응용통계연구 제 7권 1호
1994년 3월 pp. 131 - 140

정규 분포의 혼합성 판단기준

홍종선[1], 최병수[2], 엄종석[3]

요약

패턴분류대상이 되는 자료로부터 패턴식별함수를 찾고 패턴분류에 적용시키기 위하여, 우선 이러한 자료에 몇 개의 패턴이 존재하는지를 알아야 하는 문제점에 대하여 고려하였다. 수집된 자료가 단일분포로부터 발생하였는지 또는 두 개의 혼합분포로부터 발생되었는지를 판단하는 기준을 표본분산의 특성값인 분산의 부트스트랩 추정값 (bootstrap estimate) 으로 설정하였다. 새로운 판단기준의 특성에 대하여 살펴보고 시뮬레이션을 통하여 판단기준의 적합성을 연구하였다.

1. 서론

혼합분포 (mixture distribution) 문제는 Pearson (1894) 이래로 여러 학문 분야, 특히 컴퓨터 분야의 패턴인식 (pattern recognition) 등의 다양한 분야의 많은 연구자들의 관심의 대상이 되었다. 이 분야의 지금까지의 연구의 대부분은 혼합분포모형의 모수 추정에 관심을 두어 왔다. 혼합 분포의 모수를 추정하는 방법으로는 적률법 (method of moments <Derin (1987) 와 기타 논문>), 최우법 (maximum likelihood methods <Basford and McLachlan (1985) 와 기타 논문>), 그래픽 기법 (graphical approach <Postaire and Vasseur (1981) 와 기타 논문>) 등이 있다. 또한, 다변량 Gaussian 혼합분포의 모수를 추정하는 방법으로는 Fukunaga and Flick (1983) 등이 있고, 결정 이론적 접근 (decision directed approach) 으로는 Kazakos and Davission (1980) 등이 있고 Young and Coraluppi (1970)는 확률적 근사 알고리즘을 제시하였다.

이러한 연구들의 목적은 패턴 (pattern) 대상이 되는 자료로부터 패턴 식별함수 (recognition function) 를 찾고 이 식별함수를 패턴분류 (pattern classification) 에 적용하는 것이다. 그러므로 혼합분포 식별에 대한 이론은 패턴인식분야의 중요한 기초 이론중의 하나이다.

패턴 식별함수를 찾기 위해서는 우선, 이러한 자료에 몇 개의 패턴이 존재하는 지를 먼저 알아야 하는 문제점이 있다. 또한 패턴의 혼합을 전제로 하기 때문에 혼합이 되지 않은 패턴에 적용시키는 것은 무리가 있다. 그러므로 본 연구에서는 모수 추정에 앞서 패턴분류의 대상이 되는 자료가 단일 정규분포로부터 발생하였는지, 또는 두 분포들의 선형결합으로 구성된 혼합분포로부터 발생되었는지를 판단하는 기준을 마련하는데 그 목적이 있다. 그리고 이 판단기준은 모수 추정방법과 더불어 패턴분류에 많은 활용이 가능하다.

본 연구의 구성은 다음과 같다. 2절에서는 기존의 접근과는 다른 새로운 방법으로 판단 기준을 설정하고, 판단기준의 특성에 대하여 토론하고자 한다. 그리고, 시뮬레이션을 통해 이 판

1) (110-745) 서울특별시 종로구 명륜동 3가 53 성균관대학교 경상대학 통계학과 부교수.
2) (136-792) 서울특별시 성북구 삼선동 2가 389 한성대학 전산통계학과 조교수
1) (136-792) 서울특별시 성북구 삼선동 2가 389 한성대학 전산통계학과 조교수

단기준의 타당성을 3절에서 검토한다. 또한 이러한 방법을 이용하여 자료의 정규성(normality) 검정을 위한 새로운 방법을 제안하였다.

