

Testing for univariate two-component Gaussian mixture in practice

Didier Chauveau, Bernard Garel, Sabine Mercier

▶ **To cite this version:**

Didier Chauveau, Bernard Garel, Sabine Mercier. Testing for univariate two-component Gaussian mixture in practice. 2018. hal-01659771v2

HAL Id: hal-01659771
https://hal.science/hal-01659771v2

Preprint submitted on 6 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Testing for univariate two-component Gaussian mixture in practice*

Didier Chauveau†, Bernard Garel‡, Sabine Mercier§

December 5, 2018

Abstract. We consider univariate Gaussian mixtures theory and applications, and particularly the problem of testing the null hypothesis of homogeneity (one component) against two components. Several approaches have been proposed in the literature during the last decades. We focus on two different techniques, one based on the Likelihood-Ratio Test (LRT), and another one based on estimation of the parameters of the mixture grounded on some specific adaptation of the well-known EM algorithm often called the EM-test. We propose in particular a novel methodology allowing application of the LRT in actual situations, by plugging-in estimates that are assumed known in asymptotic setup. We aim to provide useful comparisons between different techniques, together with guidelines for practitioners in order to enable them to use theoretical advances for analysing actual data of realistic sample sizes. We finally illustrate these methods in an application to real data corresponding to the number of days between two events concerning ovarian response and lambing for ewes.

AMS 1991 subject classification: Primary 62F03; 62E20. Secondary 62F05.

Key words and phrases: Mixture models; likelihood ratio test; EM tests; Gaussian process.

1 Introduction

The aim of always producing the better models for data analysis can partly explain the present craze for probability distributions which can be written as a mixture. Mixture models are able to help in many circumstances. Indeed, whenever a population is constituted of K homogeneous sub populations, a K-component mixture can be proposed as an attractive model for this population. However there exist other important reasons which justify the increasing use of these models. Recently published books are entirely devoted to mixture of distributions. In particular, the books by Everitt and Hand (1981), Titterington et al. (1985), McLachlan and

*This research has been partially supported by the *Projet de Recherche d'Intérêt Régional* DURAREP2, reference 2011-00064290.

†Institut Denis Poisson, Université d'Orléans, Université de Tours, CNRS, Route de Chartres, BP 6759, 45067 Orléans cedex 2, FRANCE.didier.chauveau@univ-orleans.fr

‡Université fédérale de Toulouse-Midi Pyrénées, ENSEEIHT, 2 rue Camichel, 31071 Toulouse Cédex 7, France; garel@math.univ-toulouse.fr

§UFR Sciences Espace Société, Université Jean Jaurès, 5 allées A. Machado, 31058 Toulouse Cedex 9. mercier@univ-tlse2.fr

Basford (1988), Lindsay (1995), McLachlan and Krishnan (1997), McLachlan and Peel (2000), Böhning (2000), Frühwirth-Schnatter (2006), Schlattmann (2009) contributed to an already rich bibliography. We also have to stress that from the theoretical point of view, analysis of mixture involves many mathematical topics such as estimation and maximization for non regular models, Bayesian analysis, asymptotic distribution, stochastic processes, and so on. Among the many problems raised by mixture we find the non-identifiability of parameters, the degeneracy of the Fisher information matrix around particular points or the non-differentiability with respect to another parametrization. Then it is not surprising that we obtain non-standard asymptotic distributions for testing problems.

Beginnings of mixture models go back to the 19th century mainly with the contributions by A. Quetelet, L.A. Bertillon, S. Newcomb and K. Pearson. Behind the writings of Quetelet (1846) we find the idea that a normal distribution can be generated by a great number of other normal distributions. Analysing the heights of 9002 conscripts, Bertillon (1874) and Bertillon (1876) noted that the graphical representation of these heights gave two modes which constituted a surprise. Then he claimed that this phenomenon was due to the presence of two distinct ethnic groups. The figure he presented seems to be the first graphical representation leading to the assumption of a normal mixture. Newcomb (1882) and Newcomb (1886) addressed the problem of outliers in astronomical data. He observed that the tails of the distribution were fatter than the normal ones. He explained this non-normality by the combination of data with different scales and so, invented the contaminated normal distribution. Pearson (1894) analysed data that the zoologist Walter F. Weldon submitted to him, in particular crab forehead sizes. In his 1894 paper he graphically showed the evidence of a mixture. For him, to adjust a two-component Gaussian mixture is equivalent to carve a skewed curve into two Gaussian distributions. A way of doing so is to estimate five parameters from the five first moments. This is feasible because seeing that a mixture is a convex combination of densities, its moments are convex combination of the moments of these densities. Then Pearson found its famous ninth degree equation, a negative root of which is necessary to solve the problem. Pearson's contribution is generally thought of as the starting point of the analysis of mixtures. Then, during quite a while, the research on mixture models concentrated around improvements of this method of moments.

These contributions are related to the problem of estimation in mixture model. Another very important issue is the determination of the number of components of the mixture, which is the topic addressed in this paper. Graphical procedures have been developed. Simple examination of the histogram can bring some information. A more elaborated method has been proposed by Bhattacharya (1967). The method starts from two statements. First the logarithm of a normal density is a concave quadratic in the variable, so that its derivative is linear with negative slope. Then, when there is a lot of data and the grouping imposed by the histogram is quite fine, the histogram heights are proportional to the density. Thus, a plot of first differences of the logarithms of the histogram frequencies should display a sequence of negatively sloped linear plots, one corresponding to each components.

This concern can also be treated as a testing problem: a set of tests of k components against $k+1$ components, for instance using the likelihood ratio. After Wilks (1938), Chernoff (1954) gave the asymptotic distribution of the likelihood ratio test in the case of a regular model. Above, we gave a few reasons why mixtures do not belong to regular models. However, a few researchers began to privilege likelihood ratio for determining the number of components of a mixture and particularly the problem of testing H_0 : homogeneity ($k = 1$), against a two-component mixture. This is the very problem upon which we are going to work.

