Communications in Statistics: Simulation and Computation, 2007

SIZE AND POWER OF RECENT TESTS FOR HOMOGENEITY IN EXPONENTIAL MIXTURES [1]

Karl Mosler and Lars Haferkamp
Statistik und Ökonometrie
Universität Köln
D-50923 Köln, Germany
mosler@statistik.uni-koeln.de

Key Words: Mixture diagnosis; survival analysis; unobserved heterogeneity; overdispersion; goodness-of-fit; ADDS test; D-test; modified likelihood ratio test.

ABSTRACT

The paper investigates diagnostic procedures for finite mixture models. The problem is to decide whether given data stem from an exponential distribution or a finite mixture of such distributions. Recently three new test approaches have been proposed, the modified likelihood ratio test (MLRT) by (Chen, Chen, & Kalbfleisch, 2001), the ADDS test by (Mosler & Seidel, 2001), and the D-test by (Charnigo & Sun, 2004). The size and power of these tests are determined by Monte Carlo simulation and their relative merits are evaluated. We conclude that the ADDS test shows always not much less and under some alternatives, in particular lower contaminations, considerably more power than its competitors. Also new tables for the ADDS test are provided.

1. INTRODUCTION

In many applications mixture models arise as a natural way to model population heterogeneity; see (Lindsay, 1995), (Titterington, Smith, & Makov, 1985), and others. The

[1] We thank Richard Charnigo and Jiahua Chen for providing the codes of the MLRT and D-tests. Thanks are also to Christoph Scheicher and Wilfried Seidel for discussions and an unknown referee for useful hints.

assumption that the data are generated by a mixture of exponential distributions is widely employed in the analysis of lifetime and other duration data. This model arises from incomplete observation of an underlying conditional exponential model.

While the hazard rate of a pure exponential distribution is a constant, the hazard of a mixture of exponentials decreases. Therefore the mixture model is frequently adopted to fit the distribution of a time to 'failure' where the observed failure rate seems to decline with time. Often the mixture can be explained by competing risks: the population divides into parts which are subject to different reasons of failure (see (Prentice et al., 1978)).

(Lindsay, 1995) presents a comprehensive treatment of the theory and numerous applications of mixture models, and (McLachlan, 1995) gives a survey of mixtures of exponentials. (Böhning, Schlattmann, & Lindsay, 1992) provide computational tools for estimating such mixtures.

Assume that we observe a random sample $X_1, X_2, \ldots, X_n$ from a finite mixture of exponential distributions,

$$f(x, \lambda, p) = \sum_{j=1}^{k} \frac{\pi_j}{\lambda_j} \exp\left(\frac{x}{\lambda_j}\right), \tag{1}$$

where $\lambda_j > 0, \pi_j \geq 0, \sum_j \pi_j = 1$, and k is the number of possible mixture components. An important question is how many components are present and, in particular, whether the data are generated by a homogeneous distribution ($k = 1$), or not. We wish to test the homogeneity hypothesis

$$H_0 : \lambda_1 = \ldots = \lambda_k \tag{2}$$

against H_1 : not H_0.

It is well known that the likelihood ratio test (LRT), while being locally optimal, has a nonstandard asymptotic distribution that is difficult to implement. Other tests have been proposed for homogeneity in a mixture model, among them moment likelihood tests, dispersion score (DS) tests, which detect a mixture by its overdispersion. DS tests are also known under the name $C(\alpha)$-tests; see chapter 4 of (Lindsay, 1995). While these approaches work well, e.g., in normal mixtures, the diagnosis of exponential mixtures poses additional problems: The moment likelihood and the dispersion score tests have no power on a large class of

alternatives (Mosler & Seidel, 2001), and the calculation of the LRT statistic and the Monte-Carlo simulation of its null distribution depend heavily on the particular implementation of the EM algorithm (Seidel, Mosler, & Alker, 2000).

To test for homogeneity in exponential mixtures three alternative approaches have been recently proposed:

- The modified likelihood ratio test (MLRT) introduced by (Chen et al., 2001), which is a penalized LRT and has standard χ^2 asymptotics,
- the ADDS Test by (Mosler & Seidel, 2001), a combination of the dispersion score test with a properly chosen goodness-of-fit procedure,
- the D-test by (Charnigo & Sun, 2004), based on the L^2 distance between the estimated densities of a homogeneous and a heterogeneous model. (Charnigo & Sun, 2004) also introduce a penalized and several weighted variants of the D-test.

In this paper we will compare these test approaches and evaluate them in terms of their size and power by means of a large simulation study. Also, besides the simple D-test, two weighted D-tests and a penalized D-test are included in our comparison. Section 2 shortly surveys the test approaches. In Section 3 we check whether these tests, possibly depending on sample length, keep their nominal size, and in Section 4 their power is compared. Section 5 concludes, and the Appendix contains new tables of critical quantiles for the ADDS test.

2. TESTS

In this section the three test approaches are shortly surveyed. For details of the tests the reader is referred to the original papers.

We restrict on exponential mixtures that have at most two components, i.e., on densities

$$f(x) = (1-\epsilon)\frac{1}{\lambda_1}\exp\left(-\frac{x}{\lambda_1}\right) + \epsilon\frac{1}{\lambda_2}\exp\left(-\frac{x}{\lambda_2}\right),\ x \geq 0\,, \tag{3}$$

$\lambda_1, \lambda_2 > 0$, $0 < \epsilon < 1$. The alternative hypothesis corresponds to $\lambda_1 \neq \lambda_2$, while the null hypothesis may be signified by $\lambda = \lambda_1 = \lambda_2$ and, e.g., $\epsilon = \frac{1}{2}$. Let a random sample $X_1, \ldots, X_n$ from (3) be given.

