PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1169 2007
TAKSONOMIA 14
Klasyfikacja i analiza danych – teoria i zastosowania

Roman Huptas
Akademia Ekonomiczna w Krakowie

ESTYMACJA PARAMETRÓW ROZKŁADU NA PODSTAWIE DANYCH POGRUPOWANYCH ZA POMOCĄ ALGORYTMU EM. WYNIKI BADAŃ EMPIRYCZNYCH

1. Wstęp

Algorytm EM jest metodą iteracyjnego obliczania estymatorów największej wiarygodności (ENW) [Magiera 2002, s. 146], stosowaną do rozwiązywania problemów określanych jako problemy z niekompletnymi danymi. W każdej iteracji są wykonywane dwa kroki: krok E (*expectation step*) i krok M (*maximization step*). Z tego też powodu algorytm jest nazywany algorytmem EM [Dempster, Laird, Rubin 1977]. Na problemy z niekompletnymi danymi składają się: struktury z brakującymi danymi, modele z obciętymi rozkładami, pogrupowane czy też ocenzurowane obserwacje, a także model efektów losowych, mieszanki rozkładów, estymacja komponentów wariancyjnych, iteracyjna ważona metoda najmniejszych kwadratów, modele logarytmiczno-liniowe.

W artykule jako zastosowanie algorytmu EM przeanalizowano problem estymacji parametrów rozkładu dla danych pogrupowanych. Algorytm EM został ukazany w świetle klasycznych metod estymacji parametrów rozkładu. Przeprowadzone eksperymenty symulacyjne miały na celu zbadanie wpływu sposobu estymacji parametrów rozkładu na wyniki zastosowań testu χ^2 zgodności. Przedstawiono wyniki eksperymentów symulacyjnych, które pozwoliły porównać ze sobą założone i rzeczywiste rozmiary testu χ^2 Pearsona dla przypadku, gdy nieznane parametry estymowane są z użyciem algorytmu EM na podstawie zgrupowanych danych oraz gdy parametry estymowane są na podstawie oryginalnych, niezgrupowanych danych. W niniejszej pracy zaprezentowano także wyniki badań przeprowadzonych dla danych rzeczywistych.

2. Opis algorytmu EM

Niech $\mathbf{Y}$ będzie wektorem losowym modelującym obserwacje niekompletne, o funkcji gęstości względem miary Lebesgue'a $g(\mathbf{y}; \Psi)$, gdzie $\Psi = (\Psi_1, ..., \Psi_d)^T$ jest

wektorem nieznanych parametrów z przestrzeni parametrów $\boldsymbol{\Omega}$. Funkcja wiarygodności dla $\boldsymbol{\Psi}$ przy zadanym wektorze obserwacji $\mathbf{y}$ ma postać $L(\boldsymbol{\Psi}) = g(\mathbf{y};\boldsymbol{\Psi})$. Estymator największej wiarygodności jest definiowany jako rozwiązanie $\tilde{\boldsymbol{\Psi}}$ równania wiarygodności $\frac{\partial}{\partial \boldsymbol{\Psi}} L(\boldsymbol{\Psi}) = 0$ lub $\frac{\partial}{\partial \boldsymbol{\Psi}} \ln L(\boldsymbol{\Psi}) = 0$, w którym osiągane jest globalne maksimum $L(\boldsymbol{\Psi})$ (por. [Magiera 2002, s. 167]).

Niech $\mathbf{x}$ będzie wektorem danych kompletnych. Wektor danych obserwowanych $\mathbf{y}$ jest wówczas traktowany zarówno jako wektor danych niekompletnych, jak i funkcja danych kompletnych, tzn. $\mathbf{y} = \mathbf{y}(\mathbf{x})$ [Dempster, Laird, Rubin 1977, s. 1].