Likelihood Ratio Test

Consider, for example, the model :

$$(1-\pi)\mathcal{N}(\mu_1,\sigma_1^2)+\pi\mathcal{N}(\mu_2,\sigma_2^2) \tag{1}$$

which characterizes a univariate two-component Gaussian mixture both on the means and on the variances. In this model, homogeneity can be specified by $\pi = 0$ or 1, or by $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$. In a first attemp to test H_0 against this mixture, conjectures and simulation results about the distribution of the likelihood ratio statistic have been published.

Wolfe (1971) suggested that $(n-1-m)\lambda_n/n$ is approximately distributed as a χ^2_{2m}, where n is the sample size, λ_n is the usual likelihood ratio test statistic (LRTS) and m is the number of parameters which are different for the two components of the mixture. This gives a χ^2_2 distribution for a univariate normal mixture with equal variance and a χ^2_4 distribution for a normal mixture with different means and different variances.

In the case of univariate normal mixture with unknown but common variance, see Model **M9** below, McLachlan (1987) and Thode et al. (1988) suggested that for $n \leq 1000$ the distribution of λ_n is close to a χ^2_2, the latter having a less heavy tail. In the case of a normal mixture with different means and different variances, McLachlan (1987) found that a χ^2_6 fits well for a sample size $n = 100$. Hall and Stewart (1985) suggested using the restriction

$$\min_{1\leq i,j\leq 2}(\sigma_i/\sigma_j) \geq c \geq 0$$

and Feng and McCulloch (1996) used $\min(\sigma_1^2,\sigma_2^2) \geq c' \geq 0$. For $n = 100$, they found a distribution between a χ^2_4 and a χ^2_5 when $c' = 10^{-6}$ and between a χ^2_5 and a χ^2_6 when $c' = 10^{-10}$. The preceding results, which are given without other restrictions on the parameters, rely on Monte Carlo simulations and concern essentially finite sample distributions. Indeed, without assuming that the mixture parameter belongs to a bounded interval, λ_n tends to infinity in probability when n goes to infinity.

If we now consider asymptotic results, a first stage has been undertaken by Redner (1981). He proved that if W denotes a fixed neighbourhood of the set Γ corresponding to H_0 in the global parameter space, associated to the general model (2), then the probability that the maximum likelihood estimator (MLE) is found in W tends to one when n goes to infinity. Redner calls it convergence of the MLE in the topology of the quotient space obtained by collapsing Γ into a single point; see Ghosh and Sen (1985). The first correct expression of the asymptotic distribution was given by Ghosh and Sen (1985) for a mixture model with two components:

$$h(x;\pi,\mu_1,\mu_2) = (1-\pi)g(x,\mu_1)+\pi g(x,\mu_2), \tag{2}$$

where the component density g is general, but satisfying some regularity conditions; $1-\pi$ and $\pi \in [0,1]$ are the respective weights and μ_1 (*resp.* μ_2) is the parameter of the first (*resp.* the second) component; π, μ_1 and μ_2 are unknown. First they needed the assumption that the mixture parameters μ_1 and μ_2 belong to a bounded interval. Indeed, Hartigan (1985) proved that a statistic close to the LRTS for testing homogeneity against a Gaussian mixture of the means converges towards infinity in probability when n tends to infinity if the range of the unknown mean is unbounded. Bickel and Chernoff (1993) revisited this problem and showed that if the parameter set is unbounded, Hartigan's statistic approaches infinity with order $\log\log n$. Note that the equivalence between Hartigan's statistic and the LRTS, when H_0 is reduced to a single scalar parameter, has been proved not before quite recently by Liu and Shao (2003).

Ghosh and Sen also imposed a separation condition on the parameters of the mixture mainly in order to restore identifiability and to get an answer. They have assumed that $|\mu_2 - \mu_1| \geq c_0 > 0$. Therefore, under this constraint, H_0 is described by $\pi = 0$ or $\pi = 1$. Removing this separation condition presented a real challenge and many statisticians offered a solution, for instance, Dacunha-Castelle and Gassiat (1997), Lemdani and Pons (1999), Liu and Shao (2003), Garel (2005a). Garel (2001), Chen and Chen (2001), Garel and Goussanou (2002), Liu and Shao (2004) addressed specific mixtures in the Gaussian case. The LR approach has also being studied in a recent work Maciejowska (2013): the author proposes new hypothesis to test the homogeneity against two-component mixture model which allow to avoid the problem of identifiability.

When the parameter upon which relies the mixture is multivariate, for instance in the model (1) above, where the mixture relies both on the means and on the variances, the asymptotic distribution of the likelihood ratio is related to a Gaussian random field and the computation of percentile points becomes tricky or impossible. That is why other tests or methods have been proposed in order to assess the number of components. Let us mention, here, the seminal contribution by Donoho and Jin (2004) which provides conditions allowing a separation of the null and the alternative hypotheses. They called the associated procedure: *"The Higher Criticism Method"*, not investigated in this paper. If we are only interested by the detection of a gap from normality, classical tests of normality could be used; but it is easy to make evidence of a lack of power of these tests in our context.

EM-Test

For a two-component mixture model, Chen et al. (2001), Chen et al. (2004) modified the LRTS and derived its limiting distribution. They used a penalized likelihood, with a penalty depending on the mixture proportion π. Then, Li et al. (2009) proposed an EM-test for homogeneity, that Chen and Li (2009) mentioned in the case of a two-component Gaussian mixture. Chen and Li (2011) propose a refined method for computing a tuning parameter in the penalty used in previous papers on this EM-test approach. Then Chen et al. (2012) proposed an EM-test for testing the null hypothesis of some arbitrary fixed order under a finite mixture model.