The MLRT (Chen et al., 2001) employs the usual log-likelihood plus a penalty term,

$$l(\epsilon, \lambda_1, \lambda_2) = \sum_{i=1}^{n} \log \left[(1-\epsilon) \frac{1}{\lambda_1} \exp\left(-\frac{x}{\lambda_1}\right) + \epsilon \frac{1}{\lambda_2} \exp\left(-\frac{x}{\lambda_2}\right) \right] + C \log[4\epsilon(1-\epsilon)] \,, \qquad (4)$$

where $C > 0$ is a constant which weighs the penalty. (Chen et al., 2001) report that the MLRT is rather insensitive to C; they propose $C = \log M$ if $-M \leq \lambda_j \leq M$. Following (Charnigo & Sun, 2004) we choose a fixed $C = \log 10$ in this study. By maximizing (4) under H_0 and H_1, estimates $\hat{\lambda}_1, \hat{\lambda}_2, \hat{\epsilon}$ and, respectively, $\hat{\lambda}$ are obtained. The MLRT uses twice the ratio of penalized loglikelihoods as test statistic,

$$T_{MLRT} = 2(l(\hat{\epsilon}, \hat{\lambda}_1, \hat{\lambda}_2) - l(0.5, \hat{\lambda}, \hat{\lambda}) \,.$$

Asymptotically, under H_0 and some regularity conditions, T_{MLRT} is distributed as the fifty-fifty mixture of a χ^2_1 variable and a constant at 0. (Chen et al., 2001) demonstrate that, to detect normal location mixtures and Poisson mixtures, the MLRT develops similar or slightly better power than the DS test.

The ADDS test by (Mosler & Seidel, 2001) is a hybrid procedure that combines the DS test with a classical goodness-of-fit test, specifically, the Anderson-Darling test: Calculate the DS statistic

$$T_{DS} = \left(\frac{n(n-1)}{n+1} \right)^{\frac{1}{2}} \frac{1}{(\overline{X})^2} \left[S^2 - \frac{1}{2n} \sum_{i=1}^{n} X_i^2 \right] \qquad (5)$$

and the Anderson-Darling statistic

$$T_{AD} = \left(1 + \frac{0.6}{n}\right) \left(n - \frac{1}{n} \sum_{i=1}^{n} (2i-1) \left(\log\left(1 - e^{-X_{(i)}/\overline{X}}\right) + \frac{X_{(i)}}{\overline{X}} \right) \right), \qquad (6)$$

where $\overline{X}$ and S^2 denote the sample mean and variance and $X_{(i)}$ is the i-th order statistic. Reject H_0 if either T_{DS} or T_{AD} are too large. Under H_0 both test statistics do not depend on the parameter λ. For tables of pairs of critical quantiles for the two statistics; see the Appendix. The power of the ADDS is always at least comparable to that of a bootstrap LRT, a moment LRT and the DS test, it is much better than that of the latter two tests on a large class of alternatives; see (Mosler & Seidel, 2001).

The D-test in its original form measures the area between two densities, one fitted under H_0, and the other fitted under H_1. If the alternative is a two-component exponential scale mixture, the D-statistic is

$$T_D = \int_0^\infty \left[\frac{1-\hat{\epsilon}}{\hat{\lambda}_1} \exp\left(\frac{x}{\hat{\lambda}_1}\right) + \frac{\hat{\epsilon}}{\hat{\lambda}_2} \exp\left(\frac{x}{\hat{\lambda}_2}\right) - \frac{1}{\hat{\lambda}} \exp\left(\frac{x}{\hat{\lambda}}\right) \right]^2 dx \,, \tag{7}$$

where $\hat{\lambda}$ is an estimate of λ under H_0, $\hat{\lambda}_1, \hat{\lambda}_2$ and $\hat{\epsilon}$ are estimates of the parameters under H_1.

(Charnigo & Sun, 2004) show that, under H_0, T_D has an asymptotic distribution which is equivariant in λ. They provide tables of critical quantiles and report that, in the diagnosis of exponential scale mixtures, the simple D-test is slightly outperformed by the MLRT when n is small ($n \leq 100$). (Charnigo & Sun, 2004) therefore propose weighted forms of the D-test which put more weight to differences in the tails of the alternative densities: In place of the differential dx in the integral formula (7) they use $x\,dx$ or $x^2\,dx$. In the sequel, these weighted variants of the D-test will be signified by 'w1D' and 'w2D', respectively. Another variant of the D-test that uses a penalty term ('penD') is also considered.

3. SIZE

In a large Monte-Carlo simulation we calculate the actual size of the tests under consideration, that is, the rejection probability under H_0. We do this for different nominal sizes (= levels of significance) and sample sizes. The number of replications is always 20 000.

We investigate the ADDS test, the D-test, the weighted D-tests with weighting functions $w_1 = x$ (w1D) and $w_2 = x^2$ (w2D), the penalized D-test (penD), and the MLRT; the latter is based on its asymptotic χ^2-distribution. Following (Charnigo & Sun, 2004), the constant $C > 0$ that controls the penalty term is set to $\log(10)$.

As the MLRT and the D-tests involve estimates of the model under the alternative, in the sequel we restrict on mixture alternatives (3) that have only two components.

