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

Robert Kapłon
Politechnika Wrocławska

O LICZBIE KLAS W MODELU ANALIZY CZYNNIKOWEJ Z DWOMA ZMIENNYMI UKRYTYMI

1. Wstęp

W praktyce dość często mamy do czynienia z danymi, których charakter wskazuje na pochodzenie z wielu populacji. Zamiast więc budować jeden, ogólny model dla całej populacji, przyjmuje się, że model ten jest mieszaniną modeli pochodzących z podpopulacji[1]. Wykorzystując takie podejście, należy zmierzyć się z problemem wyboru liczby klas (podpopulacji). Niestety, w tej kwestii nie ma jednoznacznych rozstrzygnięć, gdyż podejście zmierzające do weryfikacji odpowiednich hipotez – w którym jako statystykę testową wykorzystuje się podwojony iloraz logarytmu funkcji wiarygodności – nie może być stosowane. Wynika to z niespełnienia warunków regularności, co w konsekwencji prowadzi do nieznajomości rozkładu przyjętej statystyki (por. [McLachlan, Peel 2000]). Można jednak aproksymować ten rozkład, wykorzystując podejście bootstrapowe [McLachlan 1987]. Należy pamiętać, że wadą podejścia bootstrapowego jest czasochłonność obliczeń. Dlatego próbuje się wykorzystać pewne syntetyczne mierniki oceny modeli, jak np. bayesowskie kryterium informacyjne, kryterium Akaikego.

Choć badań w zakresie przydatności syntetycznych mierników oceny jest dużo, to w głównej mierze dotyczą one modeli opartych na mieszankach rozkładów normalnych. Przedmiotem zainteresowania, jak do tej pory, nie był model analizy czynnikowej z dwoma zmiennymi ukrytymi. Dlatego celem niniejszego opracowania jest przeprowadzenie badań symulacyjnych, które staną się podstawą do sformułowania rekomendacji dotyczących wiarygodności rozważanych mierników.

[1] O celowości wprowadzenia dodatkowej zmiennej ukrytej do modelu analizy czynnikowej pisze Kapłon [2004].

2. Model analizy czynnikowej i algorytm estymacji parametrów

Kapłon [2004] przedstawił ogólny model analizy czynnikowej z dwoma zmiennymi ukrytymi wraz z procedurą estymacji. Model ten będzie podstawą przeprowadzanych badań symulacyjnych. Aby jednak skrócić czas symulacji, zrezygnowano z indeksacji wektora obserwacji kolejnymi markami oraz przyjęto, że macierz diagonalna wariancji specyficznej jest taka sama w każdej klasie. Uwzględniając to, model można przedstawić następująco.

Dane są wektory obserwacji $\mathbf{a}_1,\ldots,\mathbf{a}_N$ oraz wektory zmiennych ukrytych $\mathbf{b}_1,\ldots,\mathbf{b}_N$ i $\zeta_1,\ldots,\zeta_C$ takie, że $\mathbf{a}_i \in R^p$ oraz $\mathbf{b}_i \in R^q$, dla obserwacji $i=1,\ldots,N$. Przypisując wektor obserwacji $\mathbf{a}_i$ do określonej klasy ukrytej C_c, model analizy czynnikowej ze zmiennymi ukrytymi można wyrazić w postaci:

$$\mathbf{a}_i = \boldsymbol{\mu}_c + \boldsymbol{\Lambda}_c \mathbf{b}_i + \mathbf{e}_i.$$

Jeżeli obserwacja i należy do klasy C_c, to:

$$(\mathbf{b}_i \mid i \in C_c) \sim \mathcal{N}_q(\mathbf{0},\mathbf{I}),$$

$$(\mathbf{e}_i \mid i \in C_c) \sim \mathcal{N}_p(\mathbf{0},\boldsymbol{\Psi}),$$

$$(\mathbf{a}_i \mid \mathbf{b}_i; i \in C_c) \sim \mathcal{N}_p(\boldsymbol{\mu}_c + \boldsymbol{\Lambda}_c \mathbf{b}_i, \boldsymbol{\Psi}),$$

gdzie: $\boldsymbol{\Lambda}_c$ jest macierzą $p \times q$ ładunków czynnikowych w klasie C_c $(c=1,\ldots,C)$, $\boldsymbol{\Psi}$ jest $p \times p$ macierzą diagonalną wariancji specyficznej, $\boldsymbol{\mu}_c$ reprezentuje p-wymiarowy wektor średnich w klasie C_c, $\mathbf{e}_i$ jest p-wymiarowym czynnikiem losowym (czynnik swoisty).

