczając przez $\theta^{(k)}$ wektor parametrów dla modelu k , wzór (1) można rozwinąć następująco:

$$\begin{aligned} p(M = k|y) &= \frac{p(M = k)p(y|M = k)}{p(y)} = \frac{p(M = k) \int_{\Theta^{(k)}} p(y, \theta^{(k)}|M = k) d\theta^{(k)}}{p(y)} = \\ &= \frac{p(M = k)}{p(y)} \int_{\Theta^{(k)}} p(y|M = k, \theta^{(k)}) p(\theta^{(k)}|M = k) d\theta^{(k)}. \end{aligned}$$

Zauważmy, że jeżeli prawdopodobieństwa *a priori* $p(M = k)$ są identyczne dla każdego modelu, wówczas wpływ na prawdopodobieństwo modelu ma wyłącznie składnik:

$$p(y|M = k) = \int_{\Theta^{(k)}} p(y|M = k, \theta^{(k)}) p(\theta^{(k)}|M = k) d\theta^{(k)}. \quad (2)$$

Składnik $p(y|M = k, \theta^{(k)})$ określa funkcję wiarygodności dla modelu k , natomiast $p(\theta^{(k)}|M = k)$ oznacza rozkład *a priori* parametrów modelu k .

4. Estymacja parametrów mieszanki rozkładów Poissona

Założmy, że numer modelu M określać będzie liczbę składników mieszanki. Wówczas wektor parametrów ma postać $\theta^{(k)} = (\lambda_1, \lambda_2, \dots, \lambda_k, p_1, p_2, \dots, p_k)$. Funkcja wiarygodności jest z kolei równa:

$$p(y|M = k, \theta^{(k)}) = \prod_{i=1}^n \sum_{j=1}^k p_j \frac{e^{-\lambda_j} \lambda_j^{y_j}}{y_j!}. \quad (3)$$

Przyjęte zostanie, że rozkład *a priori* może być przedstawiony jako

$$\begin{aligned} p(\theta^{(k)}|M = k) &= p(\lambda_1, \lambda_2, \dots, \lambda_k|M = k) p(p_1, p_2, \dots, p_k|M = k) = \\ &= p(p_1, p_2, \dots, p_k|M = k) \prod_{i=1}^k p(\lambda_i|M = k). \end{aligned} \quad (4)$$

Przyjęte zostanie ponadto, że rozkład wektora wag $\mathbf{p} = (p_1, p_2, \dots, p_k)$ jest *a priori* rozkładem Dirichleta $D(1, 1, \dots, 1)$, co oznacza występowanie braku informacji *a priori* o wartościach wag. Rozkład *a priori* $p(\lambda_j | M = k)$ (dla $j = 1, 2, \dots, k$) składników wektora $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_k)$ ustalony zostanie jako rozkład *Gamma*(0,01; 0,01) (por. [1]).

Opisana zostanie obecnie metoda estymacji parametrów mieszanki jednowymiarowych rozkładów Poissona przy założeniu nieznajomości liczby jej składników (por. [6]). Wybór liczby składników mieszanki dokonany zostanie na podstawie kryterium prawdopodobieństwa *a posteriori*.

Rozkład rozpatrywanej mieszanki dany jest wzorem:

$$p(\mathbf{y} | \lambda_1, \lambda_2, \dots, \lambda_k, p_1, p_2, \dots, p_k, k) = \sum_{j=1}^k p_j \frac{e^{-\lambda_j} \lambda_j^{y_j}}{y_j!}. \quad (5)$$

Przyjmuje się, że $\sum_{j=1}^k p_j = 1$. Niech ponadto $0 < \lambda_1 < \lambda_2 < \dots < \lambda_k$, co zapewnia

identyfikowalność parametrów modelu (por. [1]). Wyznaczenie rozkładu prawdopodobieństwa liczby składników k mieszanki umożliwi wybór modelu najbardziej prawdopodobnego.