Table 1 shows simulated actual sizes of the six tests, given the sample sizes $n = 100$ and 1000 and standard significance levels $\alpha = 0.05$ and 0.01. In particular, for $n = 100$ the simple D-test rejects the null hypothesis with frequency 0.069 when the nominal size is

0.05 (factor 1.4), resp. with frequency 0.015 when the nominal size is 0.01 (factor 1.5). The penalized D-test overshoots its nominal size by up to a factor 6. The MLRT appears to be conservative (0.040 in place of 0.050) for $n = 1000$. On the other hand, the sizes of the ADDS test and the weighted D-tests come close to their nominal ones.

Further, we demonstrate how the actual sizes change with n. Figure 1 exhibits actual sizes of the ADDS test, the MLRT, the simple D-test and the penalized D-test for $100 \leq n \leq 500$ and $\alpha = 0.05$. While, by construction, the size of the ADDS test remains (approximately) equal to α when n increases, the sizes of the other two tests do not. The sizes of both D-tests converge rather slowly from above to α, and the size of the MLRT stays below α, also for large n.

n	Test	$\alpha = 0.05$	$\alpha = 0.01$
100	ADDS	0.046	0.009
	D	0.069	0.015
	w1D	0.052	0.009
	w2D	0.045	0.009
	penD	0.105	0.060
	MLRT	0.051	0.011
1000	ADDS	0.047	0.009
	D	0.054	0.013
	w1D	0.056	0.014
	w2D	0.056	0.013
	penD	0.060	0.020
	MLRT	0.046	0.010

Table 1: Actual size of the ADDS test, four D-tests, and the MLRT when nominal size is $\alpha \in \{0.01, 0.05\}$, and sample length is $n \in \{100, 1000\}$.

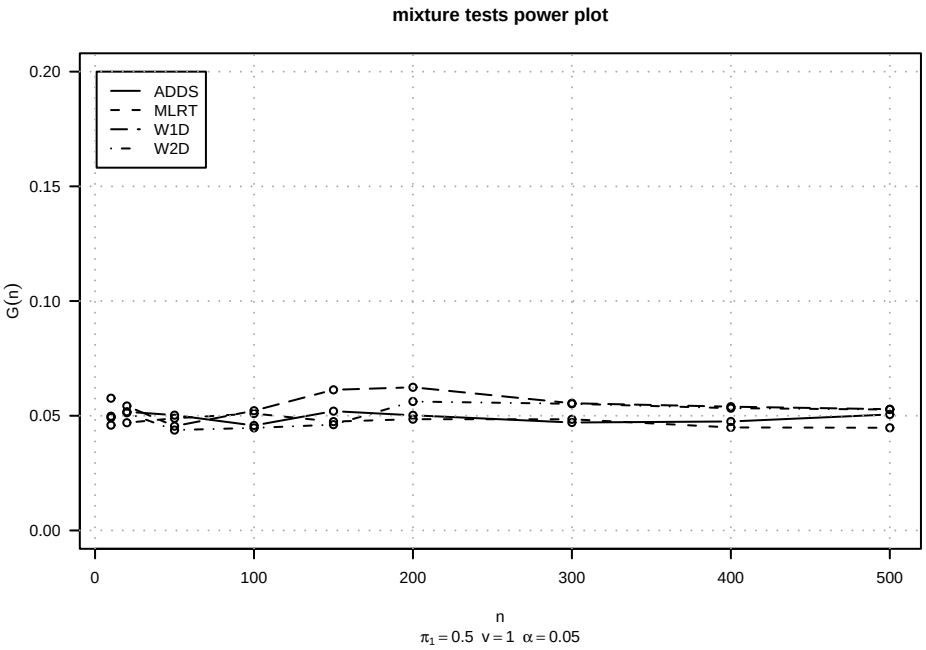

Figure 1: Actual size of ADDS test, MLRT, D-test, and penalized D-test depending on n; nominal size is $\alpha = 0.05$.

4. POWER

As the simple D-test and the penalized D-test do not keep their prescribed significance level we restrict the subsequent power study to the remaining four tests. We compare the power of the ADDS test, the MLRT and the D-tests with weighting functions $w_1 = x$ (w1D) and $w_2 = x^2$ (w2D).

The power of the tests is evaluated by Monte-Carlo simulation and compared on a large number of two-component mixture alternatives (3), sample sizes n, and significance levels α. In the sequel we present some key results of the simulation study. The number of replications is always 5000.

As the test problem is scale invariant, the ratio of the two scale parameters, $v = \frac{\lambda_2}{\lambda_1}$, will be considered only, with $\lambda_2 \geq \lambda_1$. Equation (3) becomes

$$f(x) = (1-\epsilon)\exp(-x) + \epsilon\frac{1}{v}\exp\left(-\frac{x}{v}\right),\ x \geq 0\,. \tag{8}$$

We study and compare the power of the alternative tests for increasing v, $v \geq 1$, and for

three typical cases concerning the parameter ϵ. Firstly, fifty-fifty mixtures ($\epsilon = 0.5$) and the comparative ability of the tests to detect them are considered. Next, an exponential distribution is mixed with a small portion of another exponential distribution that has v times its expectation; the resulting mixture is called an *upper contamination*. Here, the portion is ten percent. Thirdly, the same is done with a small portion of an exponential distribution having $\frac{1}{v}$ times its expectation; this is named a *lower contamination*.

Figure 2 exhibits power results on fifty-fifty mixtures for two levels of significance, $\alpha = 0.05$ and $\alpha = 0.01$. If n is large ($n = 1000$), the tests behave similarly. However, for moderate n, the linearly weighted D-test (w1D) is clearly outperformed by the others, especially when α is small. On upper contaminations, as Figure 3 demonstrates, the w1D-test is the best one, the ADDS test is second best, and the remaining two tests behave similarly.

