



ELSEVIER

Computational Statistics & Data Analysis 41 (2003) 367–378

**COMPUTATIONAL
STATISTICS
& DATA ANALYSIS**

www.elsevier.com/locate/csda

Weighted tests of homogeneity for testing the number of components in a mixture [☆]

Edward Susko*

Department of Mathematics and Statistics, Dalhousie University, Halifax, NS, Canada, B3H 3J5

Received 1 April 2002

Abstract

An important but difficult problem in fitting finite mixture models is estimating and testing the number of components in the mixture. Regularity conditions do not hold for large sample likelihood theory so that likelihood ratio tests cannot easily be implemented. However, a number of homogeneity tests have been developed to test for the presence of a mixture. Weighted versions of homogeneity tests are presented that can be used to test for the presence of additional components in a mixture. These tests are easily implemented, do not require long computational times and can incorporate covariates. Examples are given to illustrate the methodology and simulation results are presented that suggest that the tests have power comparable to the bootstrap likelihood ratio test.

© 2002 Elsevier Science B.V. All rights reserved.

Keywords: Mixture model; Finite mixtures; Homogeneity; Testing; Number of clusters

1. Introduction

Finite mixtures of distributions have been widely used to model data from a population that is suspected to be composed of several homogeneous subpopulations (see Do and McLachlan, 1984; Izenman and Sommer, 1988; Aitkin, 1999 for examples). The observations x_1, \dots, x_n are assumed to be a random sample from a *mixture density*

$$f(x, \psi) = \sum_{k=1}^m \pi_k f(x, \theta_k). \quad (1)$$

[☆] The author was supported by the Natural Sciences and Engineering Research Council of Canada.

* Tel.: +902-494-8865; fax: +902-494-5130.

E-mail address: susko@mathstat.dal.ca (E. Susko).

The subpopulation model that is frequently used as motivation for this density supposes that there are m subpopulations, that π_j is the probability of selecting an individual from subpopulation j and that $f(x, \theta_j)$, the *component density*, is the conditional density for X given that the observation is from the j th subpopulation. Since the true classification of observations into subpopulations is unobserved, the marginal density (1) must be used for the observations. Examples of common component densities $f(x, \theta)$ include normal, Poisson and binomial densities. Here $\psi = (\pi_1, \dots, \pi_{m-1}, \theta_1, \dots, \theta_m)$, the parameter π_m being $1 - \sum_{k=1}^{m-1} \pi_k$. The number of components m in the mixture is also a parameter and is what is of primary concern here.

The problem of estimating and testing the number of components has been considered by a number of authors. Part of the motivation for considering novel testing methods is that the likelihood ratio statistic for a test of $H_0: m = m_0$ against $H_A: m > m_0$ does not satisfy the regularity conditions for large sample likelihood theory and usually does not have a χ^2 limiting distribution (Hartigan, 1985). McLachlan (1987) proposes simulation to obtain the null distribution of the likelihood ratio statistic. While frequently useful, the repeated simulation required makes the test more computationally intensive than the methods proposed here and practical difficulties arise in implementing the test (Seidel et al., 2000) due to the likelihood frequently having multiple local maxima; these problems become more pronounced as m_0 increases. For estimation of the number of components, m , penalized model selection criteria such as AIC and BIC have been used (Bozdogan, 1987; Leroux, 1992) although the exact penalty terms are better motivated in regular parametric models (Kass and Raftery, 1995). Other methods have been developed such as Fowlkes (1979) and Roeder (1994) that are graphical in nature and apply to specific component densities.