Parametry p_j oznaczają wagi poszczególnych składników mieszanki (interpretowane jako prawdopodobieństwa *a priori*, że dana obserwacja pochodzi z j -tego składnika). Parametry modelu (5) traktowane będą jako zmienne losowe. W szczególności zmienną losową jest liczba składników mieszanki k .

Do każdej obserwacji y_i ($i = 1, 2, \dots, n$) pochodzącej z rozkładu (5) można przypisać nieobserwowany ciąg zmiennych losowych $Z_{i1}, Z_{i2}, \dots, Z_{ik}$, które opisywać będą numer składnika, z którego pochodzi i -ta obserwacja. Przyjmuje się, że $Z_{ij} = 1$, jeżeli i -ta obserwacja należy do składnika o numerze j , oraz $Z_{ij} = 0$ w przeciwnym wypadku. Oznacza to, że

$$p(Z_{ij} = 1) = p_j, \text{ dla } j = 1, 2, \dots, k, i = 1, 2, \dots, n. \quad (6)$$

5. Algorytm Reversible Jump Markov Chain Monte Carlo

Algorytm RJMCMC stanowi uogólnienie algorytmu Metropolisa-Hastingsa (por. [5]). Gdy nie jest zmieniany wymiar przestrzeni parametrów modelu, odpowiada on dokładnie algorytmowi Metropolisa-Hastingsa.

Ogólnie przestrzeń parametrów Θ dla zagadnień wyboru modelu o zmiennej liczbie parametrów może być przedstawiona następująco (por. [5]):

$$\Theta = \bigcap_{k=1}^{k_{\max}} \{\Theta_k\} \times \Theta_k, \quad (7)$$

gdzie $k_{\max}$ oznacza liczbę rozpatrywanych modeli, k – numer modelu, Θ_k – przestrzeń parametrów dla modelu o numerze k . Na przykład jeżeli rozpatrujemy 3 modele mające odpowiednio 1, 2 i 3 parametry, to przestrzeń $\Theta = \{1\} \times R^1 \cup \cup \{2\} \times R^2 \cup 3 \times R^3$. Oznacza to, że należy określić miarę probabilistyczną na przestrzeni, w skład której wchodzi punkty leżące na prostej, punkty leżące na płaszczyźnie i punkty z przestrzeni trójwymiarowej.

Algorytm RJMCMC umożliwia efektywne pobieranie próbek losowych z przestrzeni zadanej wzorem (7). Gdy występuje zmiana wymiaru przestrzeni parametrów, pojedyncza iteracja algorytmu przebiega w następujących etapach:

1. Niech (j, θ_j) oznacza aktualne wartości wektora parametrów θ_j modelu o numerze j , którego wymiar wektora parametrów wynosi k_j .

2. Wykonanie przejścia do stanu $(j', \theta_{j'})$ ze stanu (j, θ_j) z prawdopodobieństwem $h(j, j')$.

3. Wylosowanie wektora $\mathbf{u}$ z zadanego rozkładu $q(\mathbf{u} | \theta_{j'}, j, j')$.

4. Podstawienie $(\theta_{j'}, \mathbf{u}') = g_{j, j'}(\theta_j, \mathbf{u})$, gdzie $g_{j, j'}$ oznacza wzajemnie jednoznaczную, deterministyczną funkcję. Podstawienie to ma na celu dopasowanie wymiarów modelu aktualnego do modelu docelowego ($k_j + \dim(\mathbf{u}) = k_{j'} + \dim(\mathbf{u}')$).

Ponadto $g_{j', j} = g_{j, j'}^{-1}$.