Among the advantages of the EM-tests claimed by the authors, we find a limiting distribution under the null hypothesis which does not depend on the finiteness of the Fisher information on the mixing parameter direction. For Model (1) above, Chen and Li (2009) characterize H_0 by:

$$\pi(1-\pi) = 0 \qquad \text{or} \qquad (\mu_1, \sigma_1^2) = (\mu_2, \sigma_2^2).$$

Then the penalized log likelihood function is defined by:

$$pl_n(\pi, \mu_1, \mu_2, \sigma_1, \sigma_2) = l_n(\pi, \mu_1, \mu_2, \sigma_1, \sigma_2) + p(\pi) + p_n(\sigma_1) + p_n(\sigma_2),$$

where l_n is the usual log likelihood function, $p_n(\sigma)$ is bounded when σ is large, but goes to negative infinity as σ goes to 0, and $p(\pi)$ is maximized at $\pi = 0.5$ and goes to negative infinity as π goes to 0 or 1.

To construct the EM-test, first choose a set of $\pi_j \in (0, 0.5]$, $j = 1, ..., J$, and a positive integer K (quite often $\pi_j \in \{0.1, 0.3, 0.5\}$ and $K = 2$ or 3). For each $j = 1, 2, ..., J$, let $\pi_j^{(1)} = \pi_j$ then compute:

$$(\mu_{j1}^{(1)}, \mu_{j2}^{(1)}, \sigma_{j1}^{(1)}, \sigma_{j2}^{(1)}) = \underset{\mu_1, \mu_2, \sigma_1, \sigma_2}{\arg\max}\ pl_n(\pi_j^{(1)}, \mu_1, \mu_2, \sigma_1, \sigma_2).$$

Let $f(x;\mu,\sigma)$ be the density function of the normal $\mathcal{N}(\mu,\sigma^2)$. For $i=1,2,...,n$ and the current k, use the E-step to compute:

$$w_{ij}^{(k)} = \frac{\pi_j^{(k)} f(x_i;\mu_{j2}^{(k)},\sigma_{j2}^{(k)})}{(1-\pi_j^{(k)})f(x_i;\mu_{j1}^{(k)},\sigma_{j1}^{(k)}) + \pi_j^{(k)} f(x_i;\mu_{j2}^{(k)},\sigma_{j2}^{(k)})}$$

and use the M-step to update π and other parameters:

$$\pi_j^{(k+1)} = \arg\max_{\pi} \left\{ \left(n - \sum_{i=1}^{n} w_{ij}^{(k)} \right) \log(1-\pi) + \sum_{i=1}^{n} w_{ij}^{(k)} \log(\pi) + p(\pi) \right\}$$

and

$$(\mu_{j1}^{(k+1)},\mu_{j2}^{(k+1)},\sigma_{j1}^{(k+1)},\sigma_{j2}^{(k+1)}) = \arg\max_{\mu_1,\mu_2,\sigma_1,\sigma_2} \sum_{h=1}^{2} \left[\sum_{i=1}^{n} w_{ij}^{(k)} \log[f(x_i;\mu_h,\sigma_h)] + p_n(\sigma_h) \right].$$

The E-step and the M-step are iterated $K-1$ times. For each k and j define:

$$M_n^{(k)}(\pi_j) = 2\{pl_n(\pi_j^{(k)},\mu_{j1}^{(k)},\mu_{j2}^{(k)},\sigma_{j1}^{(k)},\sigma_{j2}^{(k)}) - pl_n(1/2,\widehat{\mu}_0,\widehat{\mu}_0,\widehat{\sigma}_0,\widehat{\sigma}_0)\}$$

where $(\widehat{\mu}_0,\widehat{\sigma}_0) = \arg\max_{\mu,\sigma} pl_n(1/2,\mu,\mu,\sigma,\sigma)$. The EM-test statistic is then defined as:

$$EM_n^{(K)} = max\{M_n^{(K)}(\pi_j) : j=1,...,J\}.$$

Reject the null hypothesis when $EM_n^{(K)}$ exceeds some critical value to be determined.

We focus on these two different techniques, namely the LRT and the EM-test based on the EM algorithm. We aim to provide useful comparisons between these techniques in several (more or less general) Gaussian mixture models, and guidelines for practitioners in order to enable them to use these methods for analysing actual data of realistic sample sizes.

Since EM-tests are intended to provide answers to pratictioners, and since they require some sort of heavy computations, the availability of public codes is important. Several versions of these EM-tests have been made publicly available in the recent MixtureInf package (Li et al., 2016) for the R statistical software (R Core Team, 2016). Two successive versions of this package have been proposed, and we used in this work the most recent one available (version 1.1, March 2016)[1]. In this current version, the function `emtest.norm` is dedicated to the test of the order of a normal mixture model. The linked references are precisely Chen and Li (2009), that was limited to the homogeneous null model vs. a two-component mixture, and Chen et al. (2012) which generalize this EM-test approach to a mixture of arbitrary order under the null hypothesis. We have also developed numerical procedures for the EM approach for some of the models that are not available in the MixtureInf package, such as Models **M2** and **M9**, in Table 1.

On the computational side, since there exists (up to our knowledge) no public codes for the LRT approach, we develop numerical procedures that will be publicly available in an upcoming version of the mixtools package (Benaglia et al., 2009) for the R statistical software (R Core Team, 2016). We compare these LRT codes with some of the codes proposed in the MixtureInf package (Li et al., 2016) (see above).

[1]We first tried another version (1.0-1 published in 2015) in which we found several errors or inconsistencies, such as negative values for the EM-test statistic and p-values 1 for Model **M8** in Table 1.

