Testing for Homogeneity in an Exponential Mixture Model*

Karl Mosler & Wilfried Seidel

In: Australian & New Zealand Journal of Statistics, 2001

Abstract

Diagnostic procedures are studied to test for homogeneity against unobserved heterogeneity in an exponential mixture model. The procedures include a dispersion score test, a likelihood ratio test, a moment likelihood approach and several goodness-of-fit tests. The empirical power of these tests is compared on a broad range of alternatives. We propose a new test, which combines the dispersion score test with a properly chosen goodness-of-fit procedure; its empirical power comes close to the power of the best of the other tests.

AMS 1991 *subject classifications.* Primary 62G10; secondary 62N05.
Key words and phrases. Mixture diagnosis, frailty model, survival analysis, hazard models, overdispersion, goodness-of-fit, likelihood tests.

1 Introduction

A duration variable, e.g. a lifetime or an unemployment spell, is observed, and we assume that its hazard rate is the product of a given hazard rate and a latent variable. The latent variable is positive, say an unobserved covariate of the unemployment spell, and may vary in the population or not. The probability distribution of the latent variable is regarded as a mixing distribution.

*This research has been partially sponsored by a grant of the Deutsche Forschungsgemeinschaft. We thank a referee and an Associate Editor for their helpful remarks and hints on additional literature. Corresponding author: Karl Mosler, Statistik und Ökonometrie, Universität Köln, D-50923 Köln, Germany (E-mail: mosler@wiso.uni-koeln.de)

Then the survival function of the observed duration appears as a mixture of powers of some baseline survival function with respect to the exponent and the model is called an *exponential mixture model* or *multiplicative frailty model*. If no grouping of the data but a known baseline survival function is given, the exponential mixture model can be transformed into a model where the baseline survival is exponential, that is, into a *mixture of exponentials model*.

A principal question is whether the latent variable really varies in the population, i.e. whether the population contains some unobserved heterogeneity. No unobserved heterogeneity means that the mixing distribution concentrates at a single point.

In this paper we present several diagnostic tests to decide whether an observed duration is an exponential mixture or not and compare their power on various mixture alternatives.

In many statistical applications mixture models arise as a natural and simple way to model population heterogeneity; see Lindsay (1995), Titterington, Smith & Makov (1985) and others. The assumption that the underlying distribution is a mixture of exponential distributions is widely invoked in the analysis of lifetime or, more general, duration data. This model arises from incomplete observation of an underlying conditional exponential model.

While the hazard rate of a pure exponential distribution is constant, the hazard of a mixture of exponentials is decreasing. 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. In many applications 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). Our model corresponds to the parametric proportional hazards model with unobserved heterogeneity (Lancaster, 1990) when no observed covariates are present.

Mendenhall & Hader (1958) apply the mixture of exponentials model to failure times of communication transmitter-receivers in aircrafts. For a similar early application to life testing of electronic tubes, see Kao (1959). Gordon (1990) uses the model to analyse a population of cancer patients which consists of two groups, those who die from the cancer and those who die from other causes. In economics mixtures of exponentials are applied, e.g., to the duration of unemployment and the theory of search; see Heckman (1991), Blossfeld, Hamerle & Meyer (1989). In biostatistics the multiplicative frailty model is used to analyze differences in lifelength between groups or individuals. See e.g. Hougaard (1984) and Aalen (1988). This has numerous applications.

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

In Section 2, we introduce the exponential mixture model. Section 3 presents four different diagnostic approaches: a score test for dispersion, a maximum likelihood approach, a moment likelihood procedure, and several goodness-of-fit tests. The power of these tests is compared in Section 4 against selected mixture distributions that are built of two and more components. We also investigate combined test procedures ('Reject the null hypothesis if at least one of two tests rejects.') and demonstrate that a test which combines the Anderson-Darling test with a dispersion score test has, in most cases, power close to the power of the best of all tests considered here. A similar result holds for a test combining the one-sided Kolmogorov-Smirnov test with a dispersion score test. Section 5 contains remarks on censoring and concludes the paper.

2 The exponential mixture model

Let T denote the duration, which is assumed to be a continuous nonnegative random variable. Let h_0 denote the given hazard, U the positive latent variable, and π the probability distribution of U. Then T, conditional on $U = u$, is assumed to have hazard rate $h(t|u) = h_0(t)/u$ and survival function

$$\Pr(T > t|U = u) = S(t|u) = S_0(t)^{1/u}, \quad t \geq 0, \tag{1}$$

where S_0 is the survival function corresponding to h_0, $S_0(t) = \exp(-\int_0^t h_0(y)dy)$. The unconditional survival function of T is

$$\Pr(T > t) = S(t) = \int_0^\infty S_0(t)^{\frac{1}{u}} \pi(du), \quad t \geq 0. \tag{2}$$

In other words, S is a mixture of powers of S_0 with π as latent distribution. This unconditional distribution of T is what we call an *exponential mixture*. If π is finite discrete, $\pi(\{u_j\}) = \pi_j$, $j = 1, \ldots, k$, then (2) specializes to

$$S(t) = \sum_{j=1}^{k} \pi_j \left(S_0(t)\right)^{1/u_j} \quad t \geq 0. \tag{3}$$

Let $\mathcal{P}$ denote the class of all distributions π on $\mathbb{R}_+ = \{t : t \geq 0\}$ that have finite fourth moments and $\int_0^\infty u d\pi(u) > 0$. Let $\mathcal{P}_k$ be the subclass of distributions that have positive mass at k or fewer points.

We assume that T can be observed but U cannot. This means that the data are not grouped according to the latent variable U. An observation T of the duration, randomly drawn from the population, has survival function (2).

We consider i.i.d. observations $T_1, \ldots, T_n$ which follow a distribution (2). The baseline survival S_0 or, equivalently, the baseline hazard h_0 is assumed to be known. We investigate and compare procedures to test for

$$H_0 : \pi(\{u\}) = 1 \quad \text{at some } \ u > 0, \tag{4}$$

i.e., $H_0 : \pi \in \mathcal{P}_1$, against various alternatives.

Our problem is closely related to the problem of detecting overdispersion in a one-parameter exponential family. In such a family the variance of a distribution is determined by its mean. Further, any mixture of distributions from the family is a dilation from the pure distribution that has the same mean as the mixture (Shaked, 1980), and therefore the mixture has larger variance than the pure distribution.

3 Different test approaches

In this section we present four different approaches to the above test problem. The tests are applied after a transformation of the data that turns the problem into one of testing for mixtures of exponential distributions. Consider

$$X = -\log\left(S_0(T)\right), \tag{5}$$

where log is the natural logarithm. If T has conditional survival function (1) then the transform X has a conditional survival function $\tilde{S}(x|u) = \exp(-x/u)$, $x \geq 0$, i.e. $X|u$ has an exponential distribution with mean u. Unconditionally, X has survival function

$$\tilde{S}(x) = \Pr(X > x) = \int_0^\infty e^{-x/u}\pi(du), \quad x \geq 0. \tag{6}$$

Thus the unconditional distribution P_X of X is a mixed exponential distribution with the same latent distribution π as before, and $\mathrm{E}(X) = u_0 = \int_0^\infty u\pi(du)$. It follows that a sample of i.i.d. variables $T_1, \ldots, T_n$ with common survival function (2) transforms into an i.i.d. sample $X_1, \ldots, X_n$ with common survival function (6), and we may confine our investigation to procedures that test the null hypothesis (4) based on a sample from a π-mixture of exponential distributions, parameterized by their means u.

Note that any mixture of exponentials has decreasing hazard rate (Shaked, 1980), while a simple exponential has constant hazard rate.

3.1 Dispersion score

As a mixture of exponentials is always more dispersed than the exponential distribution with the same mean, our first approach aims at detecting overdispersion with a dispersion score test. This method is discussed in detail in Lindsay (1995 chapter 4) and applied to survival data by Lancaster (1990 chapter 11) and others. We use a variant of Neyman and Scott's $C(\alpha)$ test (Neyman and Scott, 1966). The above H_0 is tested against $H_1 : \pi \in \mathcal{P} \setminus \mathcal{P}_1$. Consider

$$C_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 - \frac{1}{2n}\sum_{i=1}^{n} X_i^2, \quad \text{where } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{7}$$

This is the test statistic of Neyman's classic $C(\alpha)$ test, with $1/(n-1)$ in place of $1/n$.

Since u is the first moment of the exponential distribution $Exp(u)$ we get a nice interpretation of Neyman's test statistic: C_n is an unbiased estimator of the variance of π. Unbiased estimators of moments of π will be also needed in the moment likelihood statistic of Section 3.3 below. These can be obtained as follows: Conditionally on u, it holds for $k \geq 1$ that $\mathrm{E}(X^k|u) = k!u^k$, consequently

$$u^k = \mathrm{E}\left(\frac{1}{k!}X^k|u\right).$$

Let

$$m_k(\pi) = \int_0^\infty u^k \pi(du)$$

denote the k-th moment of π. A simple Fubini argument demonstrates that

$$\hat{m}_k = \frac{1}{n\ k!}\sum_{i=1}^{n} {X_i}^k \tag{8}$$

is an unbiased estimator of $m_k(\pi)$. In particular, in our model with U defined as in Section 2, we obtain that $\mathrm{var}(X) = \mathrm{E}(U^2) + \mathrm{var}(U)$ and, therefore, $\mathrm{var}(U) = \mathrm{var}(X) - m_2(\pi)/2 = \mathrm{E}(C_n)$.