The problem of testing for heterogeneity or overdispersion has received more attention than tests of the number of components. In most cases such *homogeneity tests* can be viewed as a test of $H_0: m = 1$ against $H_A: m > 1$. Note that under the null hypothesis the model for the data is a parametric one with density $f(x, \theta)$. Examples of tests include Potthoff and Whittinghill (1966a, b) and the $C(x)$ test of Neyman and Scott (1966). In the case that θ is known under the null, such tests reject when

$$\sum_{i=1}^n t(x_i, \theta) > c,$$

where $t(x, \theta)$ is a zero mean statistic under the null hypothesis. To extend these tests to a multicomponent null hypothesis we propose consideration of *weighted homogeneity tests* that reject the hypothesis $H_0: m = m_0$ at the j th component if

$$\sum_{i=1}^n p(j|x_i; \hat{\psi}) t(x_i, \theta_j) > c,$$

where the weights for the j th component, $p(j|x_i; \hat{\psi})$ are of the form

$$p(j|x; \hat{\psi}) = \hat{\pi}_j f(x, \hat{\theta}_j) \left/ \sum_{k=1}^{m_0} \hat{\pi}_k f(x, \hat{\theta}_k) \right.$$

In the subpopulation model that usually underlies the finite mixture model these weights are the chances that the observation came from the j th subpopulation.

2. Weighted homogeneity tests

The weighted homogeneity tests that we present require only the existence of a function $t(x, \theta)$ with $E_\theta[t(X, \theta)] = 0$ where expectation is with respect to the density $f(x, \theta)$. In order to expect reasonable power from the test, $t(x, \theta)$ should be chosen so that $\sum_i t(x_i, \theta)$ provides a test statistic for a test of $H_0: m = 1$ (with θ known) against $H_A: m > 1$ in (1). Neyman and Scott's $C(\alpha)$ test for homogeneity (Neyman and Scott, 1966; Lindsay, 1995, pp. 70) can be used to construct homogeneity test statistics. It gives

$$t(x, \theta) = \frac{\partial^2}{\partial \theta^2} f(x, \theta) / f(x, \theta) \tag{2}$$

and rejects if the test statistic is large. This test statistic is locally most powerful against a wide range of alternatives.

2.1. The component-wise test statistic and standard error

Given the function $t(x, \theta)$, the weighted homogeneity test statistic for the j th subpopulation is of the form

$$n^{-1/2} \sum_{i=1}^n p(j|x_i; \hat{\psi}) t(x_i, \hat{\theta}_j), \tag{3}$$

where $\hat{\psi}$ is the maximum likelihood estimate of ψ and the weights for the j th component are of the form

$$p(j|x; \psi) = \pi_j f(x, \theta_j) / \sum_{k=1}^{m_0} \pi_k f(x, \theta_k). \tag{4}$$

The motivation for the weighted homogeneity test is that these weights would tend to weight most heavily observations that are, in relative terms, most likely to have come from the j th homogeneous subpopulation and hence would give an approximate test of homogeneity for observations from this subpopulation. In the case that the observations are discrete

$$\sum_{x_i=y} p(j|x_i; \hat{\psi}) \approx n_{yj},$$

where n_{yj} is the unknown number of observations from the j th population that equal y . Thus (3) is approximately

$$\begin{aligned} n^{-1/2} \sum_{i=1}^n p(j|x_i; \hat{\psi}) t(x_i, \hat{\theta}_j) &= n^{-1/2} \sum_y t(y, \hat{\theta}_j) \sum_{x_i=y} p(j|x_i; \hat{\psi}) \\ &\approx n^{-1/2} \sum_y t(y, \hat{\theta}_j) n_{yj} \end{aligned}$$

or, up to a constant multiple, the test statistic for homogeneity for the j th population. Generally, under the null hypothesis, taking expectation with respect to the mixture density (1)

$$E \left[\sum_i p(j|X_i; \psi) t(X_i, \theta_j) \right] = n\pi_j E_{\theta_j} [t(X_i, \theta_j)] = 0.$$

The actual test statistic weights are based upon estimated parameters but if these are consistent estimators, up to first order the expectation under the null is 0.