The rest of the paper is organized as follows: Sections 2 to 5 are dedicated to further presentation of models and analyses for models **M2**, **M6**, **M8** and **M9** respectively. Section 6 presents some applications based on actual data collected for a research project from the French National Institute for Agricultural Research (INRA)[2]: we study the so-called "ram effect" on data corresponding to number of days between two events concerning ovarian response and lambing for ewes of several kinds. For such data, a mixture is sometimes suspected for biological reasons, but with not great evidence coming from the empirical distribution. Section 7, a discussion summarizes the results and derives some practical suggestions for users.

2 Description of the models and study of Model M2

Table 1 describes the models studied from the perspective of testing homogeneity vs. mixture in the literature, where the model indices (M1, M2, ...) are borrowed from previous literature. The models we actually investigate in this paper are in boldface.

Model M1 is a contaminated model. Generally, g is a smooth density, with derivatives at least up to the second order; the value of α_0 is known. As we concentrate on the Gaussian case, we study model **M2** which is a special case of M1. Like **M2**, Model M3 is a Gaussian contaminated model but on the variance. It has been studied in his Ph.D. by Saint Pierre (2003). Model M4 is also a contaminated model, but testing a three-component mixture. It has been addressed by Garel and Goussanou (2002).

Model M5 is one of the models upon which a lot of reaserchers spent a lot of time. The density g is a smooth function with derivatives up to the fourth oder for Dacunha-Castelle and Gassiat (1997) but only up to the second order for Garel (2005a). Model **M6** is the corresponding Gaussian case where the parameter is the mean. Model M7 is the case where the parameter is the variance. Model **M8** is the general model for a two-component Gaussian mixture with respect to the mean and the variance. Models **M9** and M10 are mixture models with the presence of a structural parameter, the value of which is unknown. For all these models, Garel (2001) proved or conjectured the LRT statistic, without separation condition. Tabulations are given in Garel (2005b).

2.1 Model M2

This simplest model is the standard normal $\mathcal{N}(0,1)$ contaminated by a normal distribution shifted by a mean μ,

$$H_0 : \mathcal{N}(0,1) \quad \text{vs.} \quad H_1 : (1-\pi)\mathcal{N}(0,1) + \pi\mathcal{N}(\mu,1). \tag{3}$$

This model is obviously rather artificial from a practical point of view. We start our study with it since it is the first time, up to our knowledge, that the actual quantiles for finite (realistic) sample size n are compared to the asymptotic quantiles obtained by Garel (2001). The Likelihood Ratio Test statistic (LRT) proposed by Garel (2001), Theorem 2.1, is

$$\lambda_n = \sup_{\mu\in[-a,a]\setminus\{0\}} T_n^2(\mu)\, \mathbb{1}_{\{T_n(\mu)\geq 0\}} \tag{4}$$

[2] *Projet de Recherche d'Intérêt Régional* DURAREP2.

Table 1: Description of the models studied from the perspective of testing homogeneity vs. mixture in the literature (models investigated in this paper are in boldface).

Models	H_0	H_1
1- Contaminated Models		
M1	$g(\alpha_0)$ $\alpha \in A \subset R$	$(1-\pi)g(\alpha_0) + \pi g(\alpha)$
M2	$\mathcal{N}(0,1)$	$(1-\pi)\mathcal{N}(0,1) + \pi\mathcal{N}(\mu,1)$ $\mu \in A \subset R$
M3	$\mathcal{N}(0,1)$	$(1-\pi)\mathcal{N}(0,1) + \pi\mathcal{N}(0,\sigma^2)$ $\sigma^2 \in [a,A] \subset]0,2[$
M4	$\mathcal{N}(0,1)$	$(1-\pi_1-\pi_2)\mathcal{N}(0,1)+$ $\pi_1\mathcal{N}(\mu_1,1) + \pi_2\mathcal{N}(\mu_2,1)$
2- One population against two		
M5	$g(x,\alpha)$ $\alpha \in A \subset R$	$(1-\pi)g(x,\alpha_1) + \pi g(x,\alpha_2)$
M6	$\mathcal{N}(\mu,1)$ $\mu \in A \subset R$	$(1-\pi)\mathcal{N}(\mu_1,1) + \pi\mathcal{N}(\mu_2,1)$
M7	$\mathcal{N}(0,\sigma^2)$	$(1-\pi)\mathcal{N}(0,\sigma_1^2) + \pi\mathcal{N}(0,\sigma_2^2)$ $\sigma_2^2 < 2\sigma_1^2$
M8	$\mathcal{N}(\mu,\sigma^2)$	$(1-\pi)\mathcal{N}(\mu_1,\sigma_1^2) + \pi\mathcal{N}(\mu_2,\sigma_2^2)$
3- Presence of a structural parameter		
M9	$\mathcal{N}(\mu,\sigma^2)$	$(1-\pi)\mathcal{N}(\mu_1,\sigma^2) + \pi\mathcal{N}(\mu_2,\sigma^2)$ σ^2 unknown
M10	$\mathcal{N}(\mu,\sigma^2)$	$(1-\pi)\mathcal{N}(\mu,\sigma_1^2) + \pi\mathcal{N}(\mu,\sigma_2^2)$ μ unknown

where

$$T_n(\mu) = \frac{1}{\sqrt{n(e^{\mu^2}-1)}} \sum_{i=1}^{n} \left[e^{(X_i\mu - \mu^2/2)} - 1 \right], \quad \mu \neq 0,$$

and

$$\lim_{\mu \to 0} T_n^2(\mu) = n\bar{X}^2, \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Our purpose in this section is to compute Monte-Carlo quantiles of this test statistic for realistic n (and up to "asymptotic" sample sizes), and evaluate the asymptotic behaviour w.r.t. the theoretical results. This allows us to evaluate the power of this LR-Test using these Monte-Carlo or asymptotic quantiles. We are also comparing it with an EM-test we derived using the methodology in Chen and Li (2009). Note that this particular model has not been handled by these authors.