Niech $\mathbf{X}$ będzie wektorem losowym modelującym dane kompletne i mającym funkcję gęstości względem miary Lebesgue'a $g_c(\mathbf{x};\Psi)$. Wtedy logarytm funkcji wiarygodności nieobserwowanych danych kompletnych ma postać $\ln L_c(\boldsymbol{\Psi}) = \ln g_c(\mathbf{x};\boldsymbol{\Psi})$.

Algorytm EM rozwiązuje równanie wiarygodności dla niekompletnych danych pośrednio, wykorzystując logarytm funkcji wiarygodności kompletnych danych $\ln L_c(\boldsymbol{\Psi})$. Oczywiście, funkcja $\ln L_c(\boldsymbol{\Psi})$ jest nieobserwowalna. Jest ona zastępowana przez jej warunkową wartość oczekiwaną względem wektora $\mathbf{Y}$, z użyciem bieżącej wartości parametru $\boldsymbol{\Psi}$.

Kroki E i M w (k+1)-szej iteracji są zdefiniowane następująco [McLachlan, Krishnan 1997, s. 22]:

Krok E. Obliczenie $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(k)})$, gdzie $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(k)}) = E_{\boldsymbol{\Psi}^{(k)}}\{\ln L_c(\boldsymbol{\Psi}) \mid \mathbf{Y} = \mathbf{y}\}$, a $\boldsymbol{\Psi}^{(k)}$ oznacza wartość parametru $\boldsymbol{\Psi}$ uzyskaną w k-tej iteracji algorytmu EM.

Krok M. Maksymalizacja $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(k)})$ względem $\boldsymbol{\Psi}$. Wybieramy zatem $\boldsymbol{\Psi}^{(k+1)} \in \boldsymbol{\Omega}$ takie, że dla wszystkich $\boldsymbol{\Psi} \in \boldsymbol{\Omega}$: $Q(\boldsymbol{\Psi}^{(k+1)}, \boldsymbol{\Psi}^{(k)}) \geq Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(k)})$.

Dla zadanego $\varepsilon > 0$ kroki E i M są powtarzane do momentu, gdy po raz pierwszy $\ln L(\boldsymbol{\Psi}^{(k+1)}) - \ln L(\boldsymbol{\Psi}^{(k)}) < \varepsilon$ lub inne kryterium zatrzymania zostanie spełnione, np. liczba iteracji osiągnie zadaną z góry wartość maksymalną.

3. Estymacja parametrów rozkładu na podstawie danych pogrupowanych i obciętych

Zagadnienie pogrupowanych i obciętych danych to problem z niekompletnymi danymi, do którego można zastosować algorytm EM [McLachlan, Krishnan 1997, s. 74; Dempster, Laird, Rubin 1977, s. 13].

Niech W będzie zmienną losową z przestrzeni $\mathcal{W}$ o funkcji gęstości $f(w;\boldsymbol{\Psi})$, gdzie $\boldsymbol{\Psi}$ jest wektorem nieznanych parametrów. Niech przestrzeń $\mathcal{W}$ będzie podzielona na v rozłącznych klas $\mathcal{W}_j$ $(j = 1,\ldots,v)$. Realizacje zmiennej losowej W

są niezależne, ale nie są rejestrowane. Jedynie liczby n_j tych obserwacji, które należą do klasy $\mathcal{W}_j$ dla $j=1,\ldots,r$, gdzie $r \le v$, są rejestrowane.

Niech wektor $\mathrm{y}=(n_1,\ldots,n_r)^T$ będzie wektorem danych niekompletnych (zaobserwowanych) oraz $n := n_1 + \ldots + n_r$.

Przy ustalonym n wektor y pochodzi z rozkładu wielomianowego o rozmiarze próby n z r kategoriami, a prawdopodobieństwa wystąpienia kategorii wynoszą $p_j(\Psi)/p(\Psi)$ $(j=1,\ldots,r)$, gdzie $p_j(\Psi)=\int_{\mathcal{W}_j} f(w;\Psi)\,dw$ i $p(\Psi)=\sum_{j=1}^{r} p_j(\Psi)$.