We use maximum likelihood estimation to estimate ψ . Because estimation is under the null hypothesis, the problems with regularity conditions under the alternative that give rise to non-standard results for likelihood ratio tests do not arise here. With appropriate regularity conditions on the component densities standard likelihood theory applies and we have that

$$\sqrt{n}(\hat{\psi} - \psi) \rightarrow_d N(0, I^{-1}(\psi))$$

and that

$$n^{-1/2} \sum u(x_i, \psi) \rightarrow_d N(0, I(\psi)).$$

Here $u(x, \psi)$ is the gradient of $\log[f(x, \psi)]$ and $I(\psi)$ is the expectation of minus the second derivative matrix of $\log[f(x, \psi)]$. In practice we replace $I(\psi)$ by an asymptotic equivalent

$$n^{-1} \sum u(x_i, \psi) u(x_i, \psi)^T.$$

One can show that, under the null hypothesis, (3) has a normal distribution with mean 0 and variance

$$E[\{p(j|X; \psi)t(x, \theta_j)\}^2] - r_j(\psi)^T I^{-1}(\psi) r_j(\psi), \quad (5)$$

where

$$r_j(\psi) = E[u(X, \psi)t(X, \psi)p(j|X; \psi)].$$

If ψ were not estimated, the first component in this expression would give the large sample variance. The effect of estimation is to make statistic (3) smaller in magnitude, on average, than it would be without estimation. Subtraction of the second term, which is always non-negative, adjusts for the estimation. In practice this variance can be approximated by substituting $\hat{\psi}$ and taking sample averages as approximations to the corresponding expectations. A more sophisticated approach might estimate standard errors based on simulation from the null model. Note that the mixing distribution would not have to be re-estimated in the simulations. An α level test of $H_0: m = m_0$ against $H_0: m > m_0$ would then reject at the j th component if

$$\sum_{i=1}^n p(j|x_i; \hat{\psi}) t(x_i, \hat{\theta}_j) / \text{se} \left\{ \sum_{i=1}^n p(j|x_i; \hat{\psi}) t(x_i, \hat{\theta}_j) \right\} > z_\alpha.$$

2.2. Aggregate weighted homogeneity tests

The sum of the component-wise weighted homogeneity test statistics

$$n^{-1/2} \sum_j \sum_i p(j|x_i; \hat{\psi}) t(x_i, \theta_j) \tag{6}$$

gives a natural aggregate test statistic. Similarly as with the individual test statistics, with appropriate regularity conditions, this test statistic has a normal distribution with mean 0 and variance

$$E \left[\sum_j \{p(j|X; \hat{\psi}) t(x, \theta_j)\}^2 \right] - \left(\sum_j r_j(\psi) \right)^T I^{-1}(\psi) \left(\sum_j r_j(\psi) \right). \tag{7}$$

An α level test of $H_0: m = m_0$ against $H_A: m > m_0$ rejects when the test statistic (6) divided by an estimate of its standard deviation (7) is larger than z_α .

2.3. Two-sided tests

In some cases it may be valuable to consider two sided component-wise tests; the test statistic used in the normal simulations provides an example. An appropriate aggregate test should then reject when the magnitude of the individual component-wise test statistics are large, taking into account their variances and the correlation between them. Let t_a be the vector of component-wise test statistics: the j th entry of t_a is

$$n^{-1/2} \sum_i p(j|x_i; \hat{\psi}) t(x_i, \hat{\theta}_j).$$

Then the aggregate test we propose rejects at the alpha level if

$$t_a^T \text{Cov}(t_a)^{-1} t_a > \chi_{\alpha, m_0}^2,$$

where $\text{Cov}(t_a)$ is the covariance matrix for t_a . One can show that

$$\text{Cov}(t_a)_{ij} = E[p(i|X; \hat{\psi}) p(j|X; \hat{\psi}) t(X, \theta_i) t(X, \theta_j)] - r_j(\psi)^T I^{-1}(\psi) r_j(\psi). \tag{8}$$

In practice we have estimated all expectations in standard errors using sample averages and the estimated mixing distribution parameters $\hat{\psi}$

2.4. Incorporating structural parameters

In some statistical settings one is interested in a model with component density $f(x, \theta, \beta)$. Here β is a *structural parameter*, common to each observation regardless of which subpopulation it came from. As an example, the beta-blocker data considered in Section 3 models the log-odds of the probability of death as $\theta_i + \beta$ for the treatment group at center i . Here θ_i is a random effect for the centers in the study and β is a fixed structural parameter.