Dla tak sformułowanego modelu wartości estymatorów w iteracji $t+1$ mają postać (por. [Kapłon 2004]:

$$\boldsymbol{\mu}_c^{(t+1)} = \sum_i \tau_{ic}^{(t)} \left(\mathbf{a}_i - \boldsymbol{\Lambda}_c^{(t)} \mathbf{x}_{ic}^{(t)}\right) \left[\sum_{i'} \tau_{i'c}^{(t)}\right]^{-1},$$

$$\boldsymbol{\Lambda}_c^{(t+1)} = \sum_i \tau_{ic}^{(t)} \left(\mathbf{a}_i - \boldsymbol{\mu}_c^{(t+1)}\right) \mathbf{x}_{ic}^{T(t)} \left[\sum_i \tau_{ic}^{(t)} \mathbf{Y}_{ic}^{(t)}\right]^{-1},$$

$$\boldsymbol{\Psi}^{(t+1)} = \operatorname{diag}\left[\sum_{i,c} \tau_{ic}^{(t)} \left(\mathbf{a}_i - \boldsymbol{\mu}_i^{(t+1)}\right)\left(\mathbf{a}_i^T - \boldsymbol{\mu}_c^{T(t+1)} - \mathbf{x}_{ic}^{T(t)} \boldsymbol{\Lambda}_c^{T(t+1)}\right)\right] N^{-1},$$

$$p_c^{(t+1)} = \sum_i \tau_{ic}^{(t)} N^{-1},$$

gdzie:

$$\mathbf{x}_{ic}^{(t)} = \boldsymbol{\Lambda}_c^{(t)T} \boldsymbol{\Gamma}_c^{-1(t)} \left(\mathbf{a}_i - \boldsymbol{\mu}_c^{(t)}\right) \text{ i } \boldsymbol{\Gamma}_c = \boldsymbol{\Lambda}_c \boldsymbol{\Lambda}_c^T + \boldsymbol{\Psi},$$

$$\mathbf{Y}_{ic}^{(t)} = \mathbf{I} - \boldsymbol{\Lambda}_c^{T(t)} \boldsymbol{\Gamma}_c^{-1(t)} \boldsymbol{\Lambda}_c^{(t)} + \mathbf{x}_{ic}^{(t)} \mathbf{x}_{ic}^{T(t)},$$

natomiast τ_{ic} jest prawdopodobieństwem *a posteriori* przynależności do klasy, wyznaczonym zgodnie z regułą Bayesa.

3. Kryteria wyboru liczby klas

W badaniu wykorzystano trzy grupy kryteriów. Pierwsze z nich to kryteria wywodzące się z tzw. szkoły Akaikego. Ich podstawą jest informacja Kulbacka-Leiblera (K-L), która mówi o przeciętnym stopniu rozbieżności między prawdziwym i przyjętym rozkładem. Informacja K-L nie jest bezpośrednio obserwowana, gdyż zależy od prawdziwego rozkładu oraz nieznanych parametrów w obu rozkładach. Dlatego najpierw maksymalizuje się średni poziom logarytmu wiarygodności, a następnie koryguje go się tak, aby otrzymać asymptotycznie nieobciążony estymator oczekiwanej informacji K-L (por. [Bozdogan 2000]). Okazuje się jednak, że w wielu wypadkach wyznaczenie dokładnej wartości tej korekty (obciążenia) jest praktycznie niemożliwe i stąd pojawiają się jej rozmaite aproksymacje prowadzące do różnych kryteriów.

Najczęściej wykorzystywane jest kryterium informacyjne Akaikego (*AIC*), które ma postać (por. [Akaike 1973])

$$AIC = -2\log L(\hat{\Theta}) + 2r,$$

gdzie $L(\hat{\Theta})$ jest funkcją wiarygodności, a liczbę parametrów modelu oznaczono jako r.

