On the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture with Unequal Variance

by

Z.D. Feng
Division of Public Health Sciences
Fred Hutchinson Cancer Research Center
1124 Columbia Street, MP702
Seattle, WA 98104

C.E. McCulloch
Biometrics Unit and Statistics Center
Cornell University
Ithaca, NY 14853

BU-1101-MA

August 1990
Revised February 1994

ON THE LIKELIHOOD RATIO TEST STATISTIC FOR THE NUMBER OF COMPONENTS IN A NORMAL MIXTURE WITH UNEQUAL VARIANCE

Z.D. Feng
Division of Public Health Sciences
Fred Hutchinson Cancer Research Center
1124 Columbia Street, MP702
Seattle, WA 98104

C.E. McCulloch
Biometrics Unit and Statistics Center
Cornell University
Ithaca, New York, 14853

February 1994

Abstract

An important but difficult problem in practice is to determine the number of components in a mixture normal model with unequal variances. When the likelihood ratio test statistic $-2log\lambda$ is used, it is unbounded above and fails to satisfy standard regularity conditions. A restricted maximization procedure must therefore be used which makes the procedure *ad hoc*. A consequence of this may explain the discrepancies among the simulation results of previous investigations.

Key words: Finite normal mixture; EM algorithm; Likelihood ratio test; Bootstrap; Singularity.

1 INTRODUCTION

Let $X_1, \cdots, X_n$ be a random sample of size n from a distribution $F(x, \theta)$ with density or mass function $f(x, \theta)$. A finite mixture density has the form

$$f(x, \theta) = \sum_{j=1}^{k} \pi_j f_j(x, \psi_j) \tag{1}$$

where ψ_j is the m-dimensonal parameter vector for component probability density function f_j, π_j is the mixing probability for the component j with restrictions $\sum_{j=1}^{k} \pi_j = 1$ and $\pi_j \geq 0$, and $\theta = (\pi_1, \cdots, \pi_k, \psi_1, \cdots, \psi_k)$ with dimension $p = k - 1 + km$. The number of components, k, may be known or unknown.

For testing H_0 : k' component mixture versus H_1 : k component mixture, where $k' < k$, the likelihood ratio statistic, $-2log\lambda$, does not have the usual chi-squared asymptotic distribution because: 1) under H_0, the true parameter θ_0 is on the boundary of the parameter space, 2) the distribution under H_0 are not identifiable. Wolfe (1971) performed a small scale simulation to investigate the limiting distribution of $-2log\lambda$ and suggested that $-\frac{2}{n}(n - 1 - m - \frac{k}{2})log\lambda$ has approximate limiting distribution $\chi^2_{2m(k-k')}$ where m is the dimension of ψ_j and n is the sample size.

The testing of

$$H_0 : f(x, \psi) = N(\mu, \sigma^2) \tag{2}$$

$$H_1 : f(x, \psi) = \pi N(\mu_1, \sigma_1^2) + (1 - \pi) N(\mu_2, \sigma_2^2) \tag{3}$$

using the generalized likelihood ratio test has been found to have additional difficulties. In searching for the maximum likelihood estimator under H_1, if we let $\hat{\mu}_1$ equal any observation in the sample and let $\hat{\sigma}_1^2$ approach zero, then the likelihood is unbounded above and therefore the global maximum does not exist. This was first noticed by Kiefer and Wolfowitz (1956).

Based on simulation, McLachlan (1987) suggested that χ^2_6 seemed to fit better as an approximation to the distribution of $-2log\lambda$ than χ^2_4 as suggested by Wolfe (1971). Hathaway (1985) suggested using the restriction $min_{i,j}(\sigma_i/\sigma_j) \geq c > 0$ to increase the chance of reliable convergence and claimed that by choosing a suitable c (satisfied by the true parameter) then there exists a consistent global maximum. McLachlan and Basford (1988) suggested restricting each component to have at least two observations to avoid the same difficult. However, their approach is only justifiable in estimation problems when we know that two components exist but it is not justified in the

testing problem. Both ideas essentially avoid putting a probability mass at a point. While we do not object to these approaches, there seems to be no mention of the dependency of the distribution of the test statistic on the choice of the criterion to avoid the above difficulty. We illustrate this point by a simulation.