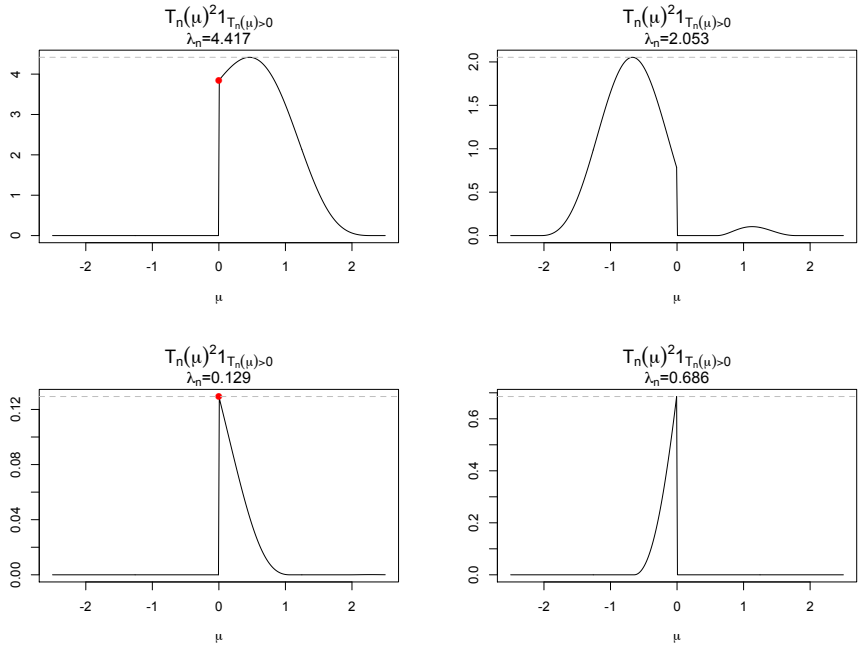


Figure 1: Some typical behaviours of $\mu \mapsto T_n^2(\mu)\mathbb{1}_{\{T_n(\mu)\geq 0\}}$ for $a = 2.5$ and simulated samples of size $n = 1000$ under H_0 for **M2**. The red dot is the limiting behaviour $T_n^2(0)$, when $T_n(0) \geq 0$.

2.2 Model M2 Monte-Carlo simulation for quantiles

2.2.1 Quantiles for the LR Test

For computing the statistic, we have to define a suitable compact $[-a, a]$. Following Garel (2001), we first tried values $a \in \{1, 2.5, 5\}$. The test statistic λ_n is easy to compute, the supremum over $\mu \in [-a, a] \setminus \{0\}$ being obtained by discretizing the interval in $k = 100$ or $k = 200$ steps. Figure 1 shows some typical behaviour of $\mu \mapsto T_n^2(\mu)\mathbb{1}_{\{T_n(\mu)\geq 0\}}$, for which we choose $a = 2.5$. It allows to see the global behaviour of the statistic by simulating several samples and retaining the different shapes obtained. In particular the discontinuity in 0 where the statistic jumps to its opposite value is visible.

Figure 2 shows the comparison between asymptotic, previously published quantiles, and Monte-Carlo quantiles computed from a large-scale experiment with $10,000$ replications, and several sample sizes from $n = 100$ up to $n = 100,000$. This experiment shows that the convergence is rather slow, but happened for $a = 1$, whereas the usage of the asymptotic quantile is questionable in cases where $a = 2.5$ or larger is used. This study hence suggests to use the Monte-Carlo quantiles in practice for any realistic (hence small) n. This very slow convergence is somehow in accordance with the rate in $\log(\log n)$ claimed by other authors as, eg, Bickel and Chernoff (1993). Note that using $a > 2.5$ is not realistic in practice for this model since, for contamination mean μ so distant from 0, the mixture structure becomes visible just looking at an histogram of the data. A more detailed explanation about the setting and impact of a is given in Section 6.1.

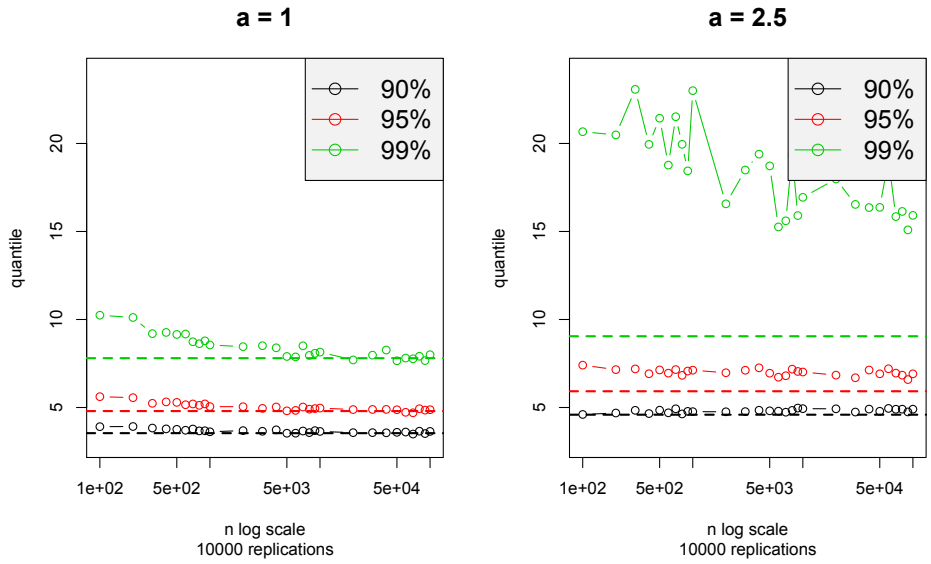


Figure 2: **M2**: Empirical quantiles for λ_n based on 10,000 Monte-Carlo replications vs. Asymptotic theoretical values from Garel (2013) (dotted lines) for three percentiles (90%, 95% et 99%) and two compact sets with $a = 1$ (left) and $a = 2.5$ (right). The x axis is in log scale, for $100 \leq n \leq 100,000$.