Główny zarzut, jaki się stawia kryterium *AIC*, to pewna tendencja do wyboru modeli, które są nazbyt rozbudowane. Dzieje się tak nawet, gdy próba jest duża. Wynika to z niewrażliwości kary ($2r$) na wzrost liczebności próby. Aby zapewnić asymptotyczną zgodność, Bozdogan [1987] proponuje pewną modyfikację (*CAIC – consistent AIC*):

$$CAIC = -2\log L(\hat{\Theta}) + r\log N + r.$$

Druga grupa kryteriów wyrosła z koncepcji podejścia bayesowskiego. Podstawą porównania dwóch modeli jest czynnik bayesowski. Tutaj również pojawiają się trudności w obliczeniu tego czynnika, dlatego przyjmuje się jego oszacowanie. Najbardziej chyba znana jest aproksymacja Schwarza [1978], która prowadzi do tzw. bayesowskiego kryterium informacyjnego (*BIC*):

$$BIC = -2\log L(\hat{\Theta}) + r\log N.$$

Kryterium to, jako aproksymacja składowej czynnika bayesowskiego, zakłada pewien szczególny rozkład *a priori* parametrów – rozkład, który zawiera tyle samo informacji, co pojedyncza obserwacja. Prowadzi to do płaskiego rozkładu (*spread out*), co niektórzy uznają za wadę [Weakliem 1999]. Jeśli brak jest informacji (wiedzy) o rozkładach *a priori*, to przyjęcie takiego rozkładu wydaje się rozsądnym rozwiązaniem (por. [Raftery 1999]).

Kryterium *BIC* nie uwzględnia poprawności klasyfikacji, dlatego ma tendencje do przeszacowywania liczby klas, bez względu na stopień ich separacji (por. [Biernacki, Celeux, Govaert 2000]). Dlatego istotnym uzupełnieniem wcześniej rozważanych mierników są kryteria klasyfikacyjne. Opierają się one na entropii, której estymator ma postać

$$EN(\hat{\boldsymbol{\tau}}) = -\sum_{i}\sum_{c} \hat{\tau}_{ic} \log \hat{\tau}_{ic},$$

i przyjmuje największą wartość równą zeru. Jeśli podział nie jest jednoznaczny, to nie ma nakładających się klas, wtedy *EN* będzie bliskie zeru.

Sama entropia nie może być podstawą wyboru liczby klas, gdyż zawsze przyjmie wartość większą dla tego z modeli, który ma ich więcej. W związku z tym Celeux i Soromenho [1996] proponują normalizację (*NEC*):

$$NEC = \frac{EN(\hat{\boldsymbol{\tau}})}{\log L_c(\hat{\boldsymbol{\Theta}}) - \log L_1(\hat{\boldsymbol{\Theta}})}.$$

W mianowniku występuje różnica między funkcjami wiarygodności: weryfikowanego modelu o liczbie klas równiej c i modelu z jedną klasą. Jak można zauważyć, kryterium to nie daje rozstrzygnięcia, gdy należy dokonać wyboru między modelem o jednej klasie a modelem o dwóch klasach.

Kolejnym kryterium, które zostanie wykorzystane, jest *CLC* (*classification likelihood information*). Entropia pełni tutaj funkcję kary nałożonej na logarytm wiarygodności [McLachlan, Peel 2000]:

$$CLC = -2\log L(\hat{\boldsymbol{\Theta}}) + 2EN(\hat{\boldsymbol{\tau}}).$$

Kryterium *CLC* daje dobre wyniki, jeśli wielkości klas są takie same. Ma jednak tendencję do przeszacowywania prawdziwej liczby klas, jeśli nie nałoży się ograniczenia odnośnie do ich wielkości [Biernacki, Govaert 1999; Biernacki, Celeux, Govaert 2000].

Kryteria informacyjne nie nakładają kary na modele nazbyt rozbudowane, tak jak ma to miejsce w *BIC*. Dlatego proponuje się kryterium *ICL BIC*, które łączy kryterium bayesowskie i klasyfikacyjne:

$$ICL\ BIC = -2\log L(\boldsymbol{\Theta}) + 2EN(\hat{\boldsymbol{\tau}}) + r\log N.$$

W tym miejscu warto odnotować, że z dwóch konkurencyjnych modeli wybiera się ten, dla którego obliczone kryterium informacyjne przyjmuje najmniejszą wartość.