Figure 4 concerns lower contaminations. All four tests have problems in detecting a lower contamination, when n is not large. This is due to the strong asymmetry of the exponential distribution: Its mass is concentrated near the left border of the support, which tends to mask any lower contamination. For $n = 100$ and $\alpha = 0.01$ (resp. 0.05), the power of all four tests does not exceed 20 % (resp. 40 %) when the scale ratio v is less than 10. For $n = 1000$, the situation improves considerably. However, for relatively large v, the ADDS test develops much better power than its competitors; this holds for different significance levels and sample sizes as well.

Figure 5 shows that also for intermediate sample sizes ($n = 200, 500$) the ADDS test outperforms the others when v is large enough. The quadratically weighted D-test should not be used unless n exceeds 500.

Next, we give a more detailed picture of the lower contamination case when the scale ratio v is relatively small ($v = 2, v = 3$). Figure 6 presents power curves depending on sample size n. Here, the w2D-test and the MLRT are best, while the w1D-test and – to a lesser extent – the ADDS test appear to be slightly inferior. To obtain at least 50 per cent power in detecting a fifty-fifty mixture that has scale ratio 2, the sample size should be at least 400 with any of these tests. In detecting a fifty-fifty mixture that has scale ratio 3, a

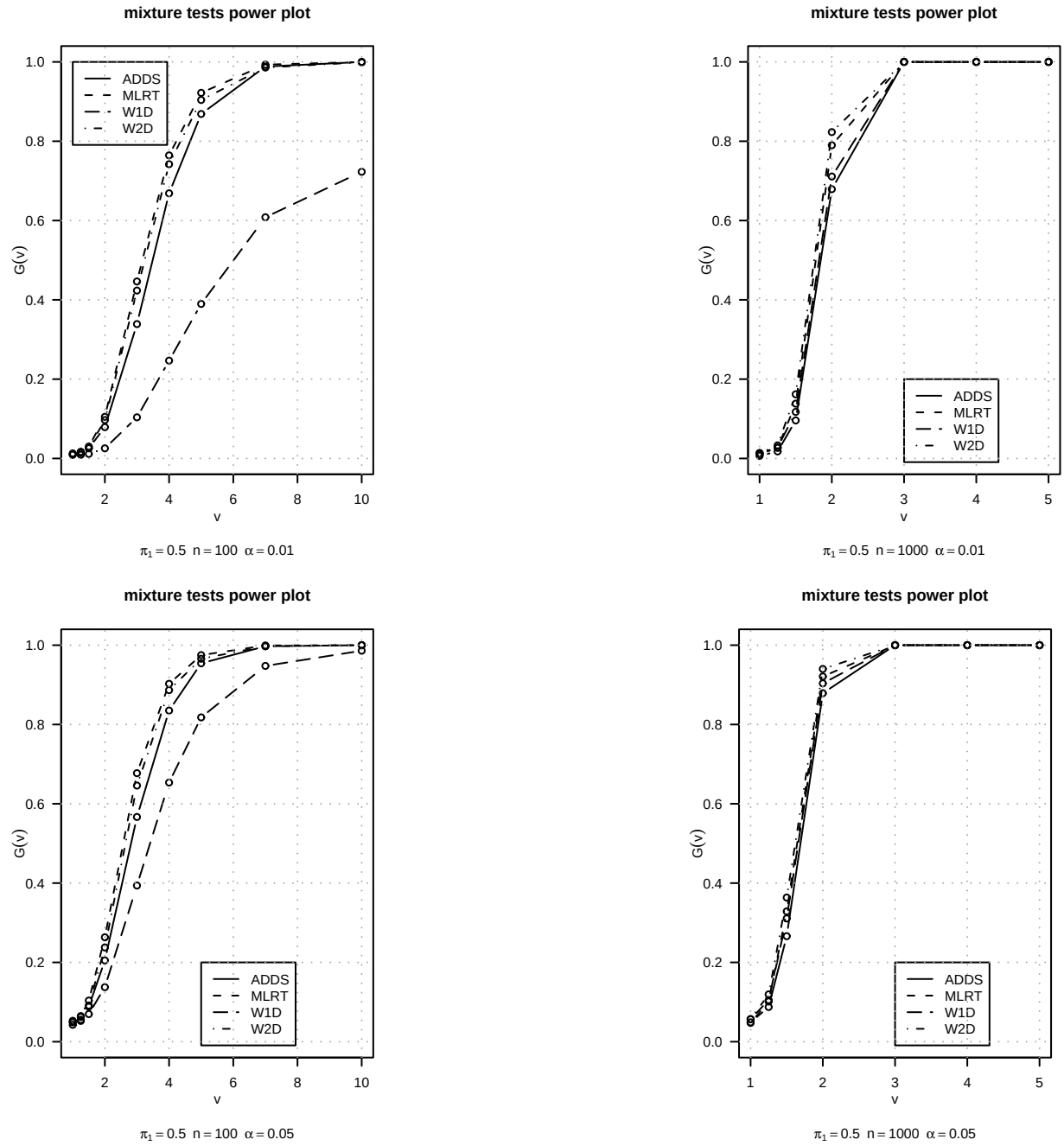

Figure 2: Power on fifty-fifty mixtures. Combined overdispersion and Anderson-Darling test (ADDS), weighted D-test (W1D / W2D), and modified likelihood ratio test (MLRT) on alternatives $f(x) = 0.5\exp(-x) + 0.5\frac{1}{v}\exp(-\frac{x}{v})$.