2.2.2 Quantiles for the EM-test

We present here the quantiles calculated from Monte-Carlo simulation for the statistic $EM_n^{(k)}$ of the EM-test proposed in Chen and Li (2009). **M2** corresponds to Example 2 in Li et al. (2009) but it seems that they do not provide the asymptotic distribution. We did not find a definition and implementation of the EM-test for this model in the package MixtureInf (Li et al., 2016) proposed by these authors, so we defined our implementation. These quantiles are computed from experiments with 10,000 replications of size n. Each experiment have been computed three times to evaluate the accuracy of the number of replications.

We choose $K = 3$ for the number of iterations in the EM algorithm and (0.1, 0.3, 0.5) for initial values for π as proposed in, e.g., Chen and Li (2009). The maximization of the initial step is done using the R function `optimize()` for this simple case. We use the penalization proposed in Chen and Li (2009), $p(\pi) = \log(1-|1-\pi|)$. Our results are in Table 2. From our Monte-Carlo experiment, a $\chi^2(1)$ limit distribution for the EM-test statistic $EM_n^{(K)}$ for **M2** seems valid, even though the convergence appears to be very slow.

2.3 Model M2 power evaluation

We have simulated under H_1 the mixture:

$$X \sim (1-\pi)\mathcal{N}(0,1) + \pi\mathcal{N}(\mu,1), \tag{5}$$

with parameters $\mu = 0.7$ and $\pi = 0.3$, already used in (Garel, 2001). These settings result in a severely overlapping mixture, as illustrated in Figure 3. Our motivation for choosing such a non-obvious mixture model is based on the idea that, if a simple histogram of the data already reveals

Table 2: Mean (and standard deviations) of three percentiles of the statistic $EM_n^{(K)}$ for different probabilities and different values of n.

n/α	90%	95%	99%
100	2.81 ($5.97\ 10^{-2}$)	3.97 ($1.77\ 10^{-2}$)	6.6 ($8.78\ 10^{-2}$)
200	2.75 ($5.8\ 10^{-2}$)	3.94 ($8.71\ 10^{-2}$)	6.89 ($22.8\ 10^{-2}$)
500	2.75 ($3.47\ 10^{-2}$)	3.93 ($3.98\ 10^{-2}$)	6.75 ($7.22\ 10^{-2}$)
1000	2.68 ($0.47\ 10^{-2}$)	3.83 ($1.90\ 10^{-2}$)	6.48 ($17.6\ 10^{-2}$)
5000	2.73 ($4.63\ 10^{-2}$)	3.90 ($10\ 10^{-2}$)	6.56 ($17.6\ 10^{-2}$)
10^4	2.71 ($0.54\ 10^{-2}$)	3.87 ($2.30\ 10^{-2}$)	6.65 ($10.3\ 10^{-2}$)
$\chi^2(1)$	2.71	3.84	6.63

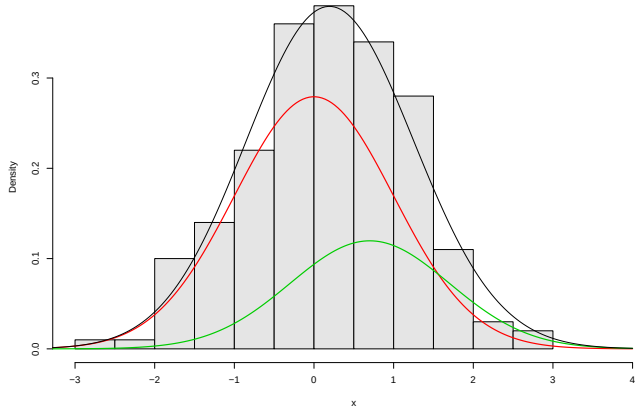


Figure 3: Typical empirical distribution of a $n = 200$ sample from the mixture model (5), with the true distributions for the two components (red, green) and the mixture (black).

a multi-modal distribution, then the test itself is not needed. Figure 4 shows that the estimated power of the LRT for model (5) is very good, even for small sample sizes, considering the difficulty of this severely overlapping mixture. However, one should be aware that this good behaviour is also a consequence of the simplicity of the model, with very few unknown parameters and a completely known distribution under H_0. Our implementation of the EM-test for **M2** shows a comparable behavior.

Since the null hypothesis for models in this paper is "homogeneous Gaussian population", one may suggest to use a standard normality tests. Intuitively, such a test with a general alternative hypothesis such as "non gaussian distribution", should be less powerful than a test dedicated to capture a mixture as the alternative. To confirm this, we have also compared the proposed tests for mixture against standard normality tests. The most common methods are Kolmogorov-Smirnov (KS) and Shapiro-Wilk (SW) tests. The standard KS test is for a null hypothesis restricted to a single, arbitrary but completely known density $H_0 : X \sim F^0$ (i.e. F^0 does not require any parameter estimation). The SW test is restricted to normality test, but for a null corresponding to the Gaussian family. In the case of **M2**, the null hypothesis is the single density $\mathcal{N}(0, 1)$, so that the KS test is more appropriate, i.e., comparison with the LRT proposed by Garel (2001) is fair, in the sense that both tests are considering the same single distribution