4. Opis eksperymentu

W eksperymencie uwzględniono cztery czynniki (w nawiasie podano ich poziomy), mogące mieć decydujący wpływ na wartości rozważanych kryteriów: liczbę klas (dwie, trzy), liczebność próby (duża, średnia), wielkość klas (podobna, różna) oraz stopień zależności między zmiennymi (duży, średni, mały). Należy wspomnieć również o liczbie zmiennych i liczbie czynników wspólnych – przyjęto, że

Tabela 1. Składowe macierzy eksperymentu oraz wartości kryteriów oceny, gdy prawdziwy model ma 2 klasy

	Liczebność		Korelacje		AIC			BIC			CAIC			CLC			ICL BIC			NEC		
Eks.	kl. 1	kl. 2	mała	duża	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1	480	120	0	0.9	0	49	1	0	**50**	0	0	**50**	0	0	**50**	0	0	**50**	0	0	**50**	0
2	300	300	0	0.9	0	50	0	0	**50**	0	0	**50**	0	0	44	6	0	**50**	0	0	**50**	0
3	160	40	0	0.9	0	48	2	0	**50**	0	0	**50**	0	0	9	41	0	**50**	0	0	**50**	0
4	100	100	0	0.9	0	50	0	0	**50**	0	0	**50**	0	0	0	50	0	**50**	0	0	**50**	0
5	480	120	0.2	0.8	0	48	2	0	**50**	0	0	**50**	0	0	2	48	0	**50**	0	0	**50**	0
6	300	300	0.2	0.8	0	**50**	0	0	**50**	0	0	**50**	0	0	42	8	0	**50**	0	0	**50**	0
7	160	40	0.2	0.8	0	**50**	0	0	**50**	0	0	**50**	0	0	5	45	0	**50**	0	0	**50**	0
8	100	100	0.2	0.8	0	**50**	0	0	**50**	0	0	**50**	0	0	0	50	0	**50**	0	0	**50**	0
9	480	120	0.3	0.6	0	**50**	0	0	**50**	0	0	**50**	0	0	**50**	0	0	**50**	0	0	**50**	0
10	300	300	0.3	0.6	0	48	2	0	**50**	0	0	**50**	0	0	47	3	0	**50**	0	0	**50**	0
11	160	40	0.3	0.6	0	1	49	0	45	5	0	**49**	1	0	0	50	0	48	2	0	**50**	0
12	100	100	0.3	0.6	0	15	35	0	**50**	0	0	**50**	0	0	0	50	0	**50**	0	0	**50**	0

Źródło: obliczenia własne.

Tabela 2. Składowe macierzy eksperymentu oraz wartości kryteriów oceny, gdy prawdziwy model ma 3 klasy

	Liczebność			Korelacje		AIC			BIC			CAIC			CLC			ICL BIC			NEC		
Eks.	kl. 1	kl. 2	kl. 3	mała	duża	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4
1	315	315	270	0	0.9	0	**50**	0	0	**50**	0	0	**50**	0	0	45	5	0	**50**	0	30	20	0
2	630	135	135	0	0.9	0	48	2	0	**50**	0	0	**50**	0	0	**50**	0	0	**50**	0	33	17	0
3	405	360	135	0	0.9	0	49	1	0	**50**	0	0	**50**	0	0	45	5	0	**50**	0	30	20	0
4	105	105	90	0	0.9	0	**50**	0	0	**50**	0	0	**50**	0	0	0	50	0	**50**	0	31	19	0
5	210	45	45	0	0.9	0	49	1	0	**50**	0	0	**50**	0	0	38	12	0	**50**	0	25	25	0
6	135	120	45	0	0.9	0	49	1	0	**50**	0	0	**50**	0	0	3	47	0	**50**	0	32	18	0
7	315	315	270	0.2	0.8	0	**50**	0	0	**50**	0	0	**50**	0	0	41	9	0	**50**	0	50	0	0
8	630	135	135	0.2	0.8	0	**50**	0	0	**50**	0	0	**50**	0	0	**50**	0	0	**50**	0	50	0	0
9	405	360	135	0.2	0.8	0	48	2	0	**50**	0	0	**50**	0	0	47	3	0	**50**	0	50	0	0
10	105	105	90	0.2	0.8	0	49	1	0	**50**	0	0	**50**	0	0	1	49	0	**50**	0	48	2	0
11	210	45	45	0.2	0.8	0	47	3	0	**50**	0	0	**50**	0	0	40	10	0	**50**	0	48	2	0
12	135	120	45	0.2	0.8	0	**50**	0	0	**50**	0	0	**50**	0	0	1	49	0	**50**	0	50	0	0
13	315	315	270	0.3	0.6	0	**50**	0	0	**50**	0	0	**50**	0	0	49	1	0	**50**	0	50	0	0
14	630	135	135	0.3	0.6	0	49	1	0	**50**	0	0	**50**	0	0	**50**	0	0	**50**	0	50	0	0
15	405	360	135	0.3	0.6	0	49	1	0	**50**	0	0	**50**	0	0	**50**	0	0	**50**	0	50	0	0
16	105	105	90	0.3	0.6	0	**49**	1	21	29	0	43	7	0	0	0	50	23	27	0	50	0	0
17	210	45	45	0.3	0.6	0	**43**	7	11	39	0	39	11	0	0	25	25	12	38	0	50	0	0
18	135	120	45	0.3	0.6	0	**49**	1	3	47	0	24	26	0	0	0	50	6	44	0	50	0	0