Further, it can be shown that C_n has an asymptotically normal distribution,

$$n^{\frac{1}{2}}\left[C_n - \left(\frac{1}{2}\mu_2' - \mu^2\right)\right] \to N(0, \tau^2)$$

$$\text{with} \quad \tau^2 = \frac{1}{4}\mu_4' - 2\mu\mu_3' + 6\mu^2\mu_2' - \frac{1}{4}{\mu_2'}^2 - 4\mu^4.$$

Here μ_r denotes the r-th central moment of the unconditional distribution of X, and μ_r' the r-th noncentral moment. Under H_0 a straightforward calculation yields that, if $\pi(\{u_0\}) = 1$,

$$\mathrm{var}(C_n) = \frac{u_0^4}{n}\ \frac{n+1}{n-1}.$$

But, as is seen from simulations, for moderately sized n ($n \leq 1000$) and given u_0 the distribution of C_n is far from being normal. More precisely, it is asymmetric with negative mode and heavy right tail. Thus the asymptotic normal distribution of C_n (as well as the chi-squared distribution of normalized and squared C_n^2) should not be used for these sample sizes.

Note that the null distribution of C_n depends on u_0. Since $\mathrm{E}(X|u_0) = u_0$, we estimate u_0 by $\overline{X}$ and use the statistic

$$O_n = \frac{C_n}{\overline{X}^2}\left(\frac{n(n-1)}{n+1}\right)^{\frac{1}{2}}, \tag{9}$$

which under H_0 has distribution independent of u_0.

The quantiles of O_n have been determined by simulation and collected in Table 1. All tables are in the Appendix.

3.2 Maximum likelihood

For our test problem next we consider the likelihood ratio test of H_0: $\pi \in \mathcal{P}_1$ against H_1: $\pi \in \mathcal{P}_2$. This test is often proposed as a test for homogeneity, see for example Lindsay (1995 chapter 4). Moreover, it is the starting point in a standard iterative procedure to determine the number of components in a mixture model, namely to test for $\pi \in \mathcal{P}_k$ versus $\pi \in \mathcal{P}_{k+1}$ for $k = 1, 2, \ldots$ (McLachlan & Basford, 1988). In this framework the test is implicitly used to check for homogeneity.

For π in $\mathcal{P}_k$, the loglikelihood function at $x = (x_1, \ldots, x_n)$ amounts to

$$l(x|\pi) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} \pi_j \frac{1}{u_j} e^{-\frac{x_i}{u_j}} \right). \tag{10}$$

For testing homogeneity, we employ a *likelihood ratio* (LR) *statistic*

$$R_n = 2\log\lambda_n = 2l(x|\hat{\pi}^{(2)}) - 2l(x|\hat{\pi}^{(1)}), \tag{11}$$

where $\hat{\pi}^{(2)} \in \mathcal{P}_2$ is a likelihood estimate for π under H_1: $\pi \in \mathcal{P}_2$ and $\hat{\pi}^{(1)} \in \mathcal{P}_1$ is one for π under H_0: $\pi \in \mathcal{P}_1$.

It is well known (Ghosh & Sen, 1985; Titterington *et al.*, 1985) that for mixtures R_n does not have the usual chi-squared asymptotics.

Generally, in a mixture setting the loglikelihood function possesses many local maxima and must be numerically maximized. Therefore any evaluation of the test statistic is the result of a numerical procedure. An often neglected consequence is that each likelihood maximizing algorithm and each particular implementation of it defines a different test (Seidel, Mosler & Alker, 2000b). Critical values are obtained by simulation only, and, obviously, the same algorithm and implementation that have been employed for simulation of quantiles must also be used for every evaluation of the test statistic.

Moreover, different maximization algorithms may result in tests with different power. In order to calculate R_n, two exponential means u_1 and u_2 and a mixing proportion p have to be estimated under H_1. We use an EM algorithm starting at $u_1 = 0.5\overline{x}$, $u_2 = 1.5\overline{x}$ and $p = 0.5$. Table 2 exhibits simulated quantiles of R_n for this particular implementation of the EM algorithm. It is shown in Seidel, Mosler & Alker (2000a) that the strategy used here yields better power than other strategies, including a multistart strategy that comes close to global maximization of the likelihood. The

reason is that if the data are simulated from a homogeneous population, global maxima often correspond to spurious solutions of the likelihood equation. The strategy proposed here usually does not converge to such spurious maxima, so it leads to smaller quantiles than other strategies. On the other hand, under true mixture populations, the global maximum is often attained at a parameter value near the 'true' one, and this value is, in most cases, identified by our strategy.

Below we report power results of our LR test on different alternatives.

3.3 Moment likelihood

Historically, mixture problems have been first analyzed by moment methods and later by likelihood procedures, which are principally more efficient; see McLachlan & Basford (1988). However, in face of the computational problems and fallacies of the likelihood ratio test, a blend of both methods appears promising. Recently, Lindsay proposed moment likelihood methods for estimating mixtures of normals that have unknown but equal variances; see Furman & Lindsay (1994), who state that these methods develop nearly the same power as an LR test.

Let us test $H_0 : \pi \in \mathcal{P}_1$ against $H_1 : \pi \in \mathcal{P}_2$. If we insert moment estimators $\tilde{\pi}^{(2)}$ and $\tilde{\pi}^{(1)}$ in (11) a *moment likelihood ratio* (MomLR) statistic is obtained.

A distribution $\tilde{\pi}^{(k)} \in \mathcal{P}_k$ is called a *moment estimate* of $\pi \in \mathcal{P}$ (Lindsay, 1989) if its first $2k-1$ moments, $m_i(\tilde{\pi}^{(k)})$, equal the respective moment estimates $\hat{m}_i$ of π,

$$m_i(\tilde{\pi}^{(k)}) = \hat{m}_i, \quad i = 1, \ldots, 2k-1.$$

Here the first three noncentral moments of U are estimated by

$$\hat{m}_1 = \overline{x}, \quad \hat{m}_2 = \frac{1}{2}\,\frac{1}{n}\sum_{i=1}^{n} x_i^2, \quad \hat{m}_3 = \frac{1}{6}\,\frac{1}{n}\sum_{i=1}^{n} x_i^3.$$

Let ϵ_u denote the point measure at u. Obviously, $\tilde{\pi}^{(1)} = \epsilon_{\overline{x}}$ is a moment estimate of π under H_0. It can be shown that a moment estimate of $\tilde{\pi}^{(2)}$ of π under H_1,

$$\tilde{\pi}^{(2)} = \gamma\epsilon_{u_1} + (1-\gamma)\epsilon_{u_2} \text{ with } u_1 \neq u_2,\ 0 < \gamma < 1, \tag{12}$$

exists if and only $\hat{m}_2 - \hat{m}_1^2 > 0$. Recall that $\hat{m}_2 - \hat{m}_1^2$ is the moment estimate of the variance of π. Therefore, if $\hat{m}_2 - \hat{m}_1^2 \leq 0$, we accept H_0. If $\hat{m}_2 - \hat{m}_1^2 > 0$ holds let u_1 and u_2 be the solutions of the equation

$$u^2(\hat{m}_2 - \hat{m}_1^2) - u(\hat{m}_3 - \hat{m}_1\hat{m}_2) + \hat{m}_1\hat{m}_3 - \hat{m}_2^2 = 0.$$

Hence u_1, u_2 and γ are given by

$$u_{1,2} = \frac{\hat{m}_3 - \hat{m}_1\hat{m}_2 \mp \sqrt{4(\hat{m}_2 - \hat{m}_1^2)(\hat{m}_2^2 - \hat{m}_1\hat{m}_3) + (\hat{m}_1\hat{m}_2 - \hat{m}_3)^2}}{2(\hat{m}_2 - \hat{m}_1)^2},$$

and $\gamma = (\hat{m}_1 - u_1)/(u_2 - u_1)$.

Since the mean is positive and $u_1 \leq u_2$, we must have $u_2 > 0$ but u_1 may be negative. If $u_1 < 0$ we construct a modified moment estimate for $\pi^{(2)}$. In the distribution (12) we choose, in the limit, $u_1 = 0$ and determine u_2 and γ in a way that matches the first two moments. Then $u_2 = \hat{m}_2/\hat{m}_1$ and $\gamma = \hat{m}_1^2/\hat{m}_2$. The corresponding exponential mixture is defined as the limit where u_1 approaches 0.