2. 혼합분포 판단기준

확률변수 X 는 패턴대상이 되는 다음과 같은 혼합분포를 따른다고 하자.

$$f(x)=\alpha f_1(x)+(1-\alpha)f_2(x), \qquad x\in(-\infty,\ \infty), \tag{1}$$

여기서 $\alpha\in(0,1)$는 혼합비율 (mixture proportion) 이고, $i=1,2$에 대하여 $f_i(\cdot)$는 평균이 $\mu_i\in(-\infty,\infty)$ 이고, 분산은 동일한 $\sigma^2\in(0,\infty)$ 인 정규분포를 따른다고 가정하자. 혼합분포에 관한 대부분의 연구는 (1)식에 존재하는 모수들의 추정에 대하여 이루어졌으며, 특히 혼합비율의 추론을 통하여 분포의 혼합성을 판단하였다. 그러나 패턴대상이 되는 자료에 몇 개의 패턴이 존재하는 지를 먼저 알아야 하는 문제가 있으므로 모수 추정에 앞서 자료가 단일분포에서 추출되었는지 또는 혼합분포에서 추출되었는지를 판단하고자 한다.

(1)식의 모형을 따르는 분포로부터 크기 n의 관찰값 $(x_1,x_2,\cdots,x_n)$ 이 있다고 하자. 이 표본이 상이한 확률밀도함수 $f_1(\cdot)$과 $f_2(\cdot)$의 혼합분포에서 추출되었는지 또는 $f_1(\cdot)$과 $f_2(\cdot)$가 동일한 밀도함수의 조건에서, 즉, 단일분포에서 추출되었는가의 여부를 판단하는 새로운 기준 (criterion) 을 설정하자.

우선 일반성을 잃지 않고 그리고 토론의 간편성을 위하여 확률변수 X를 평균과 분산에 대하여 다음과 같이 표준화한 확률변수 Z를 고려하자.

$$Z_i = \frac{(X_i-E(X))}{\sqrt{V(X)}}, \tag{2}$$

여기서 $E(X) = \alpha\mu_1 + (1-\alpha)\mu_2$,

$V(X) = \alpha(1-\alpha)(\mu_1-\mu_2)^2 + \sigma^2$.

새로운 판단기준 C^2은 다음에 따라서 구한다.

1. 표준화된 표본 $(z_1,z_2,\cdots,z_n)$ 을 구한다.

   여기서 $z_i = \dfrac{(x_i-\bar{x})}{\sqrt{(\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1)}}$.

2. 표준화된 표본으로부터 크기 n의 부트스트랩 표본 (bootstrap sample) $(y_1,y_2,\cdots,y_n)$을 취한다.

3. 표본 $(y_1,y_2,\cdots,y_n)$에서 표본분산 $S^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2/(n-1)$

   을 구한다.

4. 2번과 3번을 매우 큰 수인 B번 반복하여 $S_1^2,S_2^2,\cdots,S_B^2$ 을

구하고 이들을 통하여 C^2 을 다음과 같이 얻는다:

$$\widehat{C^2} = \mathrm{Var}(S^2)_B = \frac{1}{B-1}\sum_{j=1}^{B}(S_j^2 - \overline{S^2})^2, \tag{3}$$

여기서 $\overline{S^2} = \sum_{j=1}^{B} S_j^2/B$.

이와 같은 방법으로 정의된 새로운 판단기준 C^2은 표본분산의 분산에 대한 부트스트랩 추정값 (bootstrap estimate) 이며 (Efron (1982) 과 Efron and Tibshirani (1986) 참조), 표본 $(x_1,x_2,\cdots,x_n)$이 단일분포 ($H_0 : \mu_1 = \mu_2$), 또는 혼합분포 ($H_1 : \mu_1 \neq \mu_2$) 에서 추출되었는지를 판단하는 기준으로 이용할 수 있겠다. 여기서 우리는 귀무가설 하에서 C^2의 기대값을 구하여 보자.

$$\begin{aligned} E(C^2) &= \mathrm{Var}(S^2) \\ &= \frac{2}{n-1} \end{aligned} \tag{4}$$

이것은 이미 잘 알려져 있다. 그리고 C^2의 분포에 관한 성격에 대하여 알아보면 다음과 같다.