Funkcja wiarygodności dla niekompletnych danych ma postać:

$$L(\Psi)=\frac{n!}{\prod_{j=1}^{r} n_j!}\prod_{j=1}^{r}\left(\frac{p_j(\Psi)}{p(\Psi)}\right)^{n_j}. \tag{1}$$

Na potrzeby algorytmu EM wprowadzono wektor „brakujących" danych, tzn. $\mathrm{z}=(n_{r+1},\ldots,n_v)^T$ – wektor nieobserwowanych częstości w przypadku obciętych danych $(r<v)$ oraz $\mathrm{w}_j=(w_{j1},\ldots,w_{jn_j})^T$ $(j=1,\ldots,v)$ – wektor n_j nieobserwowanych realizacji zmiennej losowej W, które należą do klasy $\mathcal{W}_j$ dla $j=1,\ldots,v$.

Niech teraz wektor danych kompletnych ma postać $\mathbf{x}=(\mathbf{y}^T, \mathbf{z}^T, \mathbf{w}_1^T,\ldots,\mathbf{w}_v^T)^T$. Wektor losowy $\mathbf{Y}$ danych niekompletnych będzie miał rozkład wielomianowy, a tym samym funkcję wiarygodności określoną wzorem (1), jeśli wektor losowy $\mathbf{X}$ danych kompletnych będzie miał rozkład, dla którego funkcja wiarygodności będzie miała postać (z dokładnością do mnożnika niezależnego od Ψ):

$$L_c(\Psi)=\prod_{j=1}^{v}\prod_{k=1}^{n_j} f(w_{jk};\Psi), \tag{2}$$

gdzie obserwacje w_{jk} $(j=1,\ldots,v, k=1,\ldots,n_j)$ są realizacjami zmiennej losowej W dla próby o rozmiarze $n+m$, a $m=n_{r+1}+\ldots+n_v$ i jest losowe.

Konstrukcja rozkładu brakujących danych taka, aby otrzymać (2), przedstawiona jest np. w pracy [Huptas 2006].

***Krok E* dla (*k* + 1)-szej iteracji:** obliczamy $\bar{Q}(\Psi,\Psi^{(k)})=E_{\Psi^{(k)}}\{\ln L_c(\Psi)\,|\,\mathbf{Y}=\mathbf{y}\}$. Niech N_j będzie zmienną losową oznaczającą liczbę obserwacji w klasie $\mathcal{W}_j$. Ze względu na konstrukcję wektorów w_j $(j=1,\ldots,v)$, z i y, otrzymujemy (por. [McLachlan, Krishnan 1997, s. 77]):

$$\bar{Q}(\Psi,\Psi^{(k)})=\sum_{j=1}^{v} n_j^{(k)}\,Q_j(\Psi,\Psi^{(k)})+\ln\frac{(m+n-1)!\,n}{\prod_{j=1}^{v} n_j^{(k)}!}, \tag{3}$$

gdzie:

$$Q_j(\Psi,\Psi^{(k)}) = E_{\Psi^{(k)}}\{\ln f(W;\Psi) \mid W \in \mathcal{W}_j\},$$

$$n_j^{(k)} = E_{\Psi^{(k)}}\{N_j \mid \mathbf{Y}=\mathbf{y}\} = \begin{cases} n_j & \text{dla} \quad j=1,\ldots,r \\ n\, p_j(\Psi^{(k)}) / p(\Psi^{(k)}) & \text{dla} \quad j=r+1,\ldots,v \end{cases}.$$