Define

$$\psi(x_i, u_1, u_2, \gamma) = \log\left(\frac{\gamma}{u_1}e^{-x_i/u_1} + \frac{1-\gamma}{u_2}e^{-x_i/u_2}\right)$$

with the convention that $(1/0)e^{-x/0} = 0$. Then our moment likelihood statistic is given by

$$R_n^{mom} = \begin{cases} \sum_{i=1}^n (\log A_i - \log B_i) & \text{if } \hat{m}_2 - \hat{m}_1^2 > 0 \text{ and } u_1 > 0, \\ \sum_{i=1}^n (\log C_i - \log B_i) & \text{if } \hat{m}_2 - \hat{m}_1^2 > 0 \text{ and } u_1 \leq 0, \\ -\infty & \text{if } \hat{m}_2 - \hat{m}_1^2 \leq 0 \end{cases}, \qquad (13)$$

where $A_i = \psi(x_i, u_1, u_2, \gamma)$, $B_i = \psi(x_i, \hat{m}_1, 0, 1)$ and $C_i = \psi(x_i, 0, \hat{m}_2/\hat{m}_1, \hat{m}_1^2/\hat{m}_2)$. Simulated quantiles of this moment likelihood statistic are given in Table 3.

Remarks.
1. Note that, although the moment estimator $\tilde{\pi}^{(k)}$ is defined through estimated moments of the mixing distribution π, the defining equations are a rearrangement of terms of the usual sample moment formulae, that means, $\tilde{\pi}^{(k)}$ is a classical moment estimate.
2. The moment likelihood statistic R_n^{mom} is scale invariant.
3. It can be shown that in the second case R_n^{mom} is always negative, and therefore the null hypothesis is accepted.
4. Furman & Lindsay (1994) use a similar approach for normal mixtures. In a normal setting with σ^2 unknown the estimated means, u_i, of the mixtures are always in the parameter space, which is $\mathbb{R}$. Moreover, the estimated σ^2 may always be adapted to ensure $\hat{m}_2 - \hat{m}_1^2 \geq 0$.

3.4 Goodness-of-fit

Our last approach is to employ several universal goodness-of-fit procedures. We use two standard tests, Kolmogorov-Smirnov (KS) and Anderson-Darling (AD), and a test by Tiku (1980) that has been particularly advocated for mixture analysis; see Balakrishnan & Ambagaspitiya (1989).

Let $X_{(1)} \leq X_{(2)} \leq \ldots X_{(n)}$ denote the ordered sample and define $X_{(0)} = 0$. Then, under H_0,

$$V_i = \frac{\sum_{k=1}^i (n+1-k)(X_{(k)} - X_{(k-1)})}{\sum_{k=1}^n (n+1-k)(X_{(k)} - X_{(k-1)})}, \qquad i = 1, 2, \ldots, n-1, \qquad (14)$$

are the order statistics of $n-1$ variables i.i.d. uniformly on $[0,1]$; see, e.g., D'Agostino & Stephens (1986). The one-sided KS statistic is given by

$$D_n^+ = \left(\sqrt{n-1} + 0.12 + \frac{0.11}{\sqrt{n-1}}\right) \quad \max_{i=1,\ldots,n-1} \left(\frac{i}{n-1} - V_i\right). \tag{15}$$

The finite sample correction is due to D'Agostino & Stephens (1986), where critical values are found as well. We employ the one-sided version of the test as this is more powerful against general decreasing hazard rate alternatives.

The AD statistic is

$$A_n^2 = \left(1 + \frac{0.6}{n}\right) \quad \left(n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left(\log\left(1 - e^{-X_{(i)}/\overline{X}}\right) + \frac{X_{(i)}}{\overline{X}}\right)\right). \tag{16}$$

Again the finite sample correction is taken from D'Agostino & Stephens (1986), who also provide selected quantiles. H_0 is rejected if A_n^2 is too large.

Tiku's (1980) test statistic for our problem is given by

$$T_n = \sum_{i=1}^{n-1} \frac{n-i}{n-1} \frac{(n+1-i)(X_{(i)} - X_{(i-1)})}{\sum_{k=1}^{n}(n+1-k)(X_{(k)} - X_{(k-1)})}. \tag{17}$$

Its distribution exactly equals the distribution of the mean of $n-1$ i.i.d. variables that are uniform on $[0,1]$, and therefore converges rapidly to a normal with mean $1/2$ and variance $(12(n-1))^{-1}$. H_0 is rejected if $\sqrt{12(n-1)}|T_n - 1/2|$ is larger than the standard normal quantile for a two-sided test. Note that T_n is the mean of the order statistics (14) on which the above KS statistic is based. It follows that Tiku's test cannot have more power than the KS test.

4 Comparison of power and a new test

The power of each of the above tests has been evaluated on a broad range of mixture alternatives having density

$$f(x) = \int_0^\infty \frac{1}{u} \exp\left(-\frac{x}{u}\right) \pi(du), \quad x \geq 0, \tag{18}$$

with various mixing distributions π, n ranging from 10 to 10 000 and α from 0.01 to 0.1. In the sequel we report empirical power results, which have been obtained from 20 000 replications of the tests (10 000 of the LR and AD-DS tests) under different alternatives. In terms of accuracy, this means: If, at a given alternative, the empirical power of the LR test is at least 0.012 below the power of another test then it may be fairly concluded (in the worst case, with significance level 0.05 and less) that the LR test is the less powerful one.

The results reported here are part of a large power study. Similar results hold for many other values of the parameters α and n; see Mosler, Seidel & Jaschinger (1997).

4.1 Power on two-component mixtures

If $\pi \in \mathcal{P}_2$, due to the scale invariance of the test procedures, we need only consider

$$f(x) = (1-\epsilon)\exp(-x) + \frac{\epsilon}{v}\exp\left(-\frac{x}{v}\right), \quad x \geq 0, \tag{19}$$

with some $0 \leq \epsilon \leq 1$ and $v \geq 1$, which is the ratio of the two means. Here we consider $\epsilon \in \{0.1, 0.5, 0.9\}$ and $1 \leq v \leq 50$. The parameter $\epsilon = 0.5$ indicates a fifty-fifty mixture of exponentials having means 1 and v. The case $\epsilon = 0.1$ may be interpreted as an exponential distribution that has mean 1 and is slightly contaminated (at a ten percent rate) by another exponential distribution that has mean $v > 1$. We call this an *upper contamination.* On the other hand, $\epsilon = 0.9$ characterizes a mixture of two exponentials having different means 1 and v, $1 < v$, where the larger weight ($\epsilon = 0.9$) is put on the higher mean. Due to the scale invariance, this is tantamount to considering an exponential distribution with mean 1 that is slightly contaminated by an exponential distribution having mean $1/v$ with contamination rate $1-\epsilon = 0.1$. Hence the case $\epsilon = 0.9$ is regarded as a *lower contamination.*

First let us compare the power of goodness-of-fit tests. We refer to the simulation results of Balakrishnan & Ambagaspitiya (1989) who found that Tiku's test on mixtures of exponentials is on the whole more powerful than Durbin's test and the test by Shapiro and Wilks. Since Tiku's test can never be better than the above KS test, the tests by Durbin and Shapiro-Wilks are also inferior to it.

Figures 1 to 3 exhibit the power of the two remaining goodness-of-fit tests on two-component mixtures, depending on the mean ratio v, for different values of ϵ, α and n. Their power is contrasted with that of the dispersion score test and the moment likelihood procedure. From Figure 1 we see that an upper contamination is best detected by the dispersion score test, and nearly as well by the moment likelihood procedure. Next best appears the KS test, which dominates the AD test if v is not too large. This holds for both small and large samples.

Figure 2 demonstrates that for lower contaminations the ranking is reversed. Here the AD test proves to be the best. Only if the mean ratio v is below 7 the AD test is slightly outperformed by the KS. But this difference in power is hardly relevant for samples of size 1000 and more. The dispersion score test is always worse than the KS and, for v larger than 2, also worse than the AD test. In Figure 3, for fifty-fifty mixtures the two goodness-of-fit tests show no large differences in power; at $n = 100$ the KS test is slightly better than the AD.

The behaviour of the moment likelihood procedure looks rather strange in Figure 3. In many cases, even for large samples, we get the paradoxical result that the power decreases when the ratio of the means increases. While at upper contaminations the procedure is best, at fifty-fifty and other mixtures its power does not increase and even decreases with v. This paradoxical behaviour is not restricted to small

samples; see also Figure 5. It is due to the fact that, for these mixtures, case two of (13) often occurs, so that the moment likelihood test accepts the null hypothesis.

4.2 A new test approach based on goodness-of-fit and overdispersion

We have shown that, for two-component mixtures and v not too small, either the DS test or the AD test have power close to the best power among the tests considered so far. If v is small, the KS test slightly outperforms the AD test.

A detailed analysis of the simulated test decisions reveals that there are many samples on which the AD test rejects the null hypothesis while the dispersion score test does not, and there are many other samples on which the dispersion score test rejects but the AD does not. We therefore introduce a combination of the dispersion score test with the AD test (AD-DS):

$$\text{Reject } H_0 \text{ if } O_n \geq t_1 \text{ or } A_n^2 \geq t_2.$$

In choosing the two critical values there are several possibilities to achieve a given α. Here we have determined t_1 and t_2 so that the AD test and the dispersion test each individually obtain the same size $\alpha^* < \alpha$ and the combination of them obtains size α. Table 4 presents critical values t_1 and t_2 for various n and α.