정리 확률표본 X1,⋯,Xn 이 (1)식의 모형을 따르는 혼합분포로부터 추출된 것일때, $E(C^2)$은 $\alpha \in \left[\frac{3-\sqrt{3}}{6}, \frac{3+\sqrt{3}}{6}\right]$이면 $\left|\frac{\mu_1-\mu_2}{\sigma}\right|$의 단조비증가함수, 그외의 경우에서는 $\left|\frac{\mu_1-\mu_2}{\sigma}\right|$의 단조증가함수이다.

증명 (3)식에서 정의된 C^2의 기대값은

$$E(C^2) = \frac{1}{n}\left(\mu_4(0) - \frac{n-3}{n-1}\right), \tag{5}$$

여기서 $\mu_4(0) = E(Z^4)$이다. 따라서 우선 (2)식에서 표준화된 확률변수 Z의 적률모함수를 구하여 보자.

$$\begin{aligned} M_Z(t) &= \exp\left[\frac{-(\alpha\mu_1+(1-\alpha)\mu_2)t}{\sqrt{(\alpha(1-\alpha)(\mu_1-\mu_2)^2+\sigma^2)}}\right] \cdot M_X\left(\frac{t}{\sqrt{\alpha(1-\alpha)(\mu_1-\mu_2)^2+\sigma^2}}\right) \\ &= \alpha\exp\left[\frac{\sigma^2t^2}{2[\alpha(1-\alpha)(\mu_1-\mu_2)^2+\sigma^2]} + \frac{(1-\alpha)(\mu_1-\mu_2)t}{\sqrt{\alpha(1-\alpha)(\mu_1-\mu_2)^2+\sigma^2}}\right] \\ &+ (1-\alpha)\exp\left[\frac{\sigma^2t^2}{2[\alpha(1-\alpha)(\mu_1-\mu_2)^2+\sigma^2]} - \frac{\alpha(\mu_1-\mu_2)t}{\sqrt{\alpha(1-\alpha)(\mu_1-\mu_2)^2+\sigma^2}}\right] . \end{aligned}$$

그리고 적률모함수를 통하여 $\mu_4(0)$를 구하면 다음과 같음을 알 수 있다.

$$\mu_4(0)=\frac{3\sigma^4+6\alpha(1-\alpha)\sigma^2(\mu_1-\mu_2)^2+\alpha(1-\alpha)(1-3\alpha(1-\alpha))(\mu_1-\mu_2)^4}{[\alpha(1-\alpha)(\mu_1-\mu_2)^2+\sigma^2]^2}\ .$$

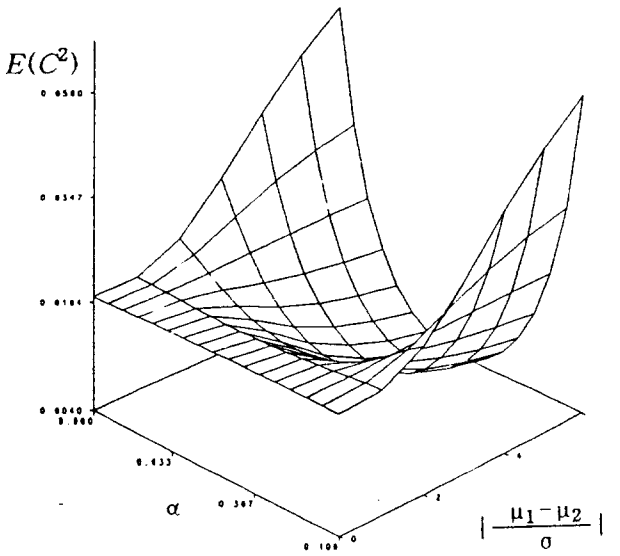
위 식의 우변을 $|(\mu_1-\mu_2)/\sigma|$에 관하여 일차 미분하면, $|(\mu_1-\mu_2)/\sigma|$ 와 α 의 함수가 되며 다음과 같음을 알 수 있다 :

$$\frac{d}{d\left|\frac{\mu_1-\mu_2}{\sigma}\right|}\mu_4(0)\ \le\ 0,\quad \text{만약 } \alpha\in\left[\frac{3-\sqrt{3}}{6},\ \frac{3+\sqrt{3}}{6}\right]$$

$$\frac{d}{d\left|\frac{\mu_1-\mu_2}{\sigma}\right|}\mu_4(0)\ >\ 0,\quad \text{그 외의 경우.}$$