Drugi człon równania (3) nie zależy od Ψ i może być opuszczony przy maksymalizacji $\bar{Q}(\Psi,\Psi^{(k)})$. Ostatecznie można zatem przyjąć, że:

$$Q(\Psi,\Psi^{(k)}) = \sum_{j=1}^{v} n_j^{(k)}\, Q_j(\Psi,\Psi^{(k)}).$$

***Krok M* dla (*k*+1)-szej iteracji**: maksymalizujemy $Q(\Psi,\Psi^{(k)})$ względem Ψ. Wartość parametru $\Psi^{(k+1)}$ będzie pierwiastkiem równania:

$$\partial Q(\Psi,\Psi^{(k)}) / \partial \Psi = 0,$$

gdzie

$$\frac{\partial}{\partial \Psi} Q(\Psi,\Psi^{(k)}) = \sum_{j=1}^{v} n_j^{(k)} \frac{\partial}{\partial \Psi} Q_j(\Psi,\Psi^{(k)}),$$

$$\frac{\partial}{\partial \Psi} Q_j(\Psi,\Psi^{(k)}) = E_{\Psi^{(k)}}\left\{\frac{\partial}{\partial \Psi} \ln f(W;\Psi) \mid W \in \mathcal{W}_j\right\}.$$

4. Eksperyment symulacyjny

Przeprowadzono eksperyment symulacyjny, w którym zastosowano algorytm EM do estymacji parametrów rozkładu dla danych pogrupowanych. Wykorzystano test χ^2 zgodności, w którym statystyka testowa ma asymptotyczny rozkład χ^2 [Magiera 2002, s. 244]. Test był przeprowadzany, gdy rozkład teoretyczny postulowany w hipotezie zerowej był ciągły i zależny od nieznanych parametrów (przyjęto rozkład normalny z nieznanymi: μ i σ^2). W celu oszacowania parametrów dla testu stosuje się dwie metody:

(a) oszacowaniami parametrów są rozwiązania układu równań wiarygodności dla obserwacji zgrupowanych (ENW są oparte na częstościach empirycznych),

(b) estymatorami parametrów są rozwiązania układu równań wiarygodności dla niezgrupowanych danych (ENW są oparte na oryginalnych obserwacjach).

Niech p_j będą częstościami teoretycznymi v klas w przypadku, gdy rozkład teoretyczny jest w pełni określony. Wówczas statystyka χ^2 Pearsona ma przy $n \to \infty$ rozkład graniczny χ^2 z $v-1$ stopniami swobody.

Niech teraz $\tilde{p}_j$ będą ENW częstości teoretycznych p_j, opartymi na częstościach empirycznych n_j (odpowiada to metodzie (a)). Wtedy, przy odpowiednich

warunkach regularności, statystyka $\tilde{\chi}^2$ Pearsona ma przy $n \to \infty$ rozkład graniczny χ^2 z $v-1-k$ stopniami swobody, gdzie k jest liczbą estymowanych parametrów.

Niech $\hat{p}_j$ będą ENW częstości teoretycznych p_j, uzyskanymi dla niezgrupowanych danych, czyli opartymi na oryginalnych obserwacjach (odpowiada to metodzie (b)). Okazuje się, że wtedy rozkład graniczny przy $n \to \infty$ statystyki testowej $\hat{\chi}^2$ leży między rozkładami granicznymi statystyk $\tilde{\chi}^2$ i χ^2. Statystyka testowa $\hat{\chi}^2$ ma więc pewien rozkład, którego kwantyle zawierają się między odpowiednimi kwantylami rozkładów χ^2_{v-k-1} i χ^2_{v-1}. Należy nadmienić, że wraz ze wzrostem liczby v klas wartości kwantyli rozkładów χ^2_{v-k-1} i χ^2_{v-1} (przy tym samym poziomie istotności α) coraz mniej się różnią.

Celem eksperymentu symulacyjnego było porównanie, przy założonych poziomach istotności, empirycznych rozmiarów testów χ^2 Pearsona uzyskanych w przypadkach, gdy nieznane parametry estymowane były na podstawie zgrupowanych danych z użyciem algorytmu EM oraz danych oryginalnych niezgrupowanych.

Estymatory były oparte na tych samych n elementowych próbach. Obliczenia były powtarzane $N = 10\,000$ razy, za każdym razem dla różnego zestawu danych.