5. Akceptacja zmiany modelu z prawdopodobieństwem $\min(1, A)$, gdzie:

$$A = \frac{p(\mathbf{y} | \theta_{j'}, j') p(\theta_{j'} | j') p(j') h(j', j) q(\mathbf{u} | \theta_{j'}, j', j)}{p(\mathbf{y} | \theta_j, j) p(\theta_j | j) p(j) h(j, j') q(\mathbf{u} | \theta_j, j, j')} \left| \frac{\partial g_{j, j'}(\theta_j, \mathbf{u})}{\partial(\theta_j, \mathbf{u})} \right|. \quad (8)$$

Jeżeli wymiar modelu j' jest wyższy niż wymiar modelu j , wówczas $\dim(\mathbf{u}') = 0$ i funkcja $g_{j, j'}$ wykonuje odwzorowanie $\theta_{j'} = g_{j, j'}(\theta_j, \mathbf{u})$. Jeżeli wymiar modelu j' jest niższy niż wymiar modelu j , wówczas $\dim(\mathbf{u}) = 0$ i odwzorowanie ma postać $(\theta_{j'}, \mathbf{u}') = g_{j, j'}(\theta_j)$.

Jeżeli nie następuje zmiana modelu, czyli wymiar modelu j jest równy wymiarowi modelu j' , wówczas nowe wartości wektora parametrów mogą być generowane przy wykorzystaniu algorytmu Gibbsa (por. [5]). Liczba iteracji algorytmu

RJMCMC w zagadnieniach praktycznych waha się w przedziale od 100 000 do 1 000 000. Prawdopodobieństwo prawdziwości modelu j jest szacowane jako liczba iteracji, w których algorytm znajdował się w stanie (j, θ_j) , podzielona przez ogólną liczbę iteracji.

W praktyce najczęściej iteracje algorytmu RJMCMC są wykonywane na przemian z iteracjami algorytmu Gibbsa.

6. Zastosowanie algorytmu RJMCMC do estymacji parametrów mieszanek

6.1. Wprowadzenie

Ustalono zostanie, że liczba składników mieszanki k będzie miała *a priori* rozkład jednostajny przy ustalonej maksymalnej liczbie składników mieszanki na poziomie $k_{\max} = 10$.

Można pokazać, że warunkowe rozkłady parametrów modelu (wykorzystywane w sytuacji, gdy stosowany jest algorytm Gibbsa) są następujące (por. [1]):

$$(z_{i1}, z_{i2}, \dots, z_{in}) | \mathbf{p}, \lambda \sim M(1, w_{i1}, w_{i2}, \dots, w_{ik}), \quad i = 1, 2, \dots, n, \quad (9)$$

gdzie $w_{ij} = \frac{p_j p(y_i | \lambda_j)}{p(y_j)} \quad (j = 1, 2, \dots, k),$

$$\mathbf{p} | (z_{11}, z_{21}, \dots, z_{n1}), \dots, (z_{1k}, z_{2k}, \dots, z_{nk}) \sim D \left(1 + \sum_{i=1}^n z_{i1}, \dots, 1 + \sum_{i=1}^n z_{ik} \right), \quad (10)$$

$$\lambda_j | (z_{1j}, z_{2j}, \dots, z_{nj}) \sim \Gamma \left(0, 01 + \sum_{i=1}^n z_{ij}, y_i, 0, 01 + \sum_{i=1}^n z_{ij} \right) I(\lambda_{j-1}, \lambda_{j+1}), \quad j = 1, 2, \dots, k. \quad (11)$$

Losowanie dla parametrów danych wzorami od (9) do (11) wykonywane będzie przy użyciu algorytmu Gibbsa.

Pojedyncza iteracja algorytmu RJMCMC zaadaptowanego do estymacji parametrów mieszanki składa się z następujących kroków:

- aktualizacja wektora wag $\mathbf{p}$ (wzór (10)),
- aktualizacja wektora parametrów λ (wzór (11)),
- aktualizacja przypisań do klas $z_{i1}, z_{i2}, \dots, z_{ik}$ (wzór (9)),
- podział pojedynczego składnika mieszanki na dwa składniki lub połączenie dwóch składników w jeden.

W kolejnym paragrafie opisany zostanie szczegółowo punkt d) algorytmu.