한편, $E(C^2)$는 (5)식에 표현되었듯이 $\mu_4(0)$의 일차함수이므로 위 정리는 증명되었다. (*)

위 정리에서 언급한 $E(C^2)$의 값은 아래 <그림>에서 나타나고 있다.



< 그림 > $E(C^2)$ 의 값

위와 같이 부트스트랩 추정값 C^2은 혼합비율 α와 표준화된 모평균의 차 $|(\mu_1-\mu_2)/\sigma|$의 변화에 따라 뚜렷한 현상의 변화를 발견할 수 있으므로 우리는 C^2을 혼합분포의 판단기준으로 설정할 수 있다.

3. 시뮬레이션 결과

자료 $(x_1,\cdots,x_n)$가 단일정규분포로부터 추출되었는지 혼합정규분포로부터 추출되었는지를 결정하기 위한 판단기준 C^2의 검정력에 관심을 두고 시뮬레이션을 하였다. 이를 위하여 먼저

H_0 : 표본은 단일정규분포에서 추출되었다. $(\mu_1=\mu_2)$

라는 귀무가설 하에서 $E(C^2)$의 95% 신뢰구간을 구하고자 한다. 여기서 신뢰구간은 백분위방법 (percentile method) 을 사용하였다 (Efron (1982) 의 10.4절 참고). 즉 표준정규분포 N(0,1)으로부터 표본크기 n을 100으로 부트스트랩 표본추출횟수 B 를 1,000으로 하는 C^2값을 10,000회 계산하여 95% 백분위 신뢰구간 (percentile confidence interval) 을 구하였다. 여기서, random number의 발생은 32bit 혼합식 합동법 (mixed congruential method)을 정규난수는 Porlar Marsaglia method를 사용하였다. 시뮬레이션 프로그램은 C언어로 작성되었다. 그 결과는 다음과 같다.

$E(C^2)$ 의 95% 백분위 신뢰구간 :
(LB= 0.01263 , UB= 0.02995).

다음으로 설정된 판단기준 C^2의 검정력을 조사하여 보자. 대립가설을

H_1 : 표본은 혼합정규분포로부터 추출되었다. $(\mu_1\neq\mu_2)$

라고 설정하여 n = 100, B = 1,000 그리고 혼합비율 α = {0.1, 0.2,⋯,0.9} 와 표준화된 모평균 차이, $|(\mu_1-\mu_2)/\sigma|$ = {0.0, 0.5, 1.0, ⋯, 6.0}에 대하여 각각 1,000회의 판단기준 C^2을 2절에서 언급한 절차에 의하여 계산하여 그 평균과 표준편차를 <표 1>에 나타내고 있다. <표 1>에 나타난 현상은 앞 절의 <그림>과 유사한 형태로 나타남을 알 수 있었고 다만 $|(\mu_1-\mu_2)/\sigma|$ 일 때의 C^2의 평균값들은 (4)식에 정리한 귀무가설 하에서 C^2의 기대값 0.02020 보다 작음을 알 수 있다. 이것은 2절에서 언급한 부트스트랩 추정값을 사용해서 발생한 부트스트랩 편의 (bootstrap bias) 이다. 그리고 <표 1>을 얻기 위하여 반복한 부트스트랩 표본추출횟수 1,000번 중 C^2값이 LB=0.01263 보다 작아서 귀무가설이 기각된 횟수, UB=0.02995 보다 커서 귀무가설이 기각된 횟수, 그리고 C^2값이 앞에서 설정한 신뢰구간에 포함되어 귀무가설을 채택한