Źródło: obliczenia własne.

liczba zmiennych będzie zawsze taka sama i równa 10, podobnie jak liczba czynników wspólnych ustalona na poziomie 4.

Duża liczebność próby to 300 obserwacji przypadających na jedną klasę, z kolei średnia liczebność to 100 obserwacji. Jeśli więc badano model z 2 klasami, to li-

czebność całej próby wyniosła 600 lub 200 obserwacji odpowiednio dla dużej i średniej liczebności.

Przyjęto relatywnie dużą różnicę między klasami. W modelu z 2 klasami wyniosła ona 60%. W wypadku trzech klas dwie z nich mają zbliżoną wielkość i różnią się od trzeciej. Procentowo przedstawia się to następująco: 70, 15 i 15% oraz 45, 45 i 15%.

W modelu analizy czynnikowej poziom korelacji ma wpływ na wyodrębniane czynniki wspólne. Te zmienne, które są najmocniej skorelowane, będą tworzyć czynnik wspólny. Zmienne te powinny być również słabo skorelowane z pozostałymi. W eksperymencie poddano to kontroli, ustalając trzy pary korelacji: (0,0.9), (0.2,0.8) oraz (0.3,0.6). Większa wartość oznacza korelację między zmiennymi, które powinny tworzyć czynnik wspólny – jeśli go nie tworzą, to korelacja jest tą wartością mniejszą.

Po zastosowaniu powyższych ustaleń otrzymano 12 i 18 wariantów modeli odpowiednio dla 2 i 3 klas. Szczegółowo przedstawiono to w tab. 1 i 2.

Dla każdego modelu (w sumie 30) wygenerowano obserwacje z wielowymiarowego rozkładu normalnego. Następnie oszacowano parametry: modelu prawdziwego oraz modeli z jedną klasą więcej i jedną klasą mniej od modelu prawdziwego. Procedurę powtórzono 50 razy, co dało 30x3x50 = 4500 wariantów modeli. Dla każdego modelu obliczono kryteria, które porównano między klasami. Model, dla którego dane kryterium miało najniższą wartość, uważany był za najlepszy.

5. Wnioski

Wśród kryteriów klasyfikacyjnych najwięcej wątpliwości budzi kryterium *NEC*. W wypadku modelu z trzema klasami najlepsze wyniki odnotowano wtedy, gdy pod względem korelacji były największe różnice (tab. 2). Niestety, nie są one jednak zadowalające, gdyż poprawność wyboru właściwego modelu wyniosła ok. 40% (średnio 20/50). W pozostałych wypadkach, gdy różnice w korelacji były mniejsze, poza dwoma wyjątkami, za każdym razem wybierany był model z dwoma klasami. Daje to podstawę, by sądzić, że kryterium *NEC* nie jest wiarygodne. Należy jeszcze odnotować, że dla prawdziwego modelu z dwoma klasami w porównaniu nie brał udziału model o jednej klasie. Mając to na względzie oraz bardzo wyraźną skłonność do preferowania przez to kryterium modeli mniej złożonych, staje się jasne, dlaczego poprawność wskazań wyniosła 100% (tab. 1).

Pewne wątpliwości budzi również kryterium *CLC*. Nie radzi sobie ono kompletnie ze wskazaniem prawdziwego modelu, jeśli liczebność próby jest średnia, a wielkość klas zbliżona. Może więc dziwić, że duże różnice w rozmiarze klas poprawiają zdolność wyboru odpowiedniego modelu (tab. 1 i 2).