Wzory na estymatory parametrów μ i σ^2 w przypadku algorytmu EM mają postać:

$$\mu^{(k+1)} = \frac{\sum_{j=1}^{v} n_j \, E_{\Psi^{(k)}}\left\{ W \mid W \in \mathcal{W}_j \right\}}{\sum_{j=1}^{v} n_j}, \tag{4}$$

$$(\sigma^2)^{(k+1)} = \frac{\sum_{j=1}^{v} n_j \, E_{\Psi^{(k)}}\left\{ (W - \mu^{(k+1)})^2 \mid W \in \mathcal{W}_j \right\}}{\sum_{j=1}^{v} n_j}. \tag{5}$$

Z kolei estymatory w przypadku danych oryginalnych obliczano następująco:

$$\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \tilde{\mu})^2 . \tag{6}$$

Rozmiary empiryczne obliczono jako $\alpha_{emp} = N_k / N$, gdzie N_k to liczba tych wartości statystyk testowych, które wpadły do obszaru krytycznego, a N oznacza liczbę powtórzeń symulacji. Za błąd symulacji przyjęto górne oszacowanie $\sqrt{Var(N_k/N)}$.

Tabele 1 i 2 prezentują wyniki uzyskane dla $N = 10\,000$ (błąd symulacji wynosił 0,004), liczności próby $n = 500$ i liczby klas $v \in \{4; 12\}$. Więcej szczegółów dotyczących eksperymentu symulacyjnego można znaleźć w pracy [Huptas 2006].

Tabela 1. Empiryczne rozmiary testów dla $v = 4$, $n = 500$ i $N = 10\,000$

Poziom istotności α	Empiryczny rozmiar testu α_{emp}	
	algorytm EM	momenty z próby
0,010	0,010	0,015
0,050	0,054	0,084
0,100	0,104	0,173

Źródło: obliczenia własne.

Tabela 2. Empiryczne rozmiary testów dla $v = 12$, $n = 500$ i $N = 10\,000$

Poziom istotności α	Empiryczny rozmiar testu α_{emp}	
	algorytm EM	momenty z próby
0,010	0,010	0,011
0,050	0,052	0,055
0,100	0,102	0,108

Źródło: obliczenia własne.

Eksperyment pokazał, że przy estymacji nieznanych parametrów na podstawie funkcji wiarygodności po zgrupowaniu danych empiryczne rozmiary testów są równe, w granicach błędu symulacji, założonym poziomom istotności niezależnie od liczby klas i liczności próbki. W estymacji parametrów na podstawie oryginalnych obserwacji empiryczne rozmiary testów odbiegają znacznie od założonych poziomów istotności, gdy liczba klas jest mała. Wzrost liczby klas powoduje, że empiryczne rozmiary testów coraz mniej się różnią od siebie.

5. Analiza danych rzeczywistych

Punktem wyjścia do badań empirycznych były eksperymenty symulacyjne opisane w punkcie 4. Celem badań było sprawdzenie za pomocą testu zgodności χ^2, czy badana próba pochodzi z rozkładu normalnego.

Dane rzeczywiste pochodzą z Głównego Urzędu Statystycznego i dotyczą wydatków gospodarstw domowych na osobę. Wstępna analiza danych pokazała, że dane mają rozkład logarytmiczno-normalny, więc do analizy dane zlogarytmowano. Próba, którą badano, liczyła $n = 506$ przypadków. Klasy zostały dobrane tak, aby liczebności teoretyczne były większe od 5 (warunek $np_j \geq 5\,(j = 1,\ldots,v)$ [Zeliaś, Pawełek, Wanat 2002]. Przedziały dolny i górny zostały domknięte z wykorzystaniem informacji o minimalnej i maksymalnej wartości w badanej próbie.