The component-wise test statistic becomes

$$n^{-1/2} \sum_i p(j|x_i; \hat{\psi}, \hat{\beta}) t(x_i, \hat{\theta}_j, \hat{\beta}).$$

The statistic $t(x, \theta_j, \beta)$ should be chosen so that $\sum_i t(x_i, \hat{\theta}_j, \hat{\beta})$ provides a good test statistic for a test of homogeneity. If β were known the $C(\alpha)$ test would be based on

$$t(x, \theta, \beta) = \frac{\partial^2}{\partial \theta^2} f(x, \theta, \beta) / f(x, \theta, \beta). \quad (9)$$

So a natural choice is (9) with $\hat{\beta}$ plugged in. The standard error for an individual component-wise test statistic is

$$E[\{p(j|X; \psi, \beta)t(X, \theta_j, \beta)\}^2] - r_j(\psi, \beta)^T I^{-1}(\psi, \beta) r_j(\psi, \beta),$$

where

$$r_j(\psi, \beta) = E[u(X, \psi, \beta)t(X, \theta_j, \beta)p(j|X; \psi, \beta)].$$

Here $u(X, \psi, \beta)$ and $I(\psi, \beta)$ are the scores and information matrices for the parameter vector (ψ, β) . Similar expressions as those given in (7) and (8) can be obtained for the standard errors for the aggregate test statistics.

2.5. Homogeneity test statistics

Common component densities for mixture models include the Poisson, binomial and normal densities. For the Poisson mixture model with component density

$$f(x, \theta) = \theta^x \exp(-\theta) / x!, \quad x = 0, 1, \dots \quad (10)$$

we use $t(x, \theta) = (x - \theta)^2 - x$ to construct the test statistic (3). A test of homogeneity based on $t(x, \theta)$ is equivalent to a $C(\alpha)$ test since $t(x, \theta) = \theta^2 (\partial^2 / \partial \theta^2) f(x, \theta) / f(x, \theta)$. In Potthoff and Whittinghill (1966b) it is shown that the one-sided test that rejects when $\sum_i t(x_i, \hat{\theta})$ is large is asymptotically optimal against the alternative that the data come from a mixture density with gamma mixing density. For the binomial component density

$$f(x_i, \theta) = \binom{m_i}{x_i} \theta^{x_i} (1 - \theta)^{m_i - x_i}, \quad x_i = 0, \dots, m_i$$

we use

$$t(x_i, \theta) = x_i(x_i - 1)/\theta + (m_i - x_i)(m_i - x_i - 1)/(1 - \theta) - m_i(m_i - 1) \quad (11)$$

to construct the test statistic (3). A test of homogeneity based on $t(x, \theta)$ is equivalent to a $C(\alpha)$ test since $t(x, \theta) = \theta(1 - \theta)(\partial^2 / \partial \theta^2) f(x, \theta) / f(x, \theta)$. In Potthoff and Whittinghill (1966a) it is shown that the one-sided test that rejects when $\sum_i t(x_i, \hat{\theta})$ is large is asymptotically optimal against the alternative that the data come from a mixture density with beta mixing density.

Test statistics for the normal component density,

$$\exp(-(x - \mu)^2/2\sigma^2)/\sqrt{2\pi}\sigma$$

need to adjust for the effects of estimation. In the case that σ^2 is a known fixed parameter common to each of the subpopulations so that $\theta = \mu$, $t(x, \theta) = (x - \theta)^2 - \sigma^2$ can be used. However, in the case $\theta = (\mu, \sigma^2)$, estimation of the σ^2 parameters implies that (3) is identically zero. We use

$$t(x, \theta) = (x - \mu)/\sigma^3 - 3(x - \mu)/\sigma. \tag{12}$$

Since the alternative has an additional normal component, the expectation is that this should be detectable as skewness at one of the fitted components under the null hypothesis.