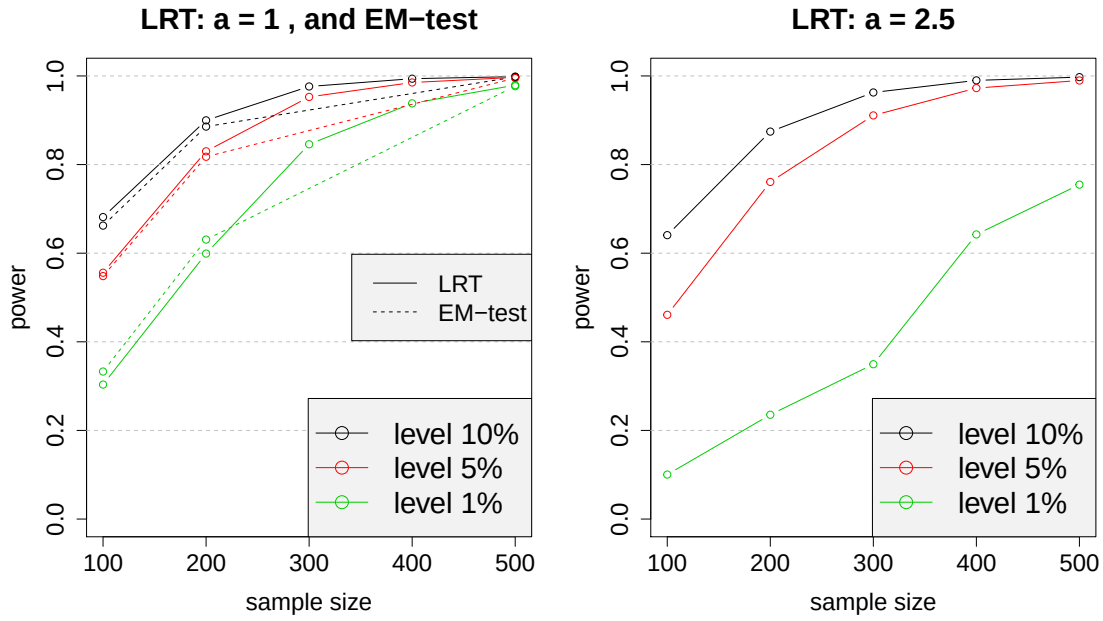


Figure 4: Power estimates for **M2**. LRT: using the quantiles obtained by Monte-Carlo as in Figure 2, for $a = 1$ (left) and $a = 2.5$ (right), $10,000$ replications. EM-test (left): for $n = 100, 200$ and 500, 2000 replications.

under the null hypothesis.

Results in Figure 5 show that the LRT achieves a better power than the KS test in this case. The conclusion holds as well for the EM-test when comparing with Figure 4. Indeed, this shows empirically that it is preferable in practice to use a test dedicated to capture an alternative hypothesis specifying a mixture, when this is expected to be the case. We have also ran the SW which achieve a very poor performance, with a power approximately equal to its level: the SW test, which tests a gaussian family, is not an option when H_0 is a single normal distribution.

3 Model M6

This model corresponds to a location mixture with same and known variance, set to 1 without loss of generality.

$$\textbf{M6} \quad H_0 : \mathcal{N}(\mu_0, 1) \quad \text{vs.} \quad H_1 : (1-\pi)\mathcal{N}(\mu_1, 1) + \pi\mathcal{N}(\mu_2, 1).$$

Garel (2001) Equation (11) page 335 proposed the following LRT statistic for this model:

$$\lambda_n = \sup_{\mu \in \Theta \setminus \{\mu_0\}} T_n^2(\mu)\, \mathbb{1}_{\{T_n(\mu) \geq 0\}} + o_p(1), \tag{6}$$

where $o_p(1)$ is a quantity which tends to 0 as n tends to infinity uniformly with respect to $\mu \in \overline{\Theta}$, where Θ is a bounded open interval and $\overline{\Theta}$ its closure. For $\mu \neq \mu_0$ we have:

$$T_n(\mu) = \frac{n^{-1/2}}{D(\mu)} \times \sum_{i=1}^{n} \left\{ \frac{\exp\left[-(1/2)(X_i - \mu)^2\right]}{\exp\left[-(1/2)(X_i - \mu_0)^2\right]} - 1 - (\mu - \mu_0)(X_i - \mu_0) \right\},$$

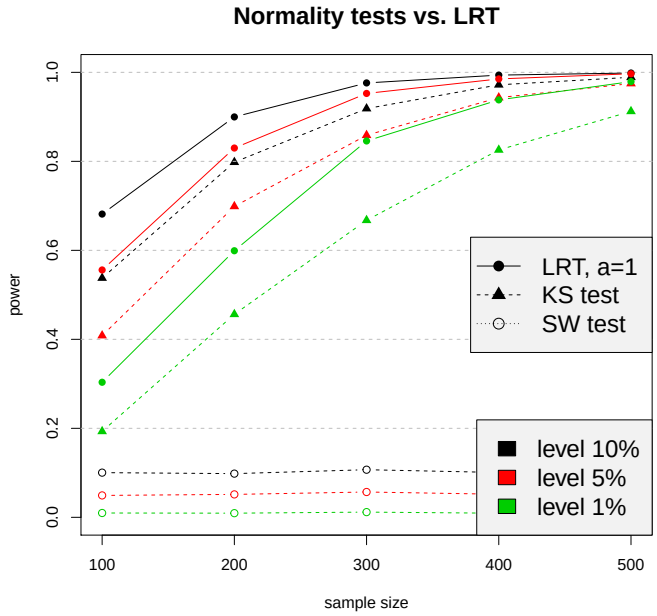


Figure 5: Power estimates for **M2**: LRT vs. the KS and SW normality tests, 10,000 replications.

where $D^2(\mu) = \exp\left[(\mu-\mu_0)^2\right] - 1 - (\mu-\mu_0)^2$, and μ_0 is the true value of μ under H_0. We have also:

$$\lim_{\mu \longrightarrow \mu_0} T_n(\mu) = (2n)^{-1/2} \sum_{i=1}^{n} \left[(X_i - \mu_0)^2 - 1\right].$$

This statistic is in its principle similar to the one detailed for **M2**, and is given up to a remainder term R_n which is $o_p(1)$ under H_0. Then, it is not surprising that this statistic includes the knowledge of the mean μ_0 under H_0. We denote the LRT statistic $\lambda_n(\boldsymbol{x}, \mu_0)$ here, where $\boldsymbol{x} = (x_1, \ldots, x_n)$ is the n-sample available to perform the test. The distribution of λ_n has been studied only asymptotically, i.e. when μ_0 is available. Of course, when one works with real data, this value is unknown and have to be estimated from the data. We propose in this work a methodology allowing for a practical application of this test in actual situations.