6.2. Podział lub połączenie składnika

W kroku d) dokonywany jest losowy wybór pomiędzy podziałem i połączeniem składników z prawdopodobieństwami b_k (dla podziału) oraz $d_k = 1 - b_k$ (dla połączenia). Prawdopodobieństwa te zależą od parametru k . Przyjęte zostanie, że $d_1 = 0$, $b_{k_{\max}} = 0$ oraz $b_k = d_k = 0,5$ dla $k = 2, 3, \dots, k_{\max} - 1$. Połączenie składników polega na losowym wyborze pary składników o numerach j_1 i j_2 , które sąsiadują ze sobą ze względu na wartość parametru λ . Składniki te są łączone i tworzony jest nowy składnik o numerze j^* (numery istniejących składników są następnie odpowiednio przenumerywane). Charakterystyki dla nowego składnika j^* wyznaczane są zgodnie z następującymi wzorami (por. [1]):

$$\begin{aligned} p_{j^*} &= p_{j_1} + p_{j_2}, \\ \lambda_{j^*} &= \frac{\lambda_{j_1} p_{j_1} + \lambda_{j_2} p_{j_2}}{p_{j^*}}. \end{aligned} \quad (12)$$

Obserwacje, które należały do składników j_1 i j_2 , zostaną przydzielane do nowego składnika j^* .

Podział składnika polega na losowym wyborze komponentu j^* , który zostaje podzielony na składniki j_1 i j_2 . Parametry oraz wagi składników j_1 i j_2 spełniają zależności dane równaniami (12). Ponieważ równania te nie wyznaczają jednoznacznie nowych parametrów, losowane są dwie wartości u_1 i u_2 zgodnie z następującymi rozkładami (por. [1]):

$$u_1 \sim \text{Beta}(2, 2), \quad u_2 \sim \text{Beta}(2, 2). \quad (13)$$

Wybór rozkładów podanych we wzorze (13) jest w zasadzie dowolny. Celem jest uzyskanie jak najwyższego prawdopodobieństwa akceptacji (wzór (8)), co powoduje, że zmiany modelu występują stosunkowo częściej, niż gdy prawdopodobieństwo jest bliskie zeru, co pozwala zredukować ogólną liczbę iteracji algorytmu. Wagi oraz parametry dla nowych składników wyznaczane są następnie zgodnie z następującymi równaniami (wynikają one z zależności (12)):

$$\begin{aligned} p_{j_1} &= p_{j^*} u_1, \\ p_{j_2} &= p_{j^*} (1 - u_1), \\ \lambda_{j_1} &= \lambda_{j^*} u_2, \\ \lambda_{j_2} &= \lambda_{j^*} \left(\frac{1 - u_1 u_2}{1 - u_1} \right). \end{aligned} \quad (14)$$

Jeżeli nowe składniki nie spełniają warunku sąsiedztwa (czyli wartość oczekiwana innego składnika należy do przedziału $(\lambda_{j1}, \lambda_{j2})$), wówczas krok algorytmu polegający na podziale składnika nie jest wykonywany.

Jakobian transformacji wektora $(p_{j^*}, \lambda_{j^*}, u_1, u_2)$ na wektor $(p_{j1}, p_{j2}, \lambda_{j1}, \lambda_{j2})$, wyznaczany na podstawie funkcji opisanych wzorem (14) ma postać (por. [1]):

$$|J| = \begin{vmatrix} u_1 & 0 & p_{j^*} & 0 \\ 1-u_1 & 0 & -p_{j^*} & 0 \\ 0 & u_2 & 0 & \lambda_{j^*} \\ 0 & \frac{1-u_1u_2}{1-u_1} & \frac{\lambda_{j^*}(1-u_2)}{(1-u_1)^2} & \frac{-\lambda_{j^*}u_1}{1-u_1} \end{vmatrix} = \frac{p_{j^*}\lambda_{j^*}}{1-u_1}. \quad (15)$$

Prawdopodobieństwo akceptacji wyrażone wzorem (8) w odniesieniu do zagadnienia estymacji parametrów mieszanki, gdy liczba składników mieszanki powiększana jest o jeden, przybiera następującą postać:

$$A = \frac{p(k')p(y|\lambda', z')p(p'|k', \delta)p(z'|p', k')p(\lambda'|k')}{p(k)p(y|\lambda, z)p(p|k, \delta)p(z|p, k)p(\lambda|k)} \times \\ \times \frac{h(k', k)}{h(k, k')} \times \frac{1}{q(u|\lambda, j, j')} \times |J|, \quad (16)$$

gdzie: $k' = k + 1$,

$$z = (z_{11}, z_{12}, \dots, z_{kn}),$$

$$\lambda = (\lambda_1, \lambda_2, \dots, \lambda_k),$$

$$y = (y_1, y_2, \dots, y_n),$$

$|J|$ – jacobian transformacji.