Zdecydowanie lepsze rezultaty osiągnęło kryterium Akaikego. Prawie dla każdego modelu poprawność wskazań była bliska 50. Jedynie dla eksperymentu 11 i

12 modelu z dwoma klasami osiągnięto niezadowalające wyniki. W tym wypadku potwierdza się ogólna opinia o *AIC*, że ma ono tendencję do wyboru modeli zbyt rozbudowanych.

Modyfikacja *AIC* w kierunku zapewnienia asymptotycznej zgodności pozwoliła uzyskać jeszcze lepsze rezultaty. Kryterium *CAIC* prawie w każdym eksperymencie właściwie wskazywało liczbę klas modelu prawdziwego. Podobnie było w wypadku kryterium bayesowskiego *BIC* oraz mieszanego *ICL BIC*. Gdy różnice w korelacji między zmiennymi były najmniejsze oraz liczebność próby była średnia (eksperyment 16, 17 i 18), wtedy te trzy kryteria – dla modelu z trzema klasami – rzadziej wskazywały poprawny model. Znacznie lepiej wypadło tutaj kryterium *AIC*.

W konkluzji należy stwierdzić, że dość wiarygodnych przesłanek do wyboru właściwego modelu dostarczają kryteria: *CAIC*, *BIC*, *ICL BIC*. Można mieć wątpliwości w sytuacji, w której liczebność próby jest na poziomie średnim, a różnica w korelacji między zmiennymi jest nieduża. Trudno jednoznacznie powiedzieć, czy różnice między wielkością klas mają jakiś wpływ na kształtowanie się kryteriów wyboru. Dlatego wydaje się, że pożądane byłyby pogłębione badania w tym zakresie.

Literatura

Akaike H. (1973), *Information Theory and an Extension of the Maximum Likelihood Principle*, [w:] B.N. Petrov, F. Csaki (red.), *Second International Symposium on Information Theory* (pp.), Academiai Kiado, Budapest, s. 267-281.

Biernacki C., Celeux G., Govaert G. (2000), *Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood*, „IEEE Transactions on Pattern Analysis and Machine Intelligence" 22 (7), s. 719-725.

Biernacki C., Govaert G., (1999) *Choosing Models in Model-Based Clustering and Discriminant Analysis*, „Journal of Statistical Computation and Simulation" 14, s. 49-71.

Bozdogan H. (1987), *Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions*, „Psychometrika" 52, s. 345--370.

Bozdogan H. (2000), *Akaike's Information Criterion and Recent Developments in Information Complexity*, „Journal of Mathematical Psychology" 44, s. 62-91.

Celeux G., Soromenho G. (1996), *An Entropy Criterion for Assessing the Number of Clusters in a Mixture Model*, „Classification Journal" 13, s. 195-212.

Kapłon R. (2004), *Estymacja parametrów modelu czynnikowego wykorzystującego klasy ukryte*, [w:] K. Jajuga, M. Walesiak (red.), *Klasyfikacja i analiza danych – teoria i zastosowania, Taksonomia 11*, Prace Naukowe AE we Wrocławiu nr 1022, AE, Wrocław.

Kass R.E., Raftery A.E. (1995), *Bayes Factors,* „Journal of the American Statistical Association" 90, s. 773-795.
McLachlan, G.J. (1987), *On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture,* „Journal of the Royal Statistical Society Series C (Applied Statistics)" 36, s. 318-324.
McLachlan G.J., Peel. D. (2000), *Finite Mixture Models*, Wiley, New York.
Raftery A.E. (1999), *Bayes Factors and BIC – Comment on "A Critique of the Bayesian Information Criterion for Model Selection"*, „Sociological Methods and Research" 27, s. 411-427.
Weakliem D.L. (1999), *A Critique of the Bayesian Information Criterion for Model Selection*, „Sociological Methods and Research" 27, s. 359-397.

DETERMINING THE NUMBER OF CLUSTERS IN FACTOR ANALYSIS WITH TWO LATENT VARIABLES

Summary

When the number of components in mixture models is not known, the problem arise. The most natural and obvious approach seems to be testing a hypothesis. Unfortunately with mixture models regularity conditions do not hold therefore minus twice the log likelihood ratio test statistic does not follow the usual chi-squared distribution. Some authors suggest using bootstrap for assessing the null distribution of that statistic. But this method is computationally very intensive. For these reasons synthetic measures are often used.

So far, many empirical comparisons have been made of these various criteria to compare the performance of the selection of the number of components in Gaussian mixture. Because factor analysis with two latent variables has not been in these comparisons therefore the aim of this paper is to fill the gap.