2 Simulation Results and Conclusions

We used 500 simulations with sample size 100 under the null hypothesis, $N(0,4)$. Figures 1, 2, and 3 are the simulated cumulative distributions of $-2log\lambda$ using the EM algorithm (Dempster, Laird and Rubin, 1977), to compute the maximum likelihood estimator from incomplete data using three different criteria. The EM algorithm has been found to be useful to find the maximum likelihood estimator in mixture models (Redner and Walker, 1984). This algorithm was programmed in GAUSS, a mathematical programming language on the IBM PC. Since there are multiple maxima and a single starting point might converge to a local maxima, a grid of 27 different starting points are used and the biggest maximum is chosen. We used grid of $\pi_1 = (.1, .3, .5), \mu_1 = (-1, 0, 1), \sigma_1^2 = (1, 2, 4)$, and fix $(\mu_2, \sigma_2^2) = (1, 4)$ as the starting points. Our experience indicates that using 27 starting points is computationally feasible and greatly improves the chance of finding the largest of the local maxima. Further increasing the number of starting points does not lead to much improvement. We compared three different criteria using a simple but equivalent one to the approach of Hathaway (1985): $min(\sigma_1^2, \sigma_2^2) \geq 10^{-6}, 10^{-10}$ and 10^{-20} respectively. The simulation indicates that when the restriction is the most stringent (in the case of $min(\sigma_1^2, \sigma_2^2) \geq 10^{-6}$), the simulated distribution of $-2log\lambda$ lies on the left of the cumulative distribution of χ_5^2. It is actually between χ_4^2 and χ_5^2. When the restriction is less stringent (as in the case of $min(\sigma_1^2, \sigma_2^2) \geq 10^{-10}$), it lies quite evenly between χ_5^2 and χ_6^2. In the least stringent case of $min(\sigma_1^2, \sigma_2^2) \geq 10^{-20}$, it is closer to χ_6^2. In the upper tail which represents the cases where the maximum is located near the sigularity point, i.e., one of the data points, it becomes much bigger than the upper tail of χ_6^2. Brooks and Morgan (1993) used a maximization techique based on simulated annealing in the normal mixture problem and reported results which also differ from the McLachlan distributions.

(FIGURES 1, 2, 3 ABOUT HERE)

The above results indicate that when a restriction is imposed to avoid the unboundedness of the likelihood, the estimation process become *ad hoc* and therefore the associated asymptotic distribution of $-2log\lambda$ also depends on the criterion of the restriction. This might explain the discrepancies

of the simulated asymptotic distribution of $-2log\lambda$. This also implies that the bootstrap procedure is a good candidate for this situation. Any reasonable criterion can be chosen to get the 'maximum likelihood estimator' and then bootstrap samples can be generated using the same criterion to compute $-2log\lambda$. Since the bootstrap distribution mimics the underlying distribution of $-2log\lambda$ computed using the criterion, we do not need to worry about the correctness of the asymptotic distribution of $-2log\lambda$. This does, however, raise the interesting issue of power comparisions. For example, using the criteria of $10^{-6}, 10^{-10}$, and 10^{-20} on the variance estimates and an alternative where $(\pi_1, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ was $(.5, 1.5, -1.5, 1, 2.5)$, the values of the power were respectively as .68, .70 and .50 for the likelihood ratio test with size 0.05 from 500 simulations. The configuration here is chosen to provide $(\mu, \sigma^2) = (0, 4)$ under null hypothesis. The simulated vlaues of power corresponding to size 0.10 are .82, .78, and .70 respectively.

To sum up, the comparison of different authors' results on the distribution of the likelihood ratio statistic for testing the number of components in a normal mixture with unequal variances is not very meaningful without explicit identification of the criteria used for computing.

ACKNOWLEDGEMENT

The authors would like to thank the referees and editor Morgan for their helpful comments.

REFERENCES

Brooks, S.P. and Morgan, B.J.T. (1993). Optimisation using simulated annealing. Manuscript, Institute of Mathematics and Statistics, University of Kent.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1-38.

Edlefsen, L.E., and Jones, S.D. (1988). *GAUSS version 2.0.* Aptch System, Inc. Kent, WA.

Hathaway, R.J. (1985). A contrained formulation of maximum-likelihood estimation for normal mixture distributions. *Annals of Statistics*, 9, 795-800.

McLachlan, G.J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. A class of statistics with asymptotically normal distribution. *Applied Statistics*, 36, 318-324.

McLachlan, G.J. and Basford, K.E. (1988). *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York.

Redner, R.A. and Walker, H.F. (1984) Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review,* 26, 195-235.

Wolfe, J.H. (1971). A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinomal distributions. *Technical Bulletin STB 72-2,* U.S. Naval Personnel and Training Research Laboratory, San Diego.