Figure 4 shows the surprising result that, for mixture alternatives of two components, the power of the combined test comes very close to the maximum power of the two tests on which it is based.

In Figure 5 the combined AD-DS test is contrasted with the likelihood ratio test and the moment likelihood test on different two-component mixtures.

For upper contaminations and fifty-fifty mixtures the combined AD-DS test has effectively the same power as the LR test. For lower contaminations, at level $\alpha = 0.05$, the AD-DS attains at least 78 percent of the LR power for any v and outperforms the LR test when v is large ($v \geq 30$). The moment likelihood again satisfactorily detects only upper contaminations. At levels other than $\alpha = 0.05$, which are not exhibited here, the qualitatitive behaviour of the tests is similar, but the tests differ in their relative power for lower contaminations. E.g. at $\alpha = 0.01$ the combined test attains 50 percent of the power of the LR test for a lower contamination.

In the same way we have combined the KS test with the dispersion score test. Table 5 exhibits the critical values. On two-component mixtures this combined test, which we name the KS-DS test, has similar power as the AD-DS test. When $v < 7$, it is slightly more powerful than the latter test. But for lower contaminations and $v \geq 7$ the AD-DS test develops considerably more power. More results on the power of the KS-DS test are found in Mosler *et al.* (1997).

Figure 1a ($\alpha = 0.05$, $n = 100$).

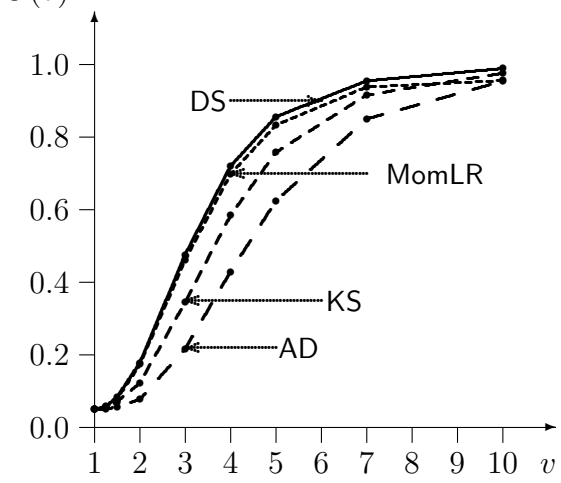

Figure 1b ($\alpha = 0.05$, $n = 1000$).

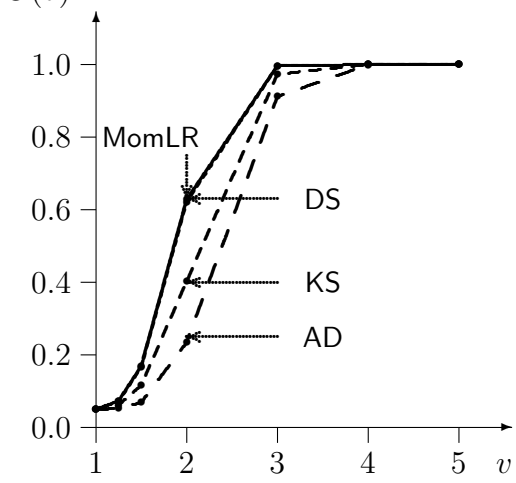

Figure 1c ($\alpha = 0.01$, $n = 100$).

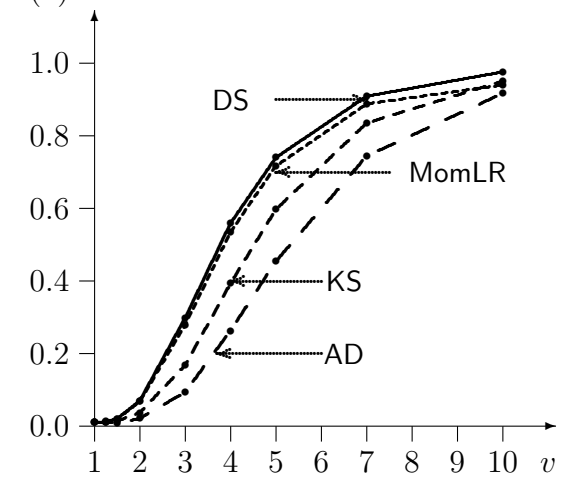

Figure 1d ($\alpha = 0.01$, $n = 1000$).

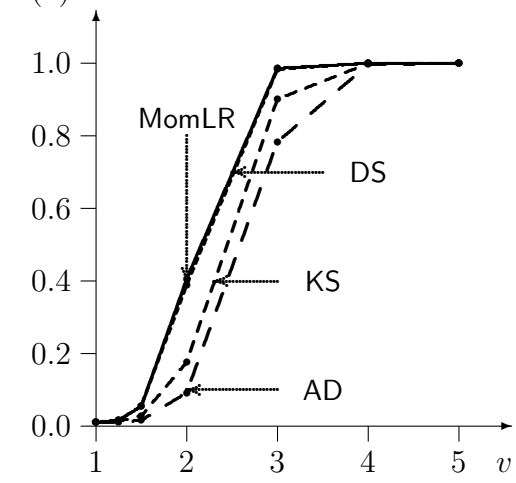

Figure 1: Power $G(v)$ of overdispersion, moment likelihood and two goodness-of-fit tests on alternative $f(x) = 0.9e^{-x} + 0.1\frac{1}{v}e^{-\frac{x}{v}}$ (upper contamination).

Figure 2a ($\alpha = 0.05$, $n = 100$).

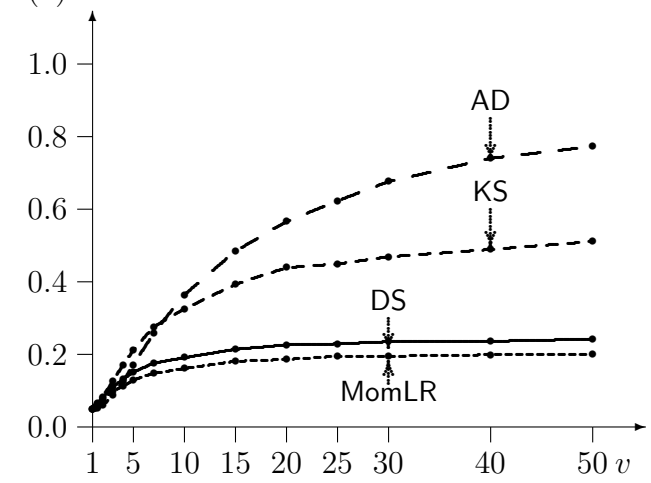

Figure 2b ($\alpha = 0.05$, $n = 1000$).

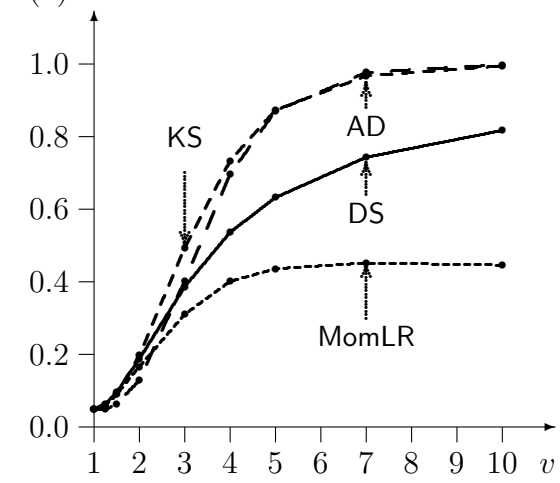

Figure 2c ($\alpha = 0.01$, $n = 100$).

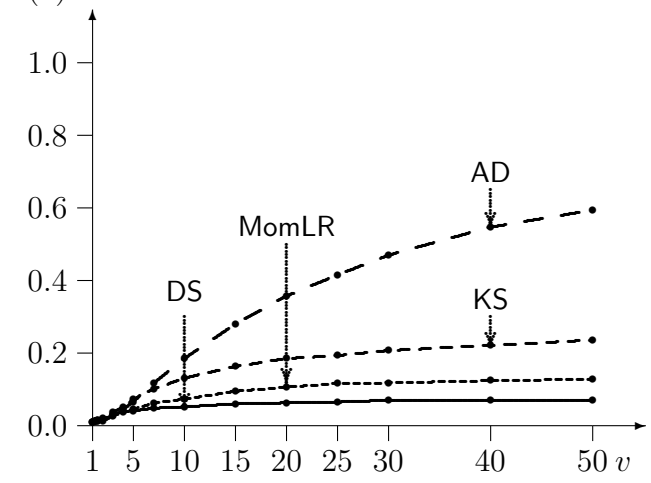

Figure 2d ($\alpha = 0.01$, $n = 1000$).