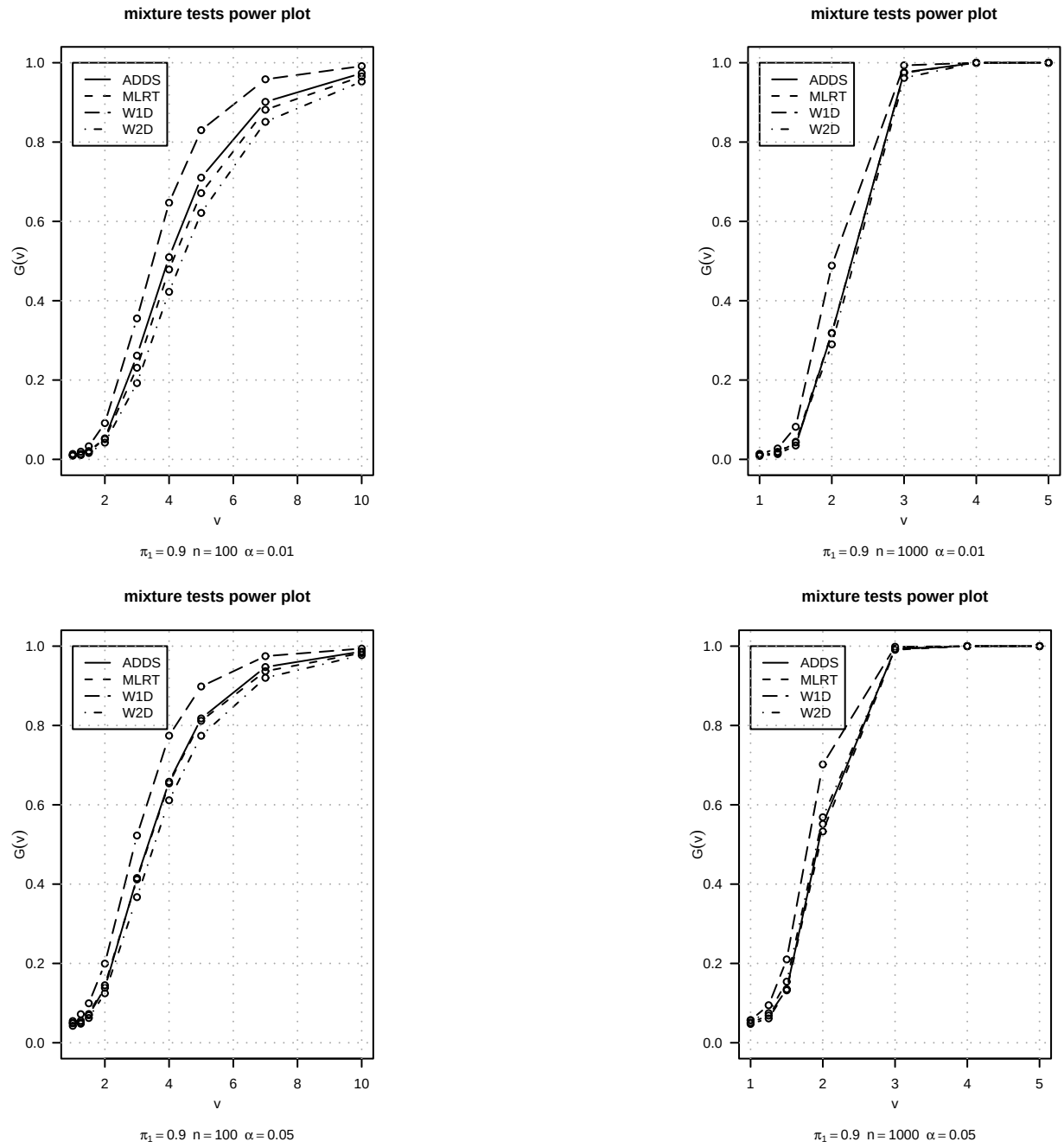

Figure 3: Power on mixtures with upper contamination. Combined overdispersion and Anderson-Darling test (ADDS), weighted D-test (W1D / W2D), and modified likelihood ratio test (MLRT) on alternatives $f(x) = 0.1\exp(-x) + 0.9v\exp(-vx)$.