Figure 1. Simulated cumulative distribution function of the likelihood ratio statistic and χ^2_4 and χ^2_5 distributions in a test of $H_0 : N(\mu, \sigma^2)$ vs $H_1 : \pi N(\mu_1, \sigma_1^2) + (1-\pi)N(\mu_2, \sigma_2^2)$ when H_0 is true. The likelihood ratio statistic is based on the maximum likelihood estimator. (sample size 100, 500 replications, variance $\geq 10^{-6}$).

Figure 2. Simulated cumulative distribution function of the likelihood ratio statistic and χ^2_5 and χ^2_6 distributions in a test of $H_0 : N(\mu, \sigma^2)$ vs $H_1 : \pi N(\mu_1, \sigma_1^2) + (1-\pi)N(\mu_2, \sigma_2^2)$ when H_0 is true. The likelihood ratio statistic is based on the maximum likelihood estimator. (sample size 100, 500 replications, variance $\geq 10^{-10}$).

Figure 3. Simulated cumulative distribution function of the likelihood ratio statistic and χ^2_5 and χ^2_6 distributions in a test of $H_0 : N(\mu, \sigma^2)$ vs $H_1 : \pi N(\mu_1, \sigma_1^2) + (1-\pi)N(\mu_2, \sigma_2^2)$ when H_0 is true. The likelihood ratio statistic is based on the maximum likelihood estimator. (sample size 100, 500 replications, variance $\geq 10^{-20}$).

Fig. 1.