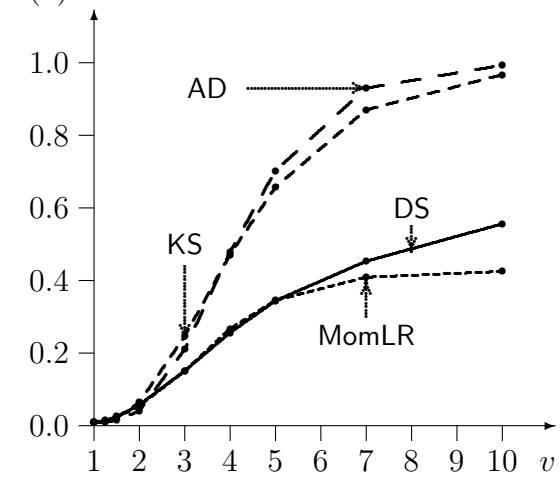

Figure 2: Power $G(v)$ of overdispersion, moment likelihood and two goodness-of-fit tests on alternative $f(x) = 0.1ve^{-vx} + 0.9e^{-x}$ (lower contamination).

Figure 3a ($\alpha = 0.05$, $n = 100$).

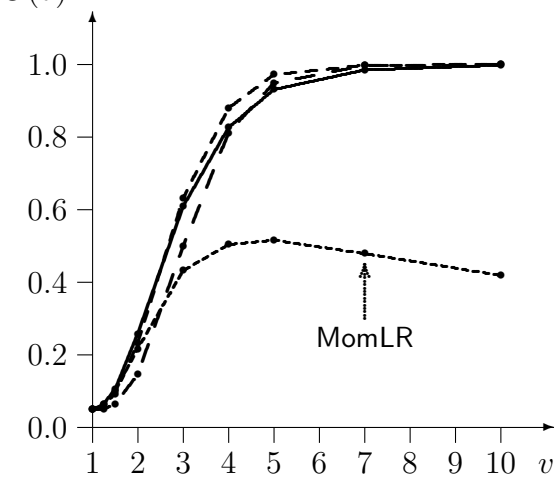

Figure 3b ($\alpha = 0.05$, $n = 1000$).

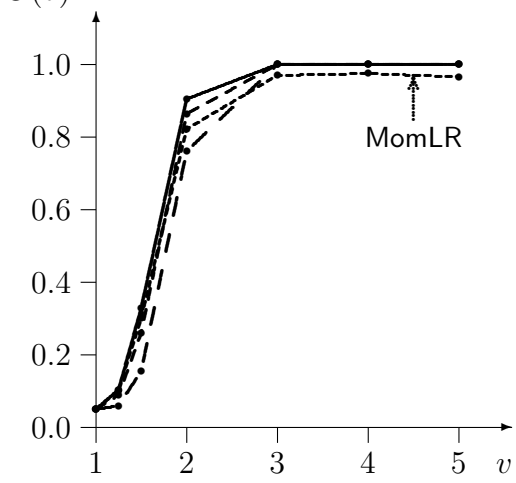

Figure 3c ($\alpha = 0.01$, $n = 100$).

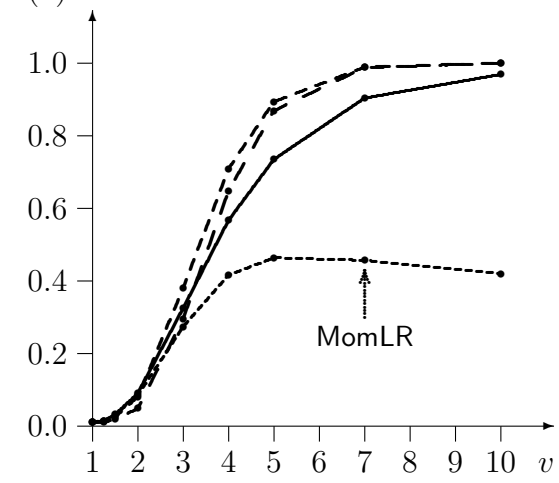

Figure 3d ($\alpha = 0.01$, $n = 1000$).

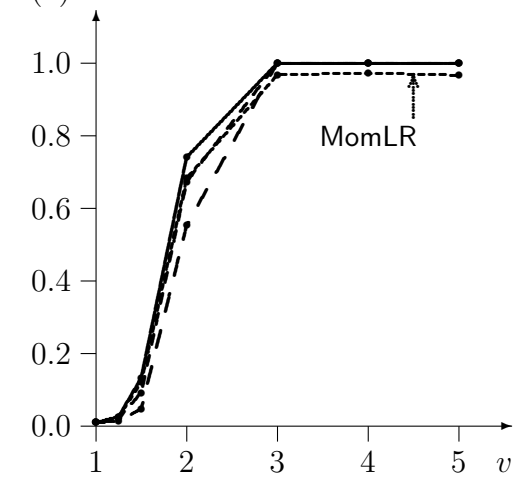

Figure 3: Power $G(v)$ of overdispersion, moment likelihood and two goodness-of-fit tests on alternative $f(x) = 0.5e^{-x} + 0.5\frac{1}{v}e^{-\frac{x}{v}}$ (fifty-fifty mixture).

Figure 4a ($\epsilon = 0.1$, $n = 100$).

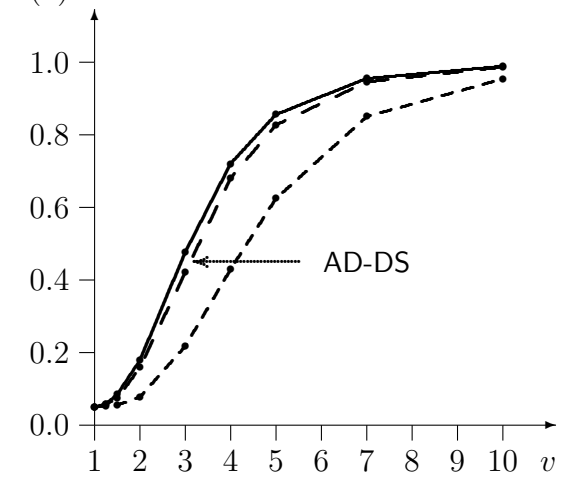

Figure 4b ($\epsilon = 0.1$, $n = 1000$).

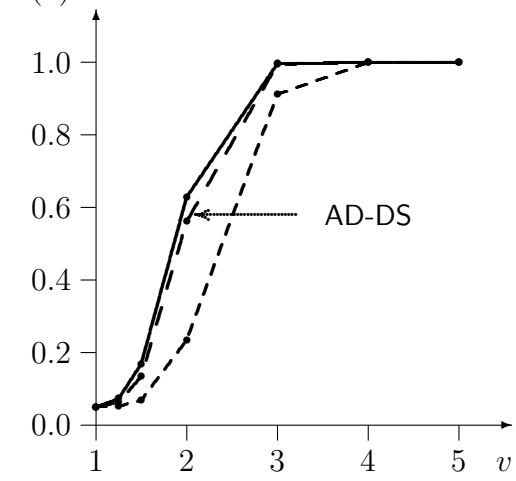

Figure 4c ($\epsilon = 0.9$, $n = 100$).

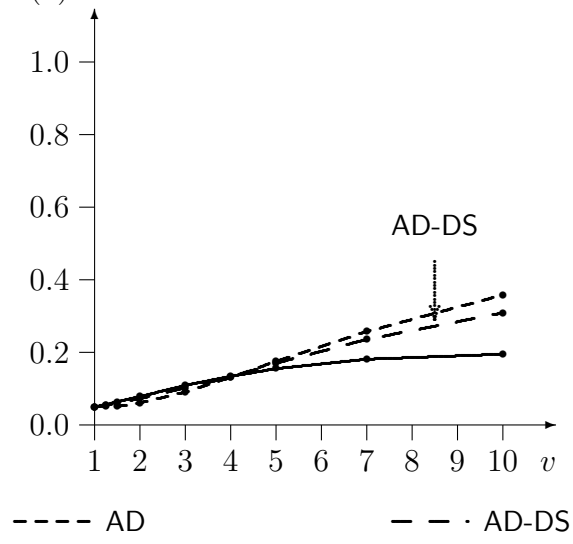

Figure 4d ($\epsilon = 0.9$, $n = 1000$).

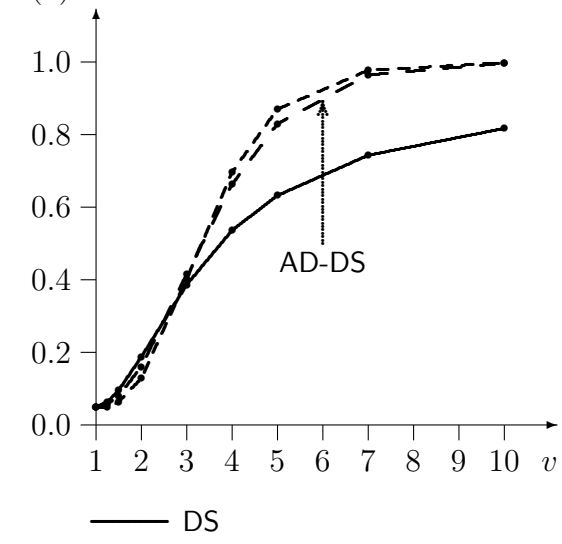

Figure 4: Power $G(v)$ of overdispersion (DS) and Anderson-Darling (AD) tests and their combination (AD-DS) on alternative $f(x) = (1-\epsilon)ve^{-vx} + \epsilon\ e^{-x}$, upper contamination ($\epsilon = 0.1$), lower contamination ($\epsilon = 0.9$); $\alpha = 0.05$.