횟수들은 <표 2>에 보이고 있다. 즉, <표 2>에서 '기각 1' ('기각 2')은 C^2값이 LB (UB)보다 작아서 (커서) H_0가 기각되는 횟수를, '채택'은 C^2값이 UB와 LB사이에 존재하는 횟수를 나타내준다. 귀무가설을 기각하는 검정력은 <표 2>의 각 칸(cell)에서 '기각 1'과 '기각 2'를 더한 빈도 수를 1,000번 반복한 수로 나눈 값이다. <표 2>에서 검정력이 어느 특정한 적은 수준 (예를 들어 0.20) 미만인 범위를 굵은 선 안쪽에 표시하였다. 그리고 <표 2>의 결과를 자세히 살펴보면, α가 0.1 또는 0.9 일 때 C^2값이 UB보다 커서 귀무가설을 기각하는 횟수가 $|(\mu_1-\mu_2)/\sigma|$의 크기에 따라 증가함을 알 수 있고, α가 0.3 부터 0.7 까지의 구간에서는 C^2값이 LB보다 작아서 귀무가설을 기각하는 횟수가 $|(\mu_1-\mu_2)/\sigma|$의 크기에 따라 증가함을 알 수 있다. 다만 <정리>에서 언급한 $(3\pm\sqrt{3})/6$에 가까운 α=0.2 또는 0.8의 경우에는 $|(\mu_1-\mu_2)/\sigma|$의 증감에 따라 유의한 변화가 없음을 알 수 있다. 따라서 <정리>과 시뮬레이션 결과를 바탕으로 다음과 같은 결론을 유도할 수 있다.

1. 판단기준 C^2값이 LB=0.01263 보다 작을 경우 :
   혼합비율이 0.5에 가까운 값을 가지면서
   혼합분포이다.

2. 판단기준 C^2값이 UB=0.02995 보다 클 경우 :
   혼합비율이 0.5에서 멀리 떨어진 값을 가지면서
   혼합분포이다.

3. 그 외의 C^2값인 경우 :
   결정을 내릴 수 없다.

4. 결론

패턴분류대상이 되는 자료가 단일분포에서 추출되었는지 또는 두 개의 분포로 이루어진 혼합분포에서 추출되었는지를 판단하는 새로운 판단기준 C^2을 부트스트랩 (bootstrap) 을 이용하여 설정하였다. " H_0 : 단일분포에서 발생된 자료이다." 라는 귀무가설 하에서 $E(C^2)$의 신뢰구간을 시뮬레이션을 통하여 구하였으며, C^2값을 이용하여 귀무가설을 기각하는 검정력을 구하였다. 그리고 표본으로부터 구한 C^2값으로 분포의 혼합성 판단기준을 유도하였다. 이 논문에서는 표본크기를 100으로 한정하여 연구하였지만 여러 종류의 표본크기에서도 비교연구가 되어야 하겠다. 또한 설정된 (1)번 모형식의 두 Gaussian 분포에서 분산의 등질성 (homogeneous variance)을 가정하였는데 이질성 (heterogeneity) 도 함께 고려하여야 할 과제이다.

그리고 (1)식에 설정된 모형에 포함된 모수들의 추정은 1절에서 언급하였듯이 많은 연구가 되었지만 본 연구의 <표 2>를 통하여 모수의 추정은 어느 정도 윤곽을 파악할 수 있다. 그러나 모수 추정과 더불어 발전된 식별함수를 찾아 패턴분류에 적용시키는 과제는 앞으로 계속 연구되어야 하겠다.

참고 문헌

[1] Basford, K. E. and McLachlan, G. J. (1985), Estimation of Allocation Rates in a cluster Analysis Context, *Journal of the American Statistical Association,* Vol. 80, 286-293.

[2] Derin, H. (1987), Estimating components of Univariate Gaussian Mixtures using Prony's method, *IEEE Transactions,* Vol. 74, 561-575.