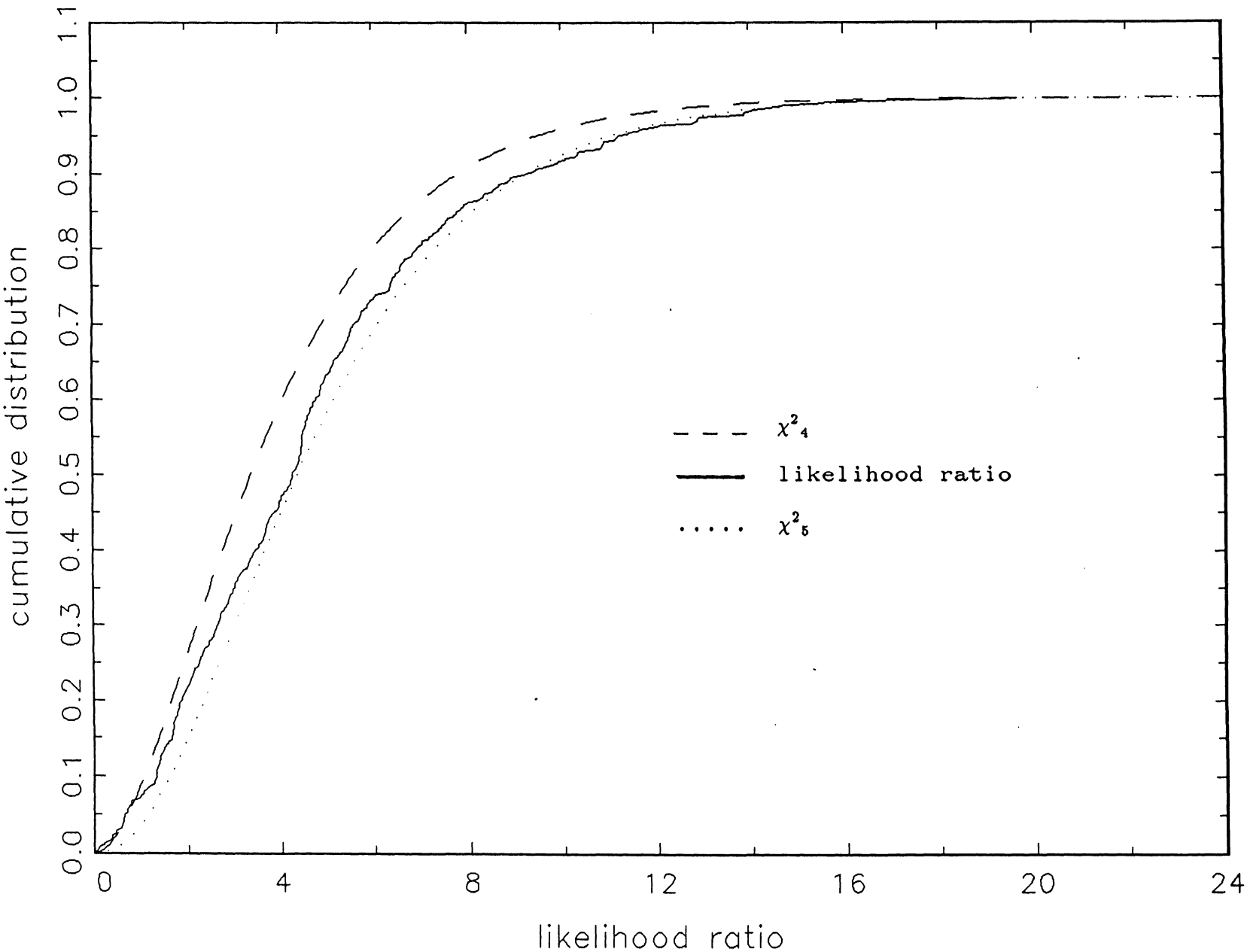

cumulative distribution
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
0
4
8
12
16
20
24
likelihood ratio
χ^2_4
likelihood ratio
χ^2_5

Fig. 2.