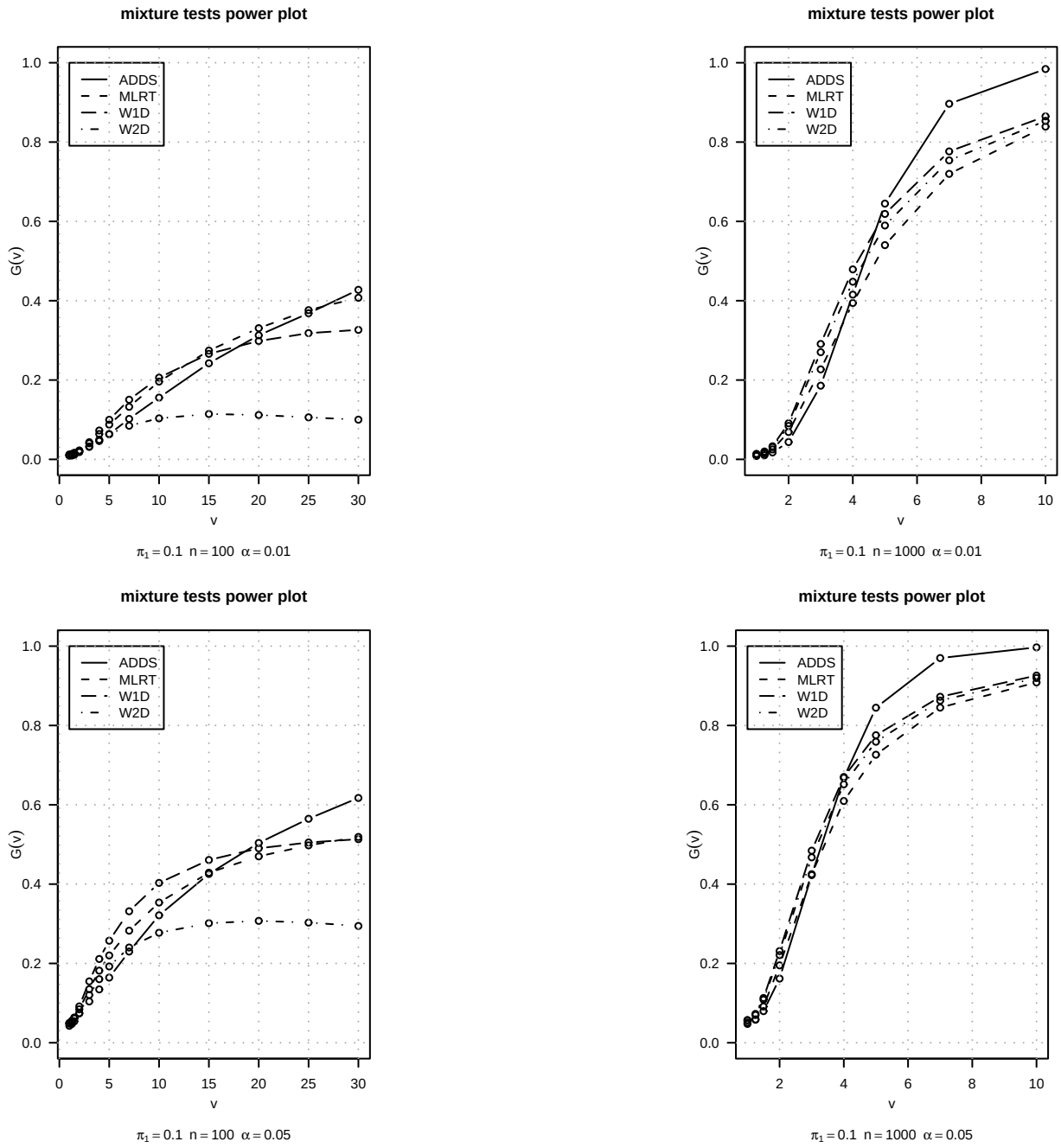

Figure 4: Power on mixtures with lower contamination. Combined overdispersion and Anderson-Darling test (ADDS), weighted D-test (W1D / W2D), and modified likelihood ratio test (MLRT) on alternatives $f(x) = 0.9\exp(-x) + 0.1v\exp(-vx)$.