Ponieważ założono jednostajny rozkład *a priori* dla liczby składników mieszanki, oznacza to, że $p(k') = p(k) = \frac{1}{k_{\max}}$. Składnik $p(y|\lambda, z)$ jest równy:

$$p(y|\lambda, z) = \prod_{j=1}^k \prod_{i: z_{ij}=1} \frac{\lambda_j^{y_i}}{y_i!} e^{-\lambda_j}. \quad (17)$$

Analogiczny wzór można podać dla składnika $p(y|\lambda', z')$.

Wyrazy $p(p|k', \delta)$ oraz $p(p|k, \delta)$ wzoru (16) oznaczają rozkłady *a priori* wag mieszanki o $k+1$ i k składnikami. Wagi te mają *a priori* rozkład Dirichleta $D(\delta, \dots, \delta)$ z parametrem $\delta = 1$.

Z kolei

$$\frac{p(\mathbf{z}'|\mathbf{p}', k')}{p(\mathbf{z}|\mathbf{p}, k)} = \frac{p_1^{n_1} \dots p_{j_1}^{l_1} p_{j_2}^{l_2} \dots p_{k+1}^{n_{k+1}}}{w_1^{n_1} \dots w_{j_*}^{l_1+l_2} \dots w_k^{n_k}} = \frac{p_{j_1}^{l_1} p_{j_2}^{l_2}}{p_{j_*}^{l_1+l_2}}, \quad (18)$$

gdzie l_1, l_2 jest to liczba obserwacji przypisanych do składników j_1 i j_2 .

Rozkłady *a priori* $p(\lambda'|k')$ i $p(\lambda|k)$ są iloczynami odpowiednio $k+1$ i k rozkładów *Gamma*(0,01; 0,01).

Do wyznaczenia prawdopodobieństwa akceptacji niezbędne jest również obliczenie prawdopodobieństwa uzyskania określonej alokacji elementów składnika podlegającego podziałowi na składniki docelowe. Prawdopodobieństwo to można obliczyć zgodnie z następującym wzorem:

$$P_{\text{alloc}} = \prod_{i: z_i(j_1)=1} \frac{p_{j_1} \frac{\lambda_{j_1}^{y_i}}{y_i!} e^{-\lambda_{j_1}}}{p_{j_1} \frac{\lambda_{j_1}^{y_i}}{y_i!} e^{-\lambda_{j_1}} + p_{j_2} \frac{\lambda_{j_2}^{y_i}}{y_i!} e^{-\lambda_{j_2}}} \times \prod_{i: z_i(j_2)=1} \frac{p_{j_2} \frac{\lambda_{j_2}^{y_i}}{y_i!} e^{-\lambda_{j_2}}}{p_{j_1} \frac{\lambda_{j_1}^{y_i}}{y_i!} e^{-\lambda_{j_1}} + p_{j_2} \frac{\lambda_{j_2}^{y_i}}{y_i!} e^{-\lambda_{j_2}}}. \quad (19)$$

Składniki $h(j', j) = d_{k+1}$, $h(j, j') = b_k P_{\text{alloc}}$, natomiast funkcja $q(\cdot)$ oznacza gęstości rozkładów (13).

W podobny sposób można wyprowadzić prawdopodobieństwo akceptacji dla połączenia dwóch składników (należy uwzględnić, że wówczas $k' = k - 1$).