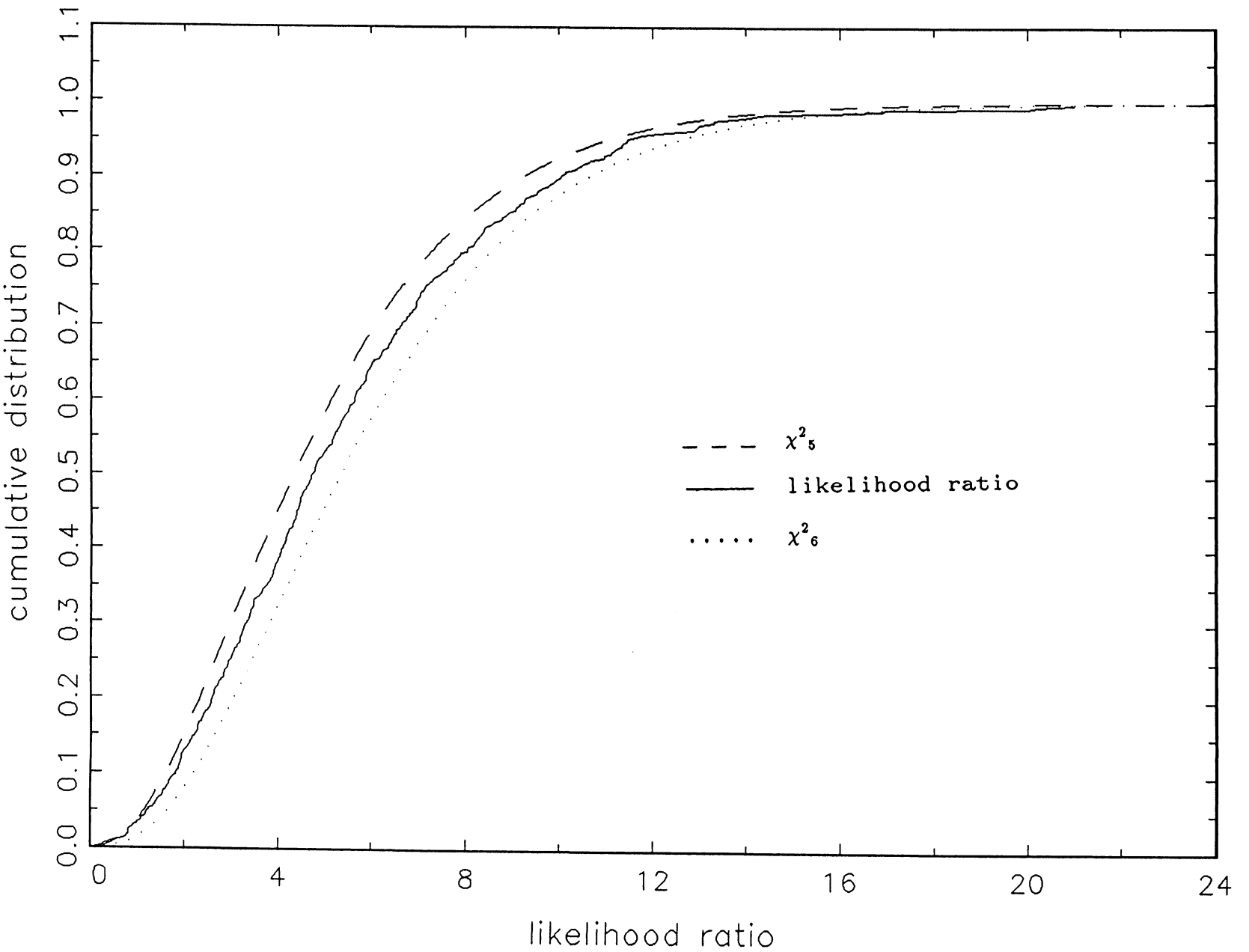
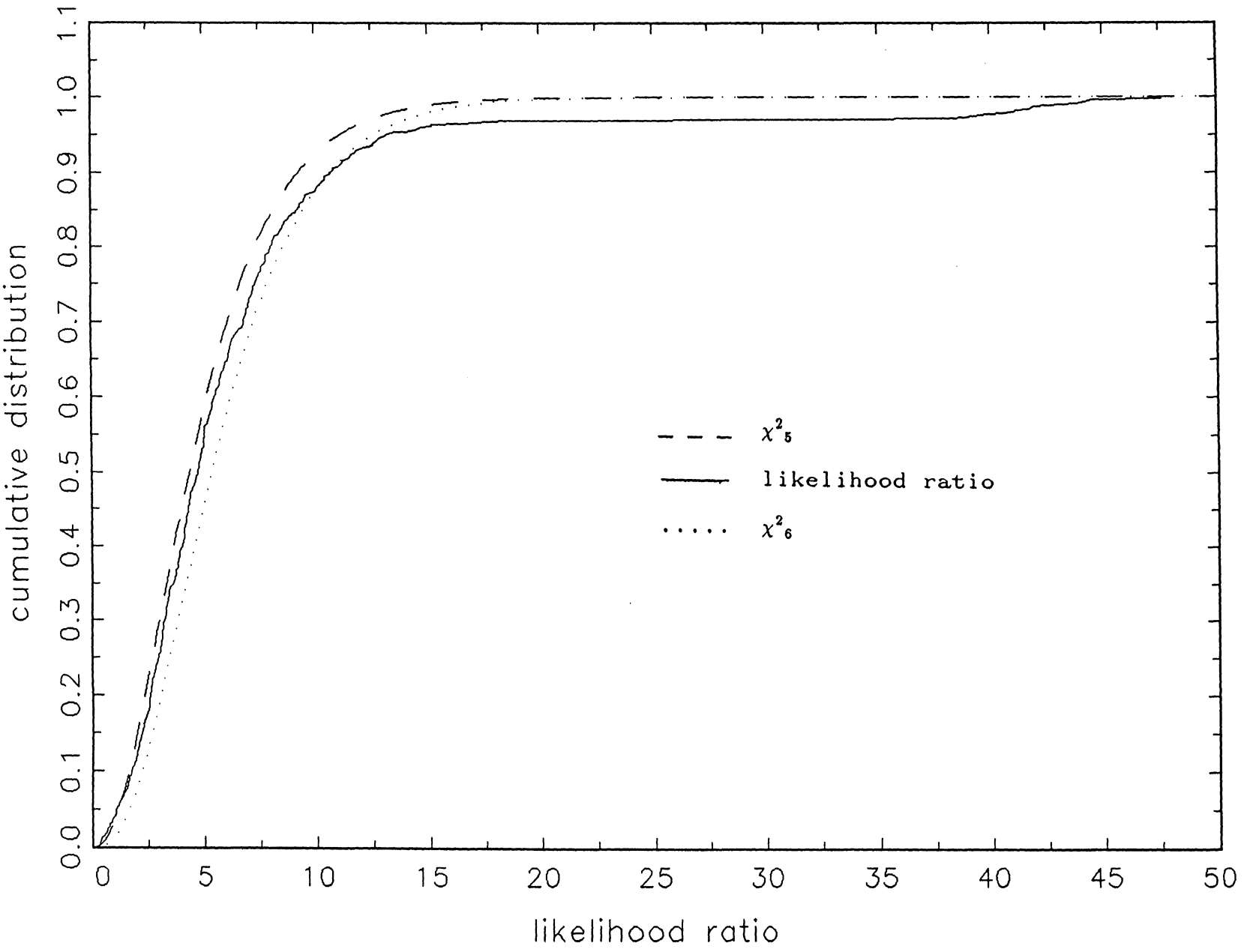

cumulative distribution
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
0
4
8
12
16
20
24
likelihood ratio
χ^2_5
likelihood ratio
χ^2_6

Fig. 3.


cumulative distribution
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
0
5
10
15
20
25
30
35
40
45
50
likelihood ratio
χ^2_5
likelihood ratio
χ^2_6