Figure 5a
($\alpha = 0.05,\ n = 200,\ \epsilon = 0.1$).
$G(v)$

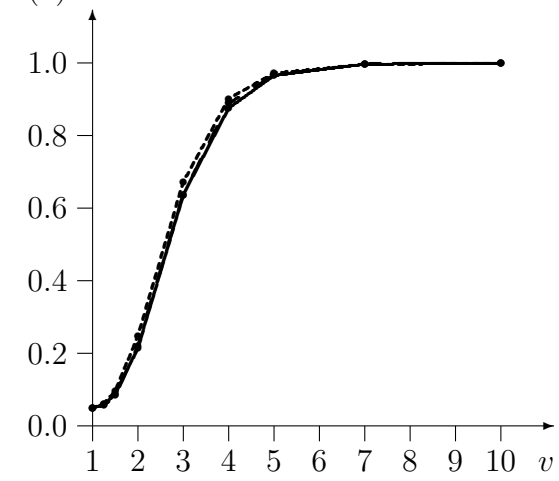

Figure 5b
($\alpha = 0.05,\ n = 200,\ \epsilon = 0.5$).
$G(v)$

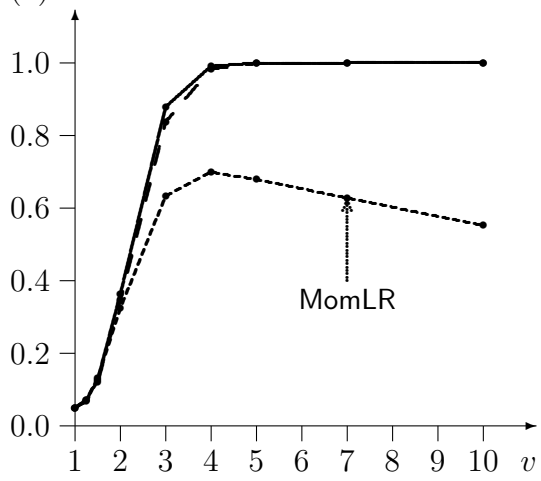

Figure 5c
($\alpha = 0.05,\ n = 200,\ \epsilon = 0.9$).
$G(v)$

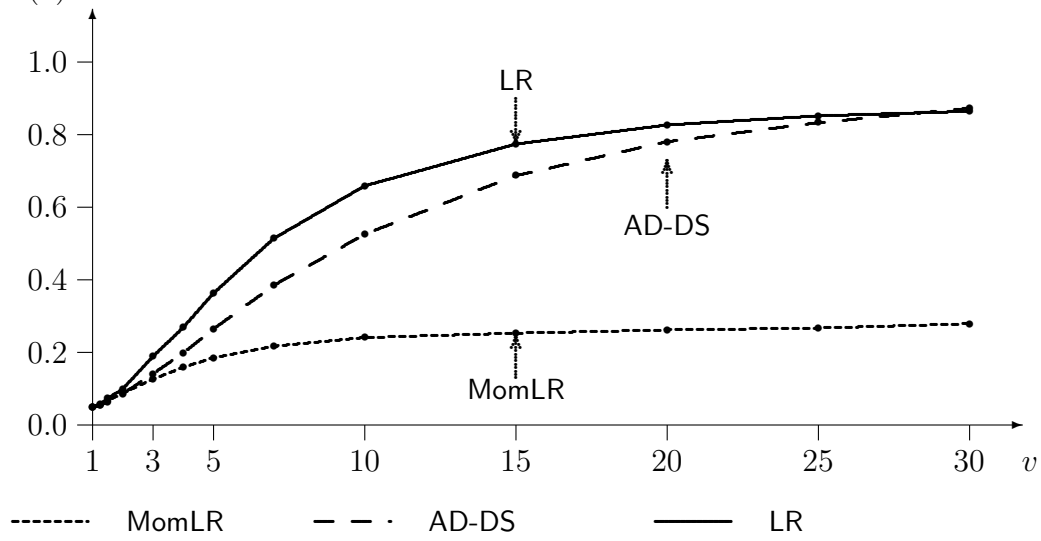

Figure 5: Power $G(v)$ of likelihood ratio and other tests on alternative $f(x) = (1-\epsilon)ve^{-vx} + \epsilon\ e^{-x}$, upper contamination ($\epsilon = 0.1$), fifty-fifty mixture ($\epsilon = 0.5$), lower contamination ($\epsilon = 0.9$).

4.3 Power for k-component mixtures, $k > 2$

We illustrate the power of the KS-DS procedure on higher mixture alternatives. (Here the AD-DS procedure yields similar results.) For three-component mixtures, Figure 6 exhibits the power of the LR (homogeneity against two-point mixtures), the KS and the combined KS-DS tests. If the three mixture weights are equal ($\epsilon_1 = \epsilon_2 = \epsilon_3 = 1/3$), the KS test and the combination have the same power, and the LR test is clearly inferior. If the weights are unequal (with $\epsilon_1 > \epsilon_3$), the combined test shows more power than the KS, and the LR remains inferior. Similar results have been obtained for mixtures of five components.

Finally, the power on three continuous mixture alternatives has been investigated. In Figure 7 the power of the LR test, the KS-DS test and the moment likelihood test is shown for sample sizes that range from 10 to 1000. Here the LR test and the combined test have similar power over the whole range of sample sizes, while the moment likelihood procedure develops comparable power only for very large n.

5 Remarks and conclusions

Different diagnostic procedures have been studied to test whether a given sample arises from an exponential mixture or a non-mixed distribution in a one-parameter family.

To simplify the exposition and keep computer simulations in workable limits we have neglected censoring. For the necessary adjustments of the above diagnostic procedures to various situations of censored or truncated data, we refer to the literature: See Schumacher (1984), Shorrack and Wellner (1986) and Andersen *et al.* (1993 chapter 5) for Kolmogorov-Smirnov and Anderson-Darling tests under random censoring and more general sampling situations, Lancaster (1990), Gray (1995), Commenges & Andersen (1995) and Jaggia (1997) for the dispersion score test under various censoring schemes, and Andersen *et al.* (1993 chapter 9) for likelihood inference in non-informatively censored frailty models. The comparative evaluation of mixture diagnostic procedures under these sampling situations is up to further investigations.

However the case of uncensored data is the starting and reference point of any such analysis. For this case we conclude from theoretical analysis and empirical power studies:

1. A properly chosen goodness-of-fit test works very well: either a one-sided Kolmogorov-Smirnov (KS) test after transformation to order statistics of a uniform distribution or an Anderson-Darling (AD) test.

2. For higher component mixtures, an LR test which tests homogeneity against two-point mixtures can have considerably less power than the KS procedure. Recall that

in the usual iterative test procedure to determine the number of mixture components this LR test forms the initial step. In order to increase the procedure's probability to identify a higher component mixture we suggest replacing the initial LR test by a KS test.

3. The KS test proposed here is clearly better than several goodness-of-fit tests that have been investigated and advocated in the literature. These are the tests by Shapiro-Wilks, Durbin and Tiku; see Balakrishnan & Ambagaspittya (1989) and the literature quoted there.

4. A moment likelihood approach, which has been successfully used in the diagnosis of normal mixtures, does not work well in exponential mixtures since the moment estimator is often infeasible.

5. The dispersion score test, which is known to be locally most powerful in any direction, makes an optimal use of the local information on the parameters. (By contrast, the likelihood ratio also uses information on the parameters that is more remote from the estimated u_0; see Lindsay (1995, pp 70 ff). However, for fifty-fifty mixtures and moderate sample sizes the dispersion score test is outperformed by goodness-of-fit procedures, and for lower contaminations it has much less power even if n is large.

6. The strength of the dispersion score test and that of a properly chosen goodness-of-fit test may be combined. If the two tests reject the null hypothesis in largely disjoint situations, the power of the test 'reject the null if one of the two tests rejects' comes close to the maximum of the two test powers. Our combined AD-DS procedure has this property; it shows a very good overall power on broad classes of alternatives.

7. Even if we restrict ourselves to the detection of mixtures of two components, the power of the LR test for homogeneity against two-component mixtures is not always the best. While for upper contaminations and fifty-fifty mixtures the AD-DS test does equally well, for lower contaminations the AD-DS test achieves at least 78 per cent of the power of the LR test and for large v, $v \geq 30$, it is even more powerful. In view of the computational load and the numerical fallacies of the LR test (see Seidel *et al.*, 2000b) the AD-DS is a good practical alternative to the LR test.

8. A similar combination of the score test with the KS test (KS-DS test) may be employed to detect two-component mixtures where the ratio v of means is not too large, $v < 7$. For these v the KS-DS procedure is slightly more powerful than AD-DS. However, in practical applications, if the class of mixture alternatives is not restricted, we suggest using the combined AD-DS procedure.