[3] Efron, B. (1982), The Jacknife, the Bootstrap, and other Resampling Plans, *Society for Industrial and Applied Mathematics,* CBMS-NSF, Vol. 38.

[4] Efron, B. and Tibshirani, R. (1986), Bootstrap Methods for Standard Errors, Confidence Intervals, and other Measures of Statistical Accuracy, *Statistical Science,* Vol. 1, No. 1, 54-77.

[5] Fukunaga, K. and Flick, T. E. (1983), Estimating of Parameters of Gaussian Mixture of Two Normal(Lognormal) Distribution, *Journal of the American Statistical Association,* Vol. 74, 561-575.

[6] Kazakos, D. and Davisson, L. D. (1980), An Improved Decision-Directed Detector, *IEEE Transactions,* IT-26, 113-115.

[7] Pearson, K. (1984), On the Dissection of Frequency Curves into Normal Curves, *Philadelphia Transactions,* PAMI-3, 163-179.

[8] Postaire, J. G. and Vasseur, C. P. A. (1981), An Approximate Solution to Normal Mixture Identification with Application to Unsupervised Pattern Classification, *IEEE Transactions,* PAMI-3, 163-179.

[9] Young, T. Y. and Coraluppi, G. C. (1979), Stochastic Estimation of a Mixture of Normal Density Functions using an Information Criterion, *IEEE Transactions,* IT-16, 259-263.

<표 1> C^2의 평균과 표준편차 :
(각 칸의 값은 평균(상단)과 표준편차(하단)으로 이루어져 있다.)
(n = 100, B=1,000)
(반복횟수 = 1,000)

α / $\lvert\frac{\mu_1-\mu_2}{\sigma}\rvert$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.0	0.01913 0.00445	0.01923 0.00456	0.01907 0.00427	0.01928 0.00439	0.01899 0.00426	0.01940 0.00444	0.01918 0.00496	0.01911 0.00436	0.01905 0.00429
0.5	0.01899 0.00438	0.01931 0.00470	0.01900 0.00438	0.01911 0.00440	0.01925 0.00480	0.01905 0.00460	0.01900 0.00427	0.01921 0.00454	0.01905 0.00463
1.0	0.01931 0.00445	0.01936 0.00458	0.01874 0.00432	0.01861 0.00405	0.01863 0.00424	0.01839 0.00405	0.01897 0.00460	0.01946 0.00447	0.01936 0.00479
1.5	0.02039 0.00522	0.01944 0.00500	0.01814 0.00387	0.01719 0.00361	0.01693 0.00369	0.01729 0.00383	0.01808 0.00405	0.01962 0.00475	0.02058 0.00509
2.0	0.02207 0.00577	0.01983 0.00458	0.01707 0.00366	0.01533 0.00313	0.01479 0.00319	0.01531 0.00321	0.01689 0.00378	0.01971 0.00461	0.02273 0.00611
2.5	0.02549 0.00659	0.02046 0.00480	0.01588 0.00331	0.01339 0.00267	0.01245 0.00242	0.01333 0.00269	0.01569 0.00336	0.01999 0.00470	0.02545 0.00692
3.0	0.02889 0.00724	0.02104 0.00516	0.01479 0.00325	0.01156 0.00240	0.01050 0.00193	0.01149 0.00234	0.01447 0.00320	0.02037 0.00477	0.02914 0.00772
3.5	0.03265 0.00782	0.02154 0.00533	0.01383 0.00318	0.00999 0.00205	0.00878 0.00160	0.00990 0.00206	0.01364 0.00329	0.02083 0.00521	0.03277 0.00802
4.0	0.03666 0.00892	0.02154 0.00596	0.01325 0.00348	0.00867 0.00190	0.00747 0.00136	0.00865 0.00189	0.01303 0.00352	0.02142 0.00557	0.03651 0.00919
4.5	0.04075 0.01076	0.02214 0.00656	0.01253 0.00371	0.00786 0.00193	0.00635 0.00112	0.00764 0.00176	0.01218 0.00354	0.02181 0.00624	0.04098 0.01120
5.0	0.04452 0.01200	0.02266 0.00682	0.01184 0.00370	0.00709 0.00197	0.00548 0.00099	0.00684 0.00173	0.01153 0.00347	0.02232 0.00656	0.04415 0.01077
5.5	0.04842 0.01323	0.02273 0.00722	0.01161 0.00395	0.00634 0.00194	0.00474 0.00085	0.00624 0.00179	0.01102 0.00369	0.02217 0.00689	0.04781 0.01341
6.0	0.05140 0.01538	0.02336 0.00806	0.01107 0.00376	0.00586 0.00179	0.00423 0.00078	0.00568 0.00171	0.01111 0.00408	0.02243 0.00711	0.05090 0.01438