3. Beta-blocker data example

The data, given in Aitkin (1999), considered in this example are the numbers of deaths among people involved in a clinical trial of beta-blockers at 22 centers and has been analyzed previously in Yusuf et al. (1985), Gelman et al. (1995) and Aitkin (1999). We adopt a constant log-odds ratio binomial model for the data:

$$\log[P(\text{Death})] = \begin{cases} \theta_i, & \text{center } i, \text{ control,} \\ \theta_i + \beta, & \text{center } i, \text{ treatment.} \end{cases}$$

The θ_i are assumed independent and identically distributed from an unknown mixing distribution. Our interest is in determining an appropriate number of components for the mixture by successively using the weighted homogeneity test. The homogeneity test statistic is (9) and we use a one sided test. The weighted homogeneity tests rejected the null at both components for the model with 2 components and the p -value for the aggregate test statistic was 0.005. The deviance for the model with 2 components was 145.23. With 3 components the aggregate test statistic had a p -value of 0.005. The p -values and the three components in the mixture were estimated as

θ	−1.61	−2.25	−2.834
Probability	0.249	0.512	0.239
p -value	0.000	0.144	0.011

The deviance was 101.29 and the β parameter was estimated as -0.2582 . The small p -value for the largest component suggests that there may be an additional component nearby. Fitting an additional component gave

θ	−1.44	−1.787	−2.258	−2.833
Probability	0.18	0.099	0.481	0.24
p -value	1.000	1.000	0.161	0.011

a deviance of 94.07726 and an estimated β of -0.2584 . An additional component is suggested by the small p -value near the smallest component. However, the p -values for the aggregate test was 0.228. Fitting an additional component gave

θ	−1.440	−1.787	−2.258	−2.684	−2.975
Probability	0.180	0.099	0.481	0.076	0.164
p -value	1.000	1.000	0.156	1.000	1.000

a deviance of 92.96 and a β of -0.2579 . The p -value for the aggregate test statistic was 0.390. Thus a maximum of 5 components is suggested by the weighted homogeneity test with the 4 component fit being favoured by the aggregate test.

4. Simulation results

We consider here the power of tests of 2 components. The proportions of rejections for $\alpha = 0.05$ level weighted homogeneity and bootstrap likelihood ratio tests are given in Tables 2–4 for three common component densities: Poisson, binomial and normal. In the normal simulations the mean was allowed to vary and a common variance parameter was estimated ($\theta_j = \mu_j$). The distributions used in each of the simulations are given in Table 1. For each of the resulting mixture densities, f_A , an approximation to the Kullback–Leibler divergence

$$\inf_{f_0 \in H_0} \int \log[f_A/f_0]f_A$$

is given and provides a measure of how distant the alternative distribution is from the null.

The weighted homogeneity tests were conducted using the homogeneity test statistics given in Section 2.5. The number of simulated data sets in each case was 1000. The proportion of rejections at either of the components can be expected to be larger than $\alpha = 0.05$ when the null hypothesis is true. These proportions are reported since they provide an upper bound on the power one might expect from any other form of aggregation than that proposed here.

The results for the bootstrap likelihood ratio test are based on parametric bootstrapping with 100 bootstrap samples for each simulated sample. The results reported are proportions of rejections over 100 simulated samples.

For each of the simulations, distribution 1 is a null distribution. Asymptotic critical values were used for the weighted homogeneity tests and critical values calculated from 100 bootstrap samples were used in the bootstrap likelihood ratio test. The results for distribution 1 give an indication of the size of these tests. No adjustment was made to make the sizes equal since none would be made in practice.