Tabele 3 i 4 prezentują wyniki analizy danych rzeczywistych w przypadkach, gdy skrajne klasy miały postacie: [5,1;5,8] i (7,8;8,5] oraz [5,1;5,4] i (7,8;8,5], a pozostałe $v-2$ klasy miały tę sama długość. Estymację wartości oczekiwanej i wariancji przeprowadzono dwiema metodami: algorytmem EM (4)-(5) oraz odpowiednimi momentami z próby (6).

Tabela 3. Wartości estymatorów i statystyk testowych dla skrajnych klas o postaci: [5,1; 5,8] i (7,8; 8,5]

Liczba klas	$\chi^2_{\nu-3;\alpha}$	Wartości estymatorów i statystyk testowych	
		algorytm EM	momenty z próby
$\nu = 4$	$\chi^2_{1;0.01} = 6,635$ $\chi^2_{1;0.05} = 3,841$ $\chi^2_{1;0.1} = 2,706$	$\mu_{EM} = 6,557262$ $\sigma^2_{EM} = 0,321128$ $\chi^2_{EM} = 3,171$	$\tilde{\mu} = 6,557755$ $\tilde{\sigma}^2 = 0,313856$ $\tilde{\chi}^2 = 3,904$
$\nu = 8$	$\chi^2_{5;0.01} = 15,086$ $\chi^2_{5;0.05} = 11,070$ $\chi^2_{5;0.1} = 9,236$	$\mu_{EM} = 6,553567$ $\sigma^2_{EM} = 0,313854$ $\chi^2_{EM} = 4,671$	$\tilde{\mu} = 6,557755$ $\tilde{\sigma}^2 = 0,313856$ $\tilde{\chi}^2 = 4,723$
$\nu = 12$	$\chi^2_{9;0.01} = 21,666$ $\chi^2_{9;0.05} = 16,919$ $\chi^2_{9;0.1} = 14,684$	$\mu_{EM} = 6,551611$ $\sigma^2_{EM} = 0,307080$ $\chi^2_{EM} = 7,043$	$\tilde{\mu} = 6,557755$ $\tilde{\sigma}^2 = 0,313856$ $\tilde{\chi}^2 = 6,930$

Źródło: obliczenia własne.

Dla liczby klas $\nu = 8$ i $\nu = 12$ przy skrajnych klasach równej i różnej długości (tab. 3 i 4) statystyki testowe χ^2_{EM} i $\tilde{\chi}^2$ są zdecydowanie mniejsze od wartości krytycznych i nieznacznie różnią się od siebie. Na wszystkich założonych poziomach istotności nie mamy podstaw do odrzucenia hipotezy zerowej, mówiącej o zgodności rozkładu badanej próby z rozkładem normalnym.

Tabela 4. Wartości estymatorów i statystyk testowych dla skrajnych klas o postaci: [5,1; 5,4] i (7,8; 8,5]

Liczba klas	$\chi^2_{\nu-3;\alpha}$	Wartości estymatorów i statystyk testowych	
		algorytm EM	momenty z próby
$\nu = 4$	$\chi^2_{1;0.01} = 6,635$ $\chi^2_{1;0.05} = 3,841$ $\chi^2_{1;0.1} = 2,706$	$\mu_{EM} = 6,564769$ $\sigma^2_{EM} = 0,315862$ $\chi^2_{EM} = 3,531$	$\tilde{\mu} = 6,557755$ $\tilde{\sigma}^2 = 0,313856$ $\tilde{\chi}^2 = 3,695$
$\nu = 8$	$\chi^2_{5;0.01} = 15,086$ $\chi^2_{5;0.05} = 11,070$ $\chi^2_{5;0.1} = 9,236$	$\mu_{EM} = 6,551458$ $\sigma^2_{EM} = 0,305467$ $\chi^2_{EM} = 6,274$	$\tilde{\mu} = 6,557755$ $\tilde{\sigma}^2 = 0,313856$ $\tilde{\chi}^2 = 6,046$
$\nu = 12$	$\chi^2_{9;0.01} = 21,666$ $\chi^2_{9;0.05} = 16,919$ $\chi^2_{9;0.1} = 14,684$	$\mu_{EM} = 6,549336$ $\sigma^2_{EM} = 0,305083$ $\chi^2_{EM} = 7,109$	$\tilde{\mu} = 6,557755$ $\tilde{\sigma}^2 = 0,313856$ $\tilde{\chi}^2 = 6,838$

Źródło: obliczenia własne.