7. Przykłady

7.1. Estymacja parametrów mieszanki dwóch rozkładów Poissona

W przykładzie wykorzystano dane uzyskane za pomocą generatora liczb losowych. 50 obserwacji wylosowano z rozkładu Poissona dla $\lambda = 1$ oraz 50 obserwacji z rozkładu Poissona dla $\lambda = 5$. Oznacza to, że cały zbiór danych pochodzi z następującej mieszanki rozkładów:

$$Y_i \sim \frac{1}{2} Po(1) + \frac{1}{2} Po(5). \quad (20)$$

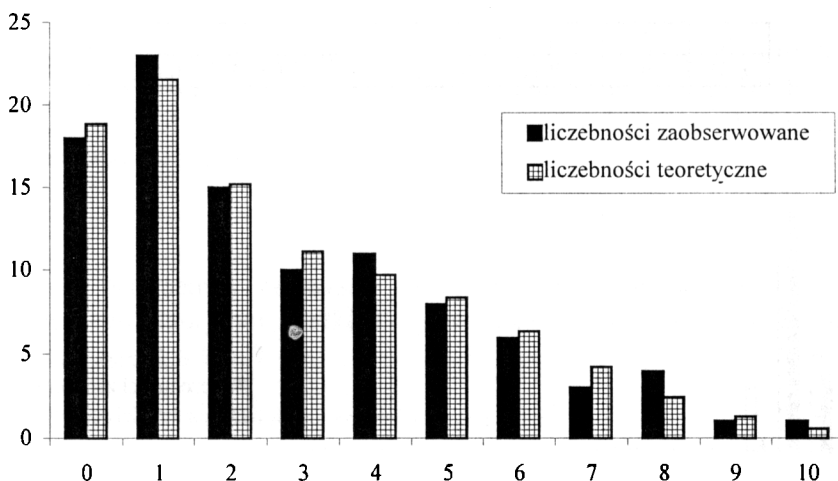
Histogram wygenerowanych danych przedstawia rys. 1.

Estymacja parametrów mieszanki wykonana została na podstawie 100 000 iteracji opisanego algorytmu. Rozkład *a posteriori* liczby składników mieszanki przedstawiony został w tab. 1.

Tabela 1. Rozkład liczby składników mieszanki

k	1	2	3	4
$p(k)$	0	0,95804	0,03819	0,00377

Najbardziej prawdopodobna liczba składników mieszanki jest zatem równa 2.



Rys. 1. Histogram danych wygenerowanych zgodnie z rozkładem (20)

Źródło: opracowanie własne.

Estymatory parametrów rozkładu uzyskane dla najbardziej prawdopodobnego wariantu dwóch składników są następujące:

$$\hat{\lambda}_1 = 1,06,$$

$$\hat{\lambda}_2 = 4,66,$$

$$\hat{p}_1 = 0,53,$$

$$\hat{p}_2 = 0,47.$$

7.2. Estymacja parametrów mieszanki trzech rozkładów Poissona

W przykładzie wykorzystano dane uzyskane za pomocą generatora liczb losowych. 100 obserwacji wylosowano z rozkładu Poissona dla $\lambda = 1$, 50 obserwacji z rozkładu Poissona dla $\lambda = 5$ i 50 obserwacji z rozkładu Poissona dla $\lambda = 12$. Oznacza to, że cały zbiór danych pochodzi z następującej mieszanki rozkładów:

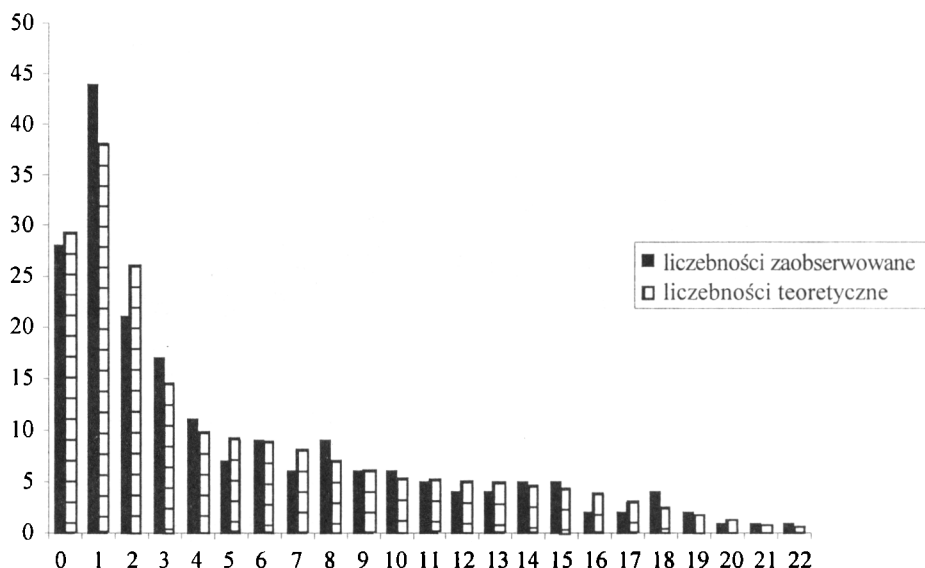
$$Y_i \sim \frac{1}{2} Po(1) + \frac{1}{4} Po(5) + \frac{1}{4} Po(12). \quad (21)$$

Histogram wygenerowanych danych przedstawia rys. 2.

Estymacja parametrów mieszanki wykonana została na podstawie 100 000 iteracji opisanego algorytmu. Rozkład *a posteriori* liczby składników mieszanki przedstawiony został w tab. 2.

Tabela 2. Rozkład liczby składników mieszanki

k	1	2	3	4	5	6
$p(k)$	0	0,0012	0,94745	0,04942	0,00179	0,00014



Rys. 2. Histogram danych wygenerowanych zgodnie z rozkładem (21) oraz wyestymowanej mieszanki trzech rozkładów Poissona

Źródło: opracowanie własne.

Najbardziej prawdopodobna liczba składników mieszanki jest zatem równa 3.

Estymatory parametrów rozkładu uzyskane dla najbardziej prawdopodobnego wariantu trzech składników są następujące:

$$\hat{\lambda}_1 = 1,29,$$

$$\hat{\lambda}_2 = 6,21,$$

$$\hat{\lambda}_3 = 13,92,$$

$$\hat{p}_1 = 0,53,$$

$$\hat{p}_2 = 0,26,$$

$$\hat{p}_3 = 0,21.$$

7.3. Estymacja parametrów mieszanki dwóch rozkładów Poissona dla małej próby

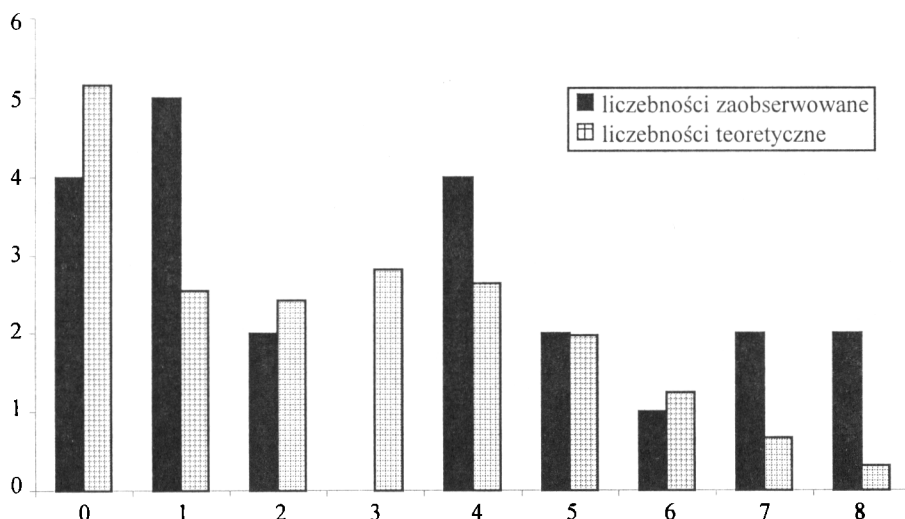
W przykładzie wykorzystano dane uzyskane za pomocą generatora liczb losowych. 10 obserwacji wylosowano z rozkładu Poissona dla $\lambda = 1$ oraz 10 obserwacji z rozkładu Poissona dla $\lambda = 5$. Oznacza to, że cały zbiór danych pochodzi z następującej mieszanki rozkładów:

$$Y_i \sim \frac{1}{2} Po(1) + \frac{1}{2} Po(5). \quad (22)$$

Estymacja parametrów mieszanki wykonana została na podstawie 100 000 iteracji opisanego algorytmu. Rozkład *a posteriori* liczby składników mieszanki przedstawiony został w tab. 3.