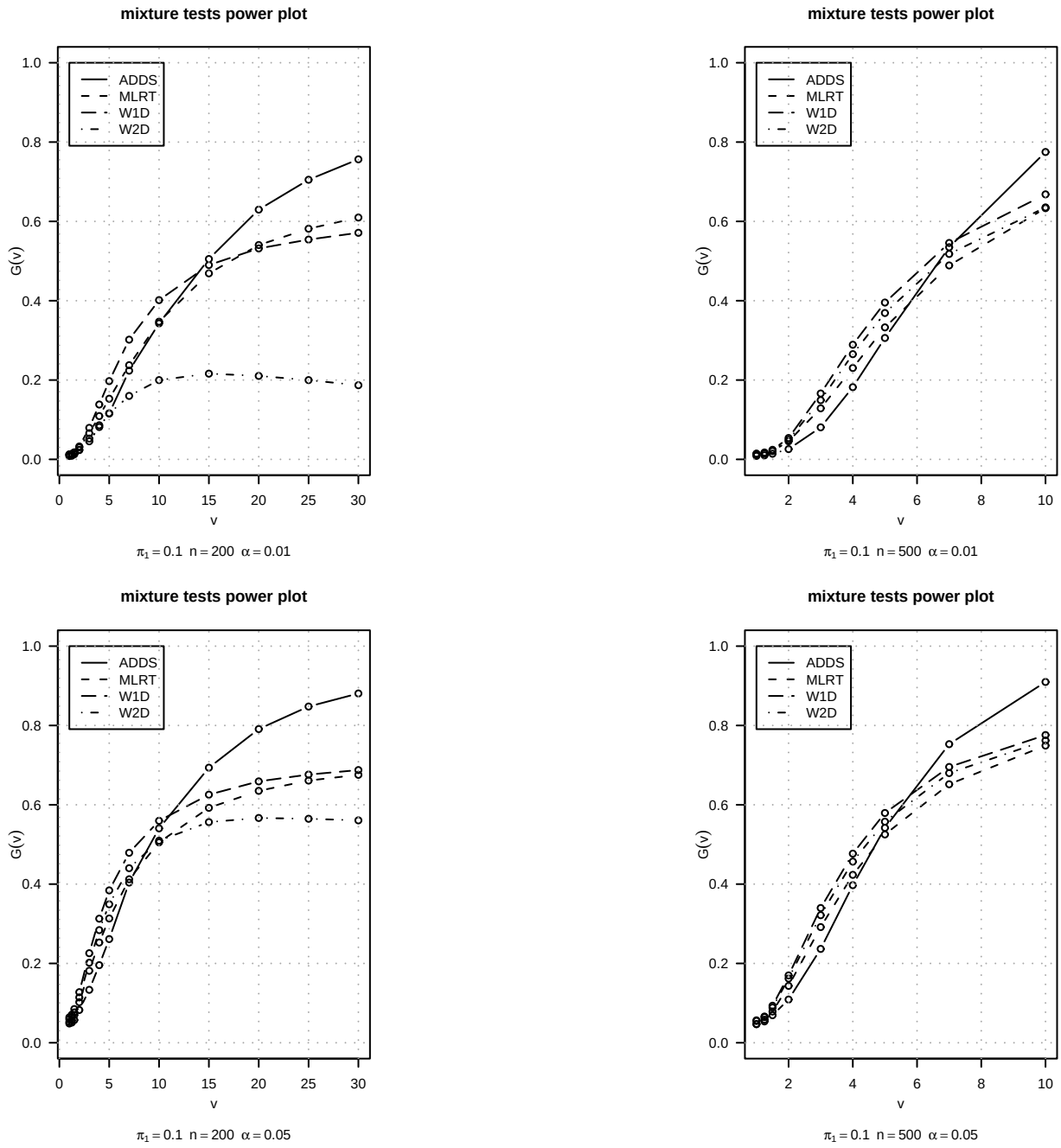

Figure 5: Power on mixtures with lower contamination. Combined overdispersion and Anderson-Darling test (ADDS), weighted D-test (W1D / W2D), and modified likelihood ratio test (MLRT) on alternatives $f(x) = 0.9\exp(-x) + 0.1v\exp(-vx)$.

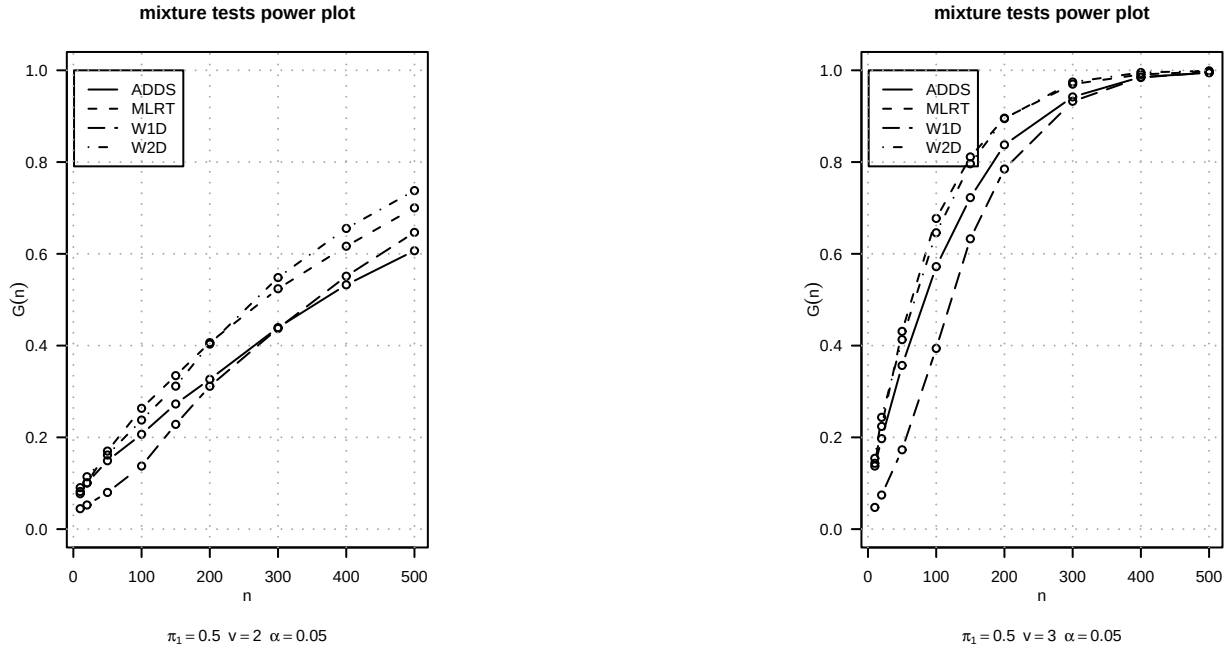

Figure 6: Power on fifty-fifty mixtures dependent on the sample size n. Combined overdispersion and Anderson-Darling test (ADDS), weighted D-test (W1D / W2D), and modified likelihood ratio test (MLRT) on alternatives $f(x) = 0.5\exp(-x) + 0.5\frac{1}{v}\exp(-\frac{x}{v})$.

sample size of 100 might be sufficient.

5. CONCLUDING REMARKS

In this paper the behaviour of several recently proposed tests for the diagnosis of exponential mixtures has been compared.

We have demonstrated that for sample lengths $n \leq 1000$ the D-test, the penalized D-test and, to a lesser extent, the MLRT do not keep their nominal size. Both D-tests are anti-conservative, the simple D-test less than the penalized one. Their anti-conservativism decreases with n, but remains relevant up to $n = 1000$. The MLRT is anti-conservative for moderate n ($n \leq 200$) and conservative for larger n, and its conservativism decreases with n. On the other hand, the two weighted versions of the D-test as well as the ADDS test keep their nominal size for all n. For this reason, the simple D-test and the penalized D-test should not be used with asymptotic quantiles unless the sample length is very large.

The relative power of the four remaining tests has been evaluated on two-component mixtures. As a result, each of the four tests has its merits.

The linearly weighted D-test (w1D) is best on upper contaminations. But it is performs rather weakly on fifty-fifty mixtures for moderately large sample sizes.