For the binomial and Poisson simulations the aggregate test gives smaller numbers of rejections than those at either component using component-wise tests but the difference is not substantial suggesting that not much is lost due to the aggregation. The power of

Table 1
The mixing distributions used in the simulations

Dist	Components			Probabilities			KL divergence	
<i>Poisson simulations</i>								
1	1.00	7.00		0.50	0.50			
2	1.00	7.00	10.00	0.25	0.50	0.25	0.007	
3	1.00	7.00	10.00	0.50	0.40	0.10	0.003	
4	1.00	7.00	8.00	0.50	0.25	0.25	0.000	
5	1.00	5.00	10.00	0.50	0.45	0.05	0.013	
<i>Normal simulations</i>								
1	−2.00	2.00		0.50	0.50			
2	−2.00	0.00	2.00	0.33	0.33	0.33	0.002	
3	−2.50	0.00	2.50	0.33	0.33	0.33	0.012	
4	−2.50	0.00	2.50	0.50	0.40	0.10	0.012	
Dist	Components			Probabilities			Size parameter	KL divergence
<i>Binomial simulations</i>								
1	0.25	0.75		0.50	0.50		40	
2	0.25	0.75	0.90	0.50	0.25	0.25	40	0.169
3	0.25	0.75	0.90	0.50	0.40	0.10	20	0.022
4	0.25	0.75	0.90	0.50	0.45	0.05	10	0.002
5	0.33	0.50	0.67	0.33	0.33	0.33	20	0.004

the weighted homogeneity tests was comparable to the power of the bootstrap likelihood ratio tests for both the binomial and Poisson simulations.

The results for distribution 1 for the normal simulations indicate that the size of the weighted homogeneity test is inflated which is likely due to the approximate nature of the asymptotics giving rise to its implementation. This makes comparison with the bootstrap likelihood ratio test which has the right size more complicated. Nevertheless the suggestion is that the size of the weighted homogeneity test is not grossly inflated and the power is large relative to the bootstrap likelihood ratio test. The proportions of rejections for the aggregate test is often larger than those at either component. This may in part be due to the approximate nature of the asymptotics but also has to do with the nature of the solution under the null. The two components in the estimated null distribution had the additional component of the alternative in the middle. Thus it was detectable at either component and consequently the aggregate test which may have involved more stable standard error estimation was able to out-perform the test at either component.

4.1. Forward selection

The weighted homogeneity tests can be used with forward selection to estimate the number of components much like in a regression setting. To investigate this, we

Table 2
The proportions of rejections for Poisson mixtures fit with two components

Distn.	<i>n</i>	Either component	Aggregate test	Bootstrap lr
1	250	0.058	0.025	0.03
	500	0.061	0.026	0.04
	1000	0.068	0.031	0.07
2	250	0.540	0.492	0.48
	500	0.803	0.784	0.71
	1000	0.976	0.974	0.97
3	250	0.272	0.237	0.25
	500	0.458	0.415	0.35
	1000	0.726	0.695	0.74
4	250	0.081	0.036	0.06
	500	0.102	0.063	0.06
	1000	0.122	0.079	0.14
5	250	0.644	0.612	0.71
	500	0.872	0.865	0.89
	1000	0.990	0.989	1.00

Table 3
The proportions of rejections for binomial mixtures fit with two components

Distn.	<i>n</i>	Either component	Aggregate test	Bootstrap lr
1	250	0.060	0.026	0.04
	500	0.067	0.034	0.05
	1000	0.065	0.035	0.09
2	250	1.000	1.000	0.91
	500	1.000	1.000	0.97
	1000	1.000	1.000	1.00
3	250	0.836	0.612	0.79
	500	0.984	0.890	0.91
	1000	1.000	0.995	0.95
4	250	0.157	0.098	0.20
	500	0.295	0.190	0.18
	1000	0.429	0.289	0.39
5	250	0.359	0.298	0.35
	500	0.623	0.552	0.49
	1000	0.874	0.845	0.82

generated 1000 samples of size 500 each from two Poisson mixtures and used forward selection with 0.1 level tests to estimate the number of components. When data

Table 4
The proportions of rejections for normal mixtures fit with two components (single variance parameter)