A final remark refers to the interpretation of a significant test result. In an ungrouped data setting, without covariates, an exponential mixture model which includes an unspecified baseline hazard rate is not identifiable since any such model

is equivalent to a proper non-mixed distribution model. In this situation it is impossible to test for mixture homogeneity if the baseline hazard is not specified at all.

In our setting, however, one has to be careful with the interpretation of a significant test result. A rejection of the null hypothesis can signal heterogeneity, a misspecified baseline hazard, omitted covariates, or other misspecifications. Clearly such tests have a more forceful interpretation with grouped data. (For a score test in a grouped data setting, see e.g. Liang, 1987).

Moreover, with ungrouped data, artifactual components caused by spurious solutions of the likelihood equations are often observed in mixture models (McLachlan and Peel, 2000a). These may lead to rejection, although the data come from a homogeneous population. However, if simulated (in contrast to asymptotic, say) critical values are used and if the same algorithm for calculating the test statistic is applied when simulating the quantiles and when performing the test, this error is a type I error which is accounted for by the level of the test. Note, however, that for likelihood ratio tests, certain strategies to maximize the likelihood function are more sensitive to spurious solutions than others (Section 3.2), and that this effect can be used to increase the power of the test on the one hand or to design a more robust test on the other.

Figure 6a
($\alpha = 0.05$, $n = 100$, $\epsilon_1 = \frac{1}{3}$, $\epsilon_2 = \frac{1}{3}$).
$G(\beta)$

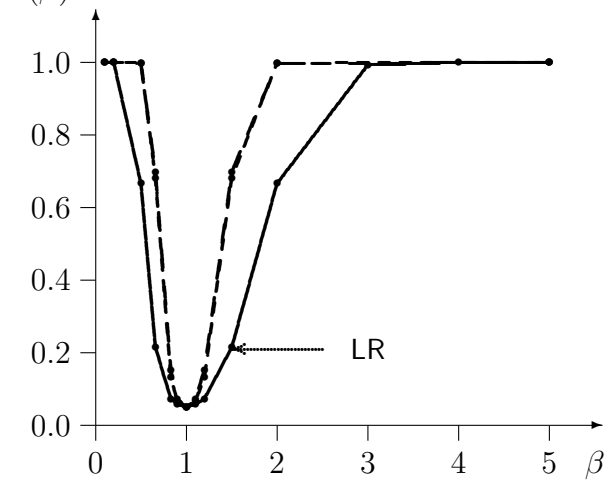

Figure 6b
($\alpha = 0.05$, $n = 1000$, $\epsilon_1 = \frac{1}{3}$, $\epsilon_2 = \frac{1}{3}$).
$G(\beta)$

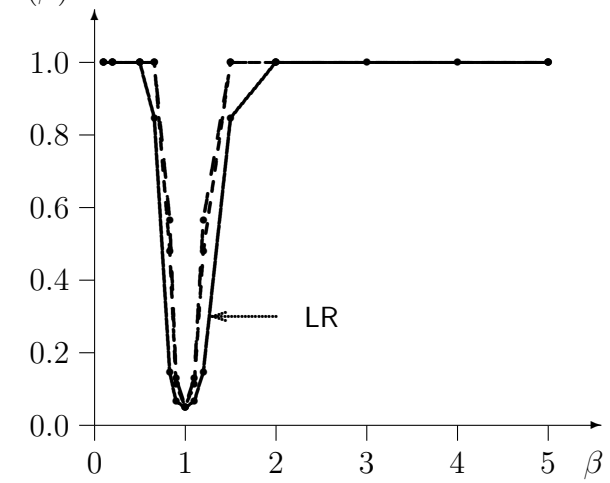

Figure 6c
($\alpha = 0.05$, $n = 100$, $\epsilon_1 = \frac{1}{5}$, $\epsilon_2 = \frac{3}{4}$).
$G(\beta)$

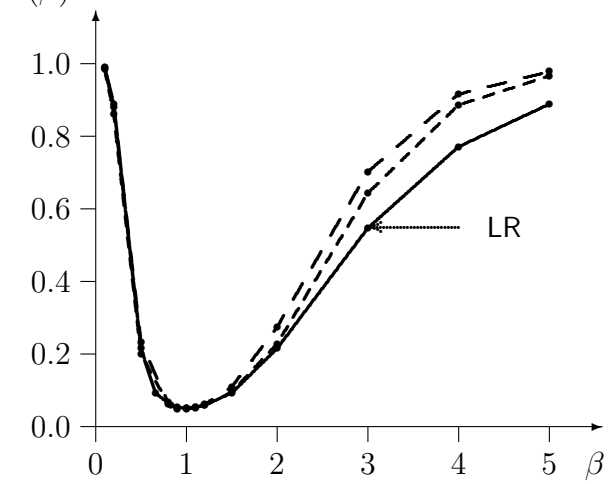

Figure 6d
($\alpha = 0.05$, $n = 1000$, $\epsilon_1 = \frac{1}{5}$, $\epsilon_2 = \frac{3}{4}$).
$G(\beta)$

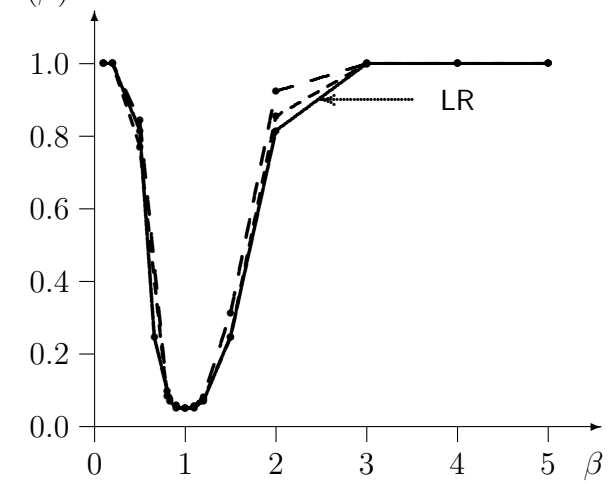

Figure 6: Power of Kolmogorov-Smirnov (KS), likelihood ratio (LR) and combined KS-overdispersion (KS-DS) tests on three-component mixtures
$f(x) = \epsilon_1 \beta e^{-\beta x} + \epsilon_2 e^{-x} + (1 - \epsilon_1 - \epsilon_2) \frac{1}{\beta} e^{-\frac{x}{\beta}}$.

Figure 7a
(π uniform on $[0,1]$).

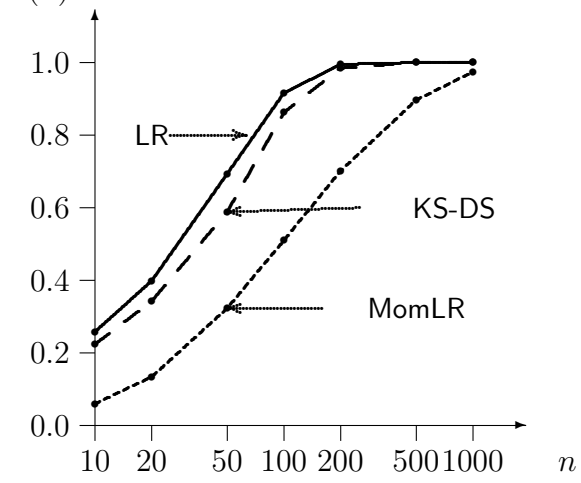

Figure 7b
(π inverse uniform on $[0,1]$).

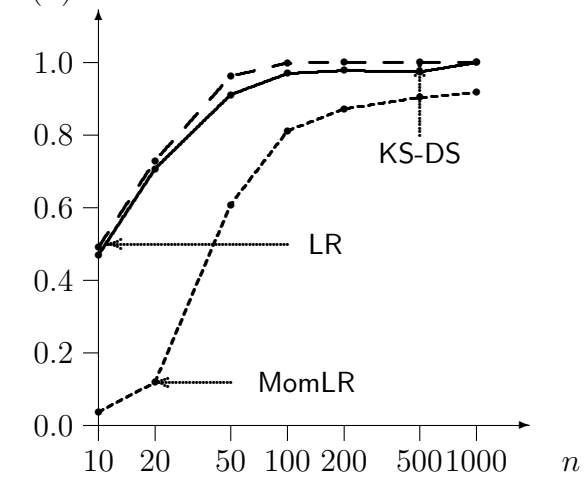

Figure 7c
(π exponential with mean 1).