W świetle wyników eksperymentu symulacyjnego najbardziej interesujące były jednak wartości statystyk testowych dla małej liczby klas. Dla liczby klas $\nu = 4$ przy równych skrajnych klasach (tab. 3) hipotezę zerową na poziomie istotności $\alpha = 0{,}05$ odrzucamy w estymacji parametrów na podstawie oryginalnych, niezgrupowanych obserwacji. Natomiast nie mamy podstaw do jej odrzucenia w przypadku estymacji nieznanych parametrów algorytmem EM. Statystyka testowa χ^2_{EM} jest znacznie mniejsza od statystyki testowej $\tilde{\chi}^2$. Dla liczby klas $\nu = 4$ przy skrajnych klasach różnej długości (tab. 4) sytuacja taka nie ma miejsca, tzn. na poziomie istotności $\alpha = 0{,}05$ w obu przypadkach nie mamy podstaw do odrzucenia hipotezy zerowej. Należy jednak nadmienić, że i tym razem statystyka testowa χ^2_{EM} jest mniejsza od statystyki testowej $\tilde{\chi}^2$.

6. Podsumowanie

Eksperyment symulacyjny pokazał, że przy małej liczbie klas empiryczne rozmiary testów w estymacji parametrów na podstawie oryginalnych danych odbiegają znacznie od założonych poziomów istotności. W przypadku estymacji parametrów algorytmem EM dla zgrupowanych danych zjawisko takie nie miało miejsca, niezależnie od liczby klas. Wyniki badań empirycznych potwierdziły, że sposób estymacji parametrów istotnie wpływa na wyniki testu zgodności rozkładu. Jeżeli więc liczba ν klas jest duża i estymujemy małą liczbę parametrów metodą największej wiarygodności bez grupowania obserwacji, to przy określaniu obszaru krytycznego można korzystać z wartości kwantyli rozkładu χ^2 z $\nu - 1 - k$ stopniami swobody. W przeciwnym razie rzeczywisty rozmiar testu może być istotnie większy niż zakładany poziom istotności.

Literatura

Dempster A.P., Laird N.M, Rubin D.B. (1977), *Maximum Likelihood from Incomplete Data via the EM Algorithm (with Discussion)*, „Journal of the Royal Statistical Society B" 39, s. 1-38.

Huptas R. (2006), *Zastosowanie algorytmu EM do estymacji parametrów rozkładu na podstawie danych pogrupowanych*, Zeszyty Naukowe AE w Krakowie (w druku).

Magiera R. (2002), *Modele i metody statystyki matematycznej*, wydanie I, GiS, Wrocław.

McLachlan G.J., Krishnan T. (1997), *The EM Algorithm and Extensions*, John Wiley and Sons Inc., New York.

Zeliaś A., Pawełek B., Wanat S. (2002), *Metody statystyczne. Zadania i sprawdziany*, PWE, Warszawa.

ESTIMATION OF PARAMETERS OF DISTRIBUTION FROM GROUPED DATA BY MEANS OF THE EM ALGORITHM. RESULTS OF EMPIRICAL RESEARCHES

Summary

In this article the Expectation-Maximization (EM) algorithm and its application are presented. The application is related to estimation of parameters of distribution in case data are grouped and may also be truncated. Results of a simulation experiment in which was used the Pearson chi-square goodness of fit test and results of empirical researches were presented. The EM algorithm as an iterative technique of estimation of parameters and classical methods of estimation were compared.