The quadratically weighted D-test (w2D) is among the best on fifty-fifty mixtures. It is slightly worse than the others on upper contaminations, and it breaks down on lower contaminations when $n \leq 200$.

The MLRT works well in detecting fifty-fifty mixtures, but is slightly outperformed by the others on upper, and even more on lower contaminations.

The ADDS clearly outperforms the others on lower contaminations when the scale ratio v is large. In all other situations it develops power not much less than the best of the four tests.

It should be noted that the above test approaches differ in that different classes of alternative hypotheses can be included. The D-tests as well as the MLRT are designed as tests against general k-component mixtures with a fixed k. The ADDS test is constructed to test against mixtures with any k and also against infinite mixtures. For power results with an

infinite mixture alternative, see (Mosler & Seidel, 2001).

As far as the overall performance is concerned, we conclude that the ADDS test shows always not much less and sometimes considerably more power than its competitors. Also, for larger n, the MLRT behaves uniformly well unless a lower contamination with large scale ratio has to be detected. Therefore, to test for homogeneity in an exponential mixture model, if no particular information on the alternative is available, the ADDS test appears to be a good choice. Other combinations of a specific mixture test with a general purpose goodness-of-fit test may prove similarly useful. A computer code of the ADDS test, written in 'R', can be obtained from the authors.

A final question is whether the ADDS test can be extended to other finite mixture models and how it compares with its competitors under different model settings. (Mosler & Scheicher, 2007) establish the ADDS test for homogeneity in a Weibull mixture model; they demonstrate that, under lower contaminations, the penalized LR and D-tests break completely down, while the ADDS test develops considerable power. In mixtures of symmetric distributions, like normal mixtures, DS procedures usually work well and, by construction, better than ADDS procedures; see the monographs by (Böhning, 2000) and (McLachlan & Peel, 2000).

TABLES OF THE ADDS TEST

The following table contains pairs of α-critical quantiles for the ADDS test, with significance levels $\alpha = 0.10, \alpha = 0.05$, and $\alpha = 0.01$. t_1 is a critical quantile of an AD test having significance level α_1, and t_2 is a critical quantile of a DS test having significance level α_2. (t_1, t_2) have been determined so that $\alpha_1 \approx \alpha_2$ and the level of the combined test is α. In particular, first a value t_1 has been fixed that depends on α but not on n, and then t_2 has been chosen from simulated ADDS test sizes to obtain the combined test size α. The simulations have been based on $N = 50000$ replications, and the results for t_1 have been slightly smoothed (by moving averages of length 2).

α	test component		n										
			10	20	30	40	50	60	80	100	150	200	250
0.1	AD	t_1	1.27										
	DS	t_2	1.36	1.48	1.55	1.58	1.60	1.64	1.67	1.69	1.69	1.68	1.68
0.05	AD	t_1	1.55										
	DS	t_2	1.85	2.00	2.10	2.13	2.13	2.15	2.18	2.18	2.16	2.15	2.16
0.01	AD	t_1	2.25										
	DS	t_2	3.25	3.52	3.56	3.57	3.53	3.49	3.40	3.28	3.25	3.21	3.15

α	test component		n									
			300	350	400	450	500	600	700	800	900	1000
0.1	AD	t_1	1.27									
	DS	t_2	1.69	1.68	1.67	1.67	1.67	1.66	1.65	1.65	1.65	1.66
0.05	AD	t_1	1.55									
	DS	t_2	2.14	2.11	2.09	2.10	2.10	2.07	2.06	2.05	2.06	2.11
0.01	AD	t_1	2.25									
	DS	t_2	3.11	3.03	2.95	2.92	2.89	2.86	2.84	2.83	2.85	2.96

Table 2: Critical values for the combination of AD and DS statistics, with the critical values of the AD component being independent of the sample size n and a moving average of length 2 for the DS component.

BIBLIOGRAPHY

Böhning, D. (2000). *Computer-assisted analysis of mixtures and applications.* London: Chapman & Hall.

Böhning, D., Schlattmann, P., & Lindsay, B. (1992). Computer-assisted analysis of mixtures (c.a.man): Statistical algorithms. *Biometrics*, *48*, 283–303.

Charnigo, R., & Sun, J. (2004). Testing homogeneity in a mixture distribution via the l^2 distance between competing models. *Journal of the American Statistical Society*, *99*, 488 – 498.

Chen, H., Chen, J., & Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society*, *B 63*, 19 – 29.

Lindsay, B. (1995). *Mixture models: Theory, geometry and applications.* Hayward, Cal.: Institute of Mathematical Statistics.

McLachlan, G. (1995). Mixtures - models and applications. In N. Balakrishnan & A. Basu (Eds.), *The exponential distribution* (pp. 307–323). Amsterdam: Gordon & Breach.

McLachlan, G., & Peel, D. (2000). *Finite mixture models.* New York: J. Wiley.

Mosler, K., & Scheicher, C. (2007). Homogeneity testing in a weibull mixture model. *Statistical Papers*. (To appear.)

Mosler, K., & Seidel, W. (2001). Testing for homogeneity in an exponential mixture model. *Australian and New Zealand Journal of Statistics*, *43*, 231–247.

Prentice, R., Kalbfleisch, J., Peterson, A., Flournoy, N., Farewell, N., & Breslow, V. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, *34*, 541–554.

Seidel, W., Mosler, K., & Alker, M. (2000). A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics*, *52*, 481–487.

Titterington, D., Smith, A., & Makov, U. (1985). *Statistical analysis of finite mixture distributions.* New York: Wiley.