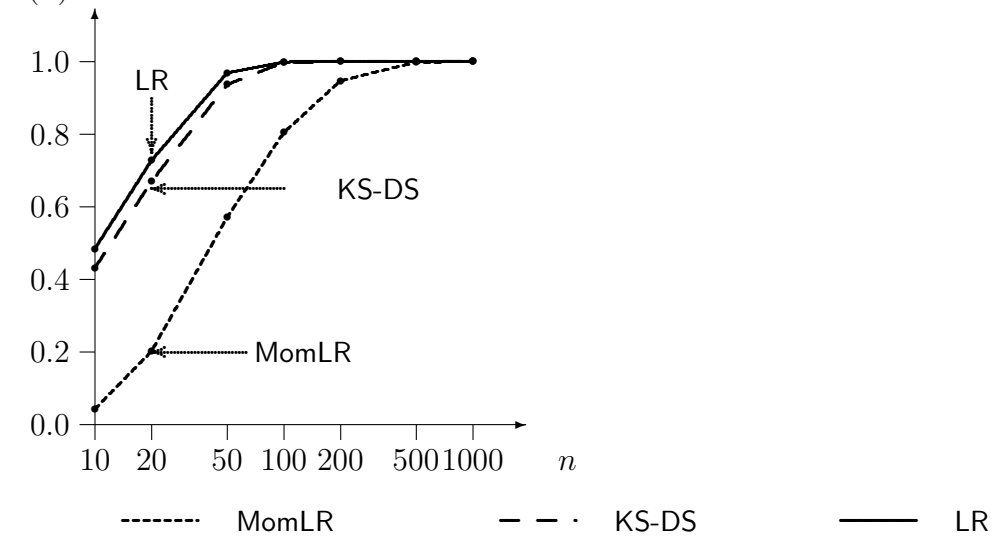

Figure 7: Power of likelihood ratio (LR), moment likelihood (MomLR) and KS-overdispersion tests on continuous mixtures where (a) π is continuous uniform on $[0,1]$, (b) π is the distribution of $\frac{1}{U}$ and U has uniform distribution on $[0,1]$, (c) π is exponential with mean 1; $\alpha = 0.05$.

Appendix

The quantiles in Tables 1 and 3 have been calculated from 10^5 replications, in Table 2 from 10^4 replications. The quantiles in Table 4 have been chosen so that the simulated size (from 50000 replications) of the test differs from the intended size by at most 10 percent.

Table 1: Quantiles of the dispersion score statistic O_n

α	n						
	10	20	50	100	200	500	1000
0.1	0.961	1.093	1.205	1.254	1.284	1.296	1.298
0.05	1.445	1.608	1.728	1.750	1.759	1.731	1.717
0.01	2.640	2.910	2.964	2.891	2.778	2.636	2.562

Table 2: Quantiles for the likelihood ratio statistic R_n.

α	n					
	10	20	50	100	200	1000
0.1	1.73	2.04	2.33	2.44	2.59	2.60
0.05	3.07	3.37	3.71	3.80	3.97	4.09
0.25	4.40	4.75	5.06	5.23	5.37	5.51
0.01	6.29	6.54	6.90	7.09	6.99	7.41

Table 3: Quantiles of the moment likelihood statistic R_n^{mom}

α	n						
	10	20	50	100	200	500	1000
0.1	0.430	0.818	1.150	1.415	1.663	1.969	2.154
0.05	0.698	1.102	1.822	2.445	3.025	3.679	4.060
0.01	0.955	2.279	6.341	9.518	12.800	16.687	18.902

Table 4: Quantiles for the combination AD-DS of O_n and A_n^2 statistics

α	test component	n						
		10	20	50	100	200	500	1000
0.1	A_n^2	1.41	1.36	1.36	1.36	1.35	1.35	1.36
	O_n	1.17	1.37	1.49	1.55	1.57	1.57	1.55
0.05	A_n^2	1.66	1.62	1.60	1.60	1.64	1.62	1.62
	O_n	1.69	1.92	2.07	2.08	2.03	2.00	1.98
0.01	A_n^2	2.36	2.26	2.26	2.26	2.26	2.26	2.26
	O_n	2.94	3.41	3.41	3.30	3.13	2.92	2.80

Table 5: Quantiles for the combination KS-DS of O_n and D_n^+ statistics

α	test component	n						
		10	20	50	100	200	500	1000
0.1	D_n^+	1.13	1.14	1.16	1.16	1.17	1.17	1.17
	O_n	1.17	1.37	1.49	1.55	1.57	1.57	1.55
0.05	D_n^+	1.29	1.30	1.31	1.31	1.32	1.32	1.32
	O_n	1.69	1.92	2.07	2.08	2.03	2.00	1.98
0.01	D_n^+	1.58	1.58	1.61	1.61	1.61	1.61	1.61
	O_n	2.94	3.41	3.41	3.30	3.13	2.92	2.85

References

Aalen, O.O. (1988). Heterogeneity in survival analysis. *Statistics in Medicine* **7** 1121–1137.

Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag, New York.

Balakrishnan, N. & Ambagaspitiya, R.S. (1989). An empirical power comparison of three tests of exponentiality under mixture- and outlier-models. *Biometrical J.* **31** 49–66.

Blossfeld, H.-P., Hamerle, A. & Meyer, K.-U. (1989). *Event History Analysis. Statistical Theory and Application in the Social Sciences.* Erlbaum, Hillsdale, NJ.

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

COMMENGES, D. & ANDERSEN, P.K. (1995). Score test of homogeneity for survival data. *Lifetime Data Analysis* **1** 145–160.

D'AGOSTINO, R.B. & STEPHENS, M.A. (1986). *Goodness-of-fit techniques.* Dekker, New York.

FURMAN, W.D. & LINDSAY, B.G. (1994). Testing for the number of components in a mixture of normal distributions using moment estimators. *Computational Statistics & Data Analysis* **17** 473–492.

GHOSH, J.K. & SEN, P.K. (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. *Proceedings of the Berkeley Conference in Honor of J. Neyman & J. Kiefer, Vol 2,* (L.M. LeCam, R.A. Olshen, eds.), Wadsworth, Belmont CA, pp.789–806.

GORDON, N.H. (1990). Application of the theory of finite mixtures for the estimation of 'cure' rates of treated cancer patients. *Statistics in Medicine* **9** 397–407.

GRAY, R.J. (1995). Tests for variation over groups in survival data. *J. American Statistical Association* **90** 198–203.

HECKMAN, J.J. (1991). A nonparametric method-of-moments estimator for the mixture-of-exponentials model and the mixture-of-geometrics model. *Nonparametric and Semiparametric Methods in Econometrics and Statistics, Proceedings of the 5th International Symposium in Economic Theory and Econometrics* (Barnett, Powell, Tauchen, eds.), Cambridge, 243–258.

HOUGAARD, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika* **71** 75–83.

JAGGIA, S. (1997). Alternative forms of the score test for heterogeneity in a censored exponential model. *Review of Economics and Statistics* **79** 340–343.

KAO, J.H.K. (1959). A graphical estimation of mixed Weibull parameters in life-testing of electron tubes. *Technometrics* **1** 389–407.

LANCASTER, T. (1990). *The Econometric Analysis of Transition Data.* Cambridge University Press, Cambridge.

LIANG, K-Y. (1987). A locally most powerful test for homogeneity with many strata. *Biometrika* **74** 259–264.

LINDSAY, B.G. (1989). Moment matrices: Applications in mixtures. *Annals of Statistics* **17** 722–740.

LINDSAY, B.G. (1995). *Mixture Models: Theory, Geometry and Applications.* Institute of Mathematical Statistics, Hayward, Cal.

MCLACHLAN, G.J. (1995). Mixtures - models and applications. *The Exponential Distribution* (N. Balakrishnan, A.P. Basu, eds.), Gordon & Breach, Amsterdam, pp. 307–323.

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

McLachlan, G.J. & Peel, D. (2000) Computing Issues for the EM Algorithm in Mixture Models, *Proceedings of the 31st Interface*, pp. 421-430.

Mendenhall, W. & Hader, R.J. (1958). Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. *Biometrika* **45** 504–520.

Mosler, K., Seidel, W. & Jaschinger, C. (1997). A Power Comparison of Homogeneity Tests in Mixtures of Exponentials. *Discussion Papers in Statistics and Econometrics* **3/97**. Universität Köln.

Neyman, J. & Scott, E.L. (1966). On the use of $C(\alpha)$ optimal tests of composite hypotheses. *Bull. Inst. Int. Statist.* **41 I** 477–497.

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

Schumacher, M. (1984). Two-sample tests of Cramér-von Mises- and Kolmogorov-Smirnov-type for randomly censored data. *International Statistical Review* **52** 263–281.

Seidel, W., Mosler, K. & Alker, M. (2000a). Likelihood ratio tests based on subglobal optimization: A power comparison in exponential mixture models. *Statistical Papers* **41** 85–98.

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

Shaked, M. (1980). On mixtures from exponential families. *Journal of the Royal Statistical Society, Ser. B* **42** 192–198.

Shorack, G.R. & Wellner, J.A. (1986). *Empirical Processes.* Wiley, New York.

Tiku, M.L. (1980). Goodness of fit statistics based on the spacings of complete or censored samples. *Australian Journal of Statistics* **22** 260–275.

Titterington, D.M., SMITH, A.F.M. & Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Wiley, New York.

Karl Mosler
Seminar für Wirtschafts- und Sozialstatistik
Universität zu Köln
D–50923 Köln, Germany

Wilfried Seidel
Institut für Statistik und Quantitative Ökonomik
Universität der Bundeswehr Hamburg
D–22039 Hamburg, Germany