Tabela 3. Rozkład liczby składników mieszanki

k	1	2	3	4	5	6
$p(k)$	0,24187	0,68693	0,06398	0,0063	0,00087	0,00005



Rys. 3. Histogram danych wygenerowanych zgodnie z rozkładem (22) oraz wyestymowanej mieszanki dwóch rozkładów Poissona

Źródło: opracowanie własne.

Na podstawie danych z tab. 3 można zauważyć, że wyniki są mniej jednoznaczne niż w przykładzie z punktów 7.1 i 7.2.

Estymatory parametrów rozkładu uzyskane dla najbardziej prawdopodobnego wariantu dwóch składników są następujące:

$$\hat{\lambda}_1 = 0,28, \hat{\lambda}_2 = 3,76, \hat{p}_1 = 0,32, \hat{p}_2 = 0,68.$$

8. Podsumowanie

Opisana bayesowska procedura wyboru liczby składników mieszanki rozkładu Poissona umożliwia wybór modelu, który w kontekście zebranych danych jest najbardziej prawdopodobny. W niektórych sytuacjach prawdopodobieństwa kilku modeli mogą być do siebie zbliżone. Może to oznaczać, że zebrane dane są niewystarczające liczne. Istnieje również możliwość, że zastosowany model mieszanki rozkładów Poissona jest nieadekwatny i być może lepszym rozwiązaniem jest zastosowanie mieszanki innego typu.

Konieczność ustalenia liczby składników mieszanki jest zagadnieniem często wstępującym w praktyce. Jednym z zagadnień, w którym możliwe jest zastosowanie opisaney procedury, jest estymacja liczby szkód w portfelu ubezpieczeń.

Literatura

- [1] Dellaportas P., Karlis D., Xekalaki E., *Bayesian Analysis of Finite Poisson Mixtures*, Technical Report Athens University of Economics and Business, 1997.
- [2] Green P., *Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination*, „Biometrika” 1995, nr 82, 711–732.
- [3] Han C., Carlin B.P., *MCMC Methods for Computing Bayes Factors: a Comparative Review*, JASA, 96, 1122–1132, 2001.
- [4] *Highly Structured Stochastic Systems*, P.J. Green, N. Hjort, S. Richardson (red.), University Press, Oxford 2003.
- [5] Robert C.P., Casella G., *Monte Carlo Statistical Methods*, Second Edition, Springer 2004.
- [6] Viallefont V., Richardson S., Green P.J., *Bayesian Analysis of Poisson Mixtures*, „Journal of Nonparametric Statistics” 2002 vol. 14, iss. 1-2, s. 181–202.

BAYESIAN PROCEDURE FOR NUMBER OF COMPONENTS CHOICE IN A MIXTURE OF DISTRIBUTIONS

Summary

The paper describes the Bayesian procedure for the selection of the number of components in a mixture. The numerical implementation involves the Reversible Jump Markov Chain Monte Carlo algorithm (RJMCMC).

The Bayesian approach assumes that the model is treated as an unknown parameter over which we can assume a probability distribution. The model and in this case, the number of components in a mixture, is selected according to the criterion of the maximum posterior probability.

The first part of the paper presents an analytical solution of the problem for the simplified example of a mixture of binomial distributions. Then the RJMCMC algorithm is described and its implementation for the problem of the choice of the number of components for the Poisson mixture.

Dariusz Biskup – dr, adiunkt w Katedrze Statystyki Akademii Ekonomicznej we Wrocławiu.