Distn.	n	Either component	Aggregate test	Bootstrap lr
1	250	0.186	0.135	0.02
	500	0.166	0.107	0.04
	1000	0.124	0.067	0.03
2	250	0.360	0.415	0.10
	500	0.352	0.467	0.31
	1000	0.477	0.632	0.39
3	250	0.705	0.789	0.50
	500	0.811	0.919	0.80
	1000	0.971	0.994	0.99
4	250	0.803	0.785	0.58
	500	0.942	0.919	0.93
	1000	0.999	0.995	0.99

was generated from a mixture with rate parameters 1 and 7, and with equal π_j the distribution of estimated components over 1000 simulated samples was

1	2	3	4	5	6	7+
0.000	0.879	0.095	0.025	0.010	0.000	0.000

When data was generated from a mixture with rate parameters 1, 7 and 10, with probabilities 0.5, 0.25 and 0.25, respectively, the distribution was

1	2	3	4	5	6	7+
0.000	0.218	0.728	0.047	0.006	0.001	0.000

4.2. Concluding remarks

The weighted homogeneity test described here is computationally efficient and easily implemented. Simulations suggest that it has power comparable to that of the much more computationally intensive bootstrap likelihood ratio test.

The component-wise tests were motivated as approximate separate homogeneity tests for the unobserved subpopulations in the mixture. Thus rejection at a particular component should suggest the presence of an additional component nearby which did seem to be the case in the simulations. The aggregate test statistic adjusts for the inflation of type I error that would occur if the null was rejected whenever any one of the component-wise tests rejected. The simulations suggest that there is not much loss in power due to aggregation.

An S-Plus function to compute the test statistic based upon an input component density and homogeneity function $t(x, \theta)$ is available upon request from the author.

References

- Aitkin, M., 1999. Meta-analysis by random effect modelling in generalized linear models. *Statist. Med.* 18, 2343–2351.
- Bozdogan, H., 1987. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* 52, 345–370.
- Do, K., McLachlan, G.J., 1984. Estimation of mixing proportions: a case study. *Appl. Statist.* 33, 134–140.
- Fowlkes, E.B., 1979. Some methods for studying the mixture of two normal (lognormal) distributions. *J. Amer. Statist. Assoc.* 74, 561–575.
- Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. *Bayesian Data Analysis*. Chapman and Hall, London.
- Hartigan, J.A., 1985. A failure of likelihood asymptotics for normal mixtures, in: *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman*, Vol. 2, pp. 807–810.
- Izenman, A.J., Sommer, C., 1988. Philatelic mixtures and multimodal densities. *J. Amer. Statist. Assoc.* 83, 941–953.
- Kass, R.E., Raftery, A.E., 1995. Bayes factors and model uncertainty. *J. Amer. Statist. Assoc.* 90, 773–795.
- Leroux, B., 1992. Consistent estimation of a mixing distribution. *Ann. Statist.* 20, 1350–1360.
- Lindsay, B.G., 1995. *Mixture Models: Theory, Geometry and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, Hayward: Institute for Mathematical Statistics.
- McLachlan, G.J., 1987. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Statist.* 36, 318–324.
- Neyman, J., Scott, E.L., 1966. On the use of $C(\alpha)$ optimal tests of composite hypotheses. *Bull. Inst. Internat. Statist.* 41 (1), 477–497.
- Potthoff, R.F., Whittinghill, M., 1966a. Testing for homogeneity. I. The binomial and multinomial distributions. *Biometrika* 53, 167–182.
- Potthoff, R.F., Whittinghill, M., 1966b. Testing for homogeneity. II. The Poisson distribution. *Biometrika* 53, 183–190.
- Roeder, K., 1994. A graphical technique for determining the number of components in a mixture of normals. *J. Amer. Statist. Assoc.* 89, 487–495.
- Seidel, W., Mosler, K., Alker, M., 2000. A cautionary note on likelihood ratio tests in mixture models. *Ann. Inst. Statist. Math.* 52, 481–487.
- Yusaf, S., Peto, R., Lewis, J., Collins, R., Sleight, P., 1985. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in Cardiovascular Diseases* 27, 335–371.