<표 2> C^2의 값이 각 영역에 해당되는 횟수 :

(각 칸의 값은 '기각 1'(상단)과 '채택'(중단) 그리고 '기각 2'(하단)로 이루어져 있다)

(n = 100, B=1,000)

(반복횟수 = 1,000)

α / $\lvert\frac{\mu_1-\mu_2}{\sigma}\rvert$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.0	31 944 25	34 941 25	24 956 20	32 942 26	28 950 22	20 956 24	25 946 29	28 950 22	29 953 18
0.5	28 948 24	29 947 24	28 952 20	33 944 23	30 933 37	32 947 21	28 954 18	34 938 28	31 944 25
1.0	30 942 28	29 946 25	37 942 21	41 945 14	50 933 17	34 952 14	36 935 29	32 942 26	34 940 26
1.5	18 938 44	28 943 29	42 947 11	77 920 3	86 909 5	74 916 10	37 950 13	18 948 34	21 941 38
2.0	8 894 98	24 943 33	80 916 4	172 827 1	259 740 1	195 804 1	99 893 8	26 941 33	4 879 117
2.5	3 786 211	9 948 43	153 846 1	451 549 0	596 404 0	438 562 0	196 802 2	23 949 28	2 798 200
3.0	1 607 392	11 935 54	254 745 1	718 282 0	865 135 0	723 277 0	294 706 0	20 940 40	0 606 394
3.5	0 410 590	16 914 70	411 589 0	890 110 0	981 19 0	904 96 0	420 580 0	26 927 47	0 395 605
4.0	0 227 773	18 897 85	492 507 1	971 29 0	999 1 0	965 35 0	514 485 1	27 897 76	0 243 757
4.5	0 142 858	34 856 110	573 425 2	985 15 0	1000 0 0	988 12 0	625 375 0	34 873 93	0 135 865
5.0	0 78 922	35 821 144	637 362 1	982 18 0	1000 0 0	989 11 0	673 327 0	22 855 123	0 75 925
5.5	0 46 954	31 815 154	685 315 0	988 12 0	1000 0 0	995 5 0	710 289 1	42 836 122	0 53 947
6.0	0 36 964	38 792 170	716 284 0	993 7 0	1000 0 0	995 5 0	712 287 1	48 832 120	0 31 969

A Criterion for Identification of the Mixture Normal Distribution

C. S. HONG[4], B. S. CHOI[5], J. S. UM[6]

In order to find the identification function from data and to apply the identification function for the pattern classification, we consider the existing problem of the number of patterns in such data. In this paper, a new criteria for the identification of Gaussian mixture distribution could be established as a characteristic of the sample variance, which is a bootstrap estimate of the sample variance. We examine the properties and fittness of the criteria through a large scale of computer simulations.

4) 3-53 Myungryun Dong, Chongro Gu, SEOUL 110-745, KOREA Associate Professor, Department of Statistics, Sung Kyun Kwan University.

5) 2-389 Samsun Dong, Sungbuk Gu, SEOUL 136-792 KOREA, Assistant Professor, Department of Computer Science & Statistics, Han Sung University.

6) 2-389 Samsun Dong, Sungbuk Gu, SEOUL 136-792 KOREA, Assistant Professor, Department of Computer Science & Statistics, Han Sung University.