3.1 Model M6: LRT with plug-in estimate

Our approach consists in plugging-in an estimate of μ_0, i.e. computing some $\hat{\mu}_0$ and using it in $\lambda_n(\boldsymbol{x}, \hat{\mu}_0)$. There are several issues when using such a plug-in estimate, in particular (i) How to estimate μ_0, and under which constraint (H_0 or the general model); (ii) shall we separate the available sample in two subsamples, or use the whole sample twice?

The subsampling strategy consists in separating the available data in two samples, say $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, compute the estimate $\hat{\mu}_0(\boldsymbol{x}_1)$ using the first sample, and plug it to compute the LRT statistic $\lambda_n(\boldsymbol{x}_2, \hat{\mu}_0(\boldsymbol{x}_1))$. On the other way, using the same sample twice means that the data $\boldsymbol{x}$ are used first to get the estimate $\hat{\mu}_0(\boldsymbol{x})$, and then re-used to compute the statistic i.e. $\lambda_n(\boldsymbol{x}, \hat{\mu}_0(\boldsymbol{x}))$. This solution is more appealing since the sample size used in both procedure is n instead of $n/2$ (if 50% of the initial sample is selected for $\boldsymbol{x}_1$). But it implies a dependence between the plug-in estimate and the test statistic, that may have unpredictable effects. In the case of the LRT for

M6, the theoretical properties (asymptotic distribution of λ_n under H_0) is preserved since the weak consistency of the estimate $\hat{\mu}_0$ is the only required condition. We thus choose to use the sample $\boldsymbol{x}$ twice here, but we had also experiment the subsampling procedure for comparisons that are not presented here.

3.1.1 Estimation of μ_0

Since μ_0 is the expectation of the data under H_0, a straightforward procedure suggests to estimate it by the classical Maximum Likelihood Estimate (MLE) under H_0, that is the sample empirical mean $\hat{\mu}_0(\boldsymbol{x}) = \bar{x}_n$. This will work perfectly if the data come from homogeneity. However, if the data come from a mixture, this estimate will not converge to μ_0 and the power of the test collaps dramatically. This fact has been verified by a simulation experiment that shows a power approximately equal to the level of the test (the results are not provided here for brevity). Instead, we thus propose to estimate μ_0 by the mean associated to the largest component weight of the fitted mixture, i.e. the one representing the "prominent gaussian" closest to the model under H_0. This estimation needs consequently to be performed with a version of the EM algorithm for a two-component Gaussian mixture, with the constraint σ^2 known (and set to 1), imposed in **M6**. Hence the parameter of the model is $\boldsymbol{\theta} = (\pi, \mu_1, \mu_2)$. It is well known that the EM algorithm fits the parameter of a mixture up to a permutation of the labels (the "label-switching" problem). Thus we define the estimate

$$\hat{\mu}_0(\boldsymbol{x}) = \hat{\mu}_1 \mathbb{1}_{\{\hat{\pi}<1/2\}} + \hat{\mu}_2 \mathbb{1}_{\{\hat{\pi}\geq 1/2\}}.$$

Several constrained EM-algorithm have been proposed in Chauveau and Hunter (2013), and are already available in the mixtools package (Benaglia et al., 2009) for the R statistical software (R Core Team, 2016), including this very simple situation (known and same variance for the two components).

Initialization of EM algorithms A crucial remaining question is the EM initialization, which is known to influence the estimates. See, e.g., Bordes and Chauveau (2016), Section 5.1 for a discussion about several initialization methods depending on the complexity of the model. For a bimodal and univariate mixture the simplest procedure, called *splitting the data*, consists in breaking the data in two blocs (left and right) associated to the two components ideally using a visible separation between two modes from an histogram, and using as initialization the estimates of the component parameters from these blocs. An automatic, data-driven implementation when the histogram is not obviously bimodal (as in our case here) is as follows:

- Define a finite set $\mathcal{S}$ of k points within the observations range (typically a grid)
- for each $s \in \mathcal{S}$:
  1. Split the data $\boldsymbol{x}$ into two subsamples $\boldsymbol{x}^1 = \{x_i : x_i \leq s\}$ and $\boldsymbol{x}^2 = \{x_i : x_i > s\}$
  2. Compute the MLE from each subsample, the empirical mean μ_j^0 of sample j in the present normal case. Set $\pi^0 = \#\{x_i : x_i \leq s\}/n$, to obtain an initial parameter $\boldsymbol{\theta}^0 = (\pi^0, \mu_1^0, \mu_2^0) = \boldsymbol{\theta}^0(s)$
  3. Run the EM algorithm from that $\boldsymbol{\theta}^0(s)$ to get an estimate $\hat{\boldsymbol{\theta}}(s)$.
- retain the estimate achieving the largest log-likelihood: $\hat{\boldsymbol{\theta}} = \text{argmax}_{s\in\mathcal{S}}\, \ell_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}(s))$.

In all our experiments below, we choose a grid of $k = 10$ points between the empirical quantiles of order 10% and 90% of the data $\boldsymbol{x}$. This means that 10 EM algorithms are ran until convergence for each replication of the data.

3.1.2 Quantiles for the LRT with plug-in estimate

For the setting $a = 1$, i.e. the compact $[0, a]$ as in Garel (2013), the empirical quantiles for $\lambda_n(\boldsymbol{x}, \hat{\mu}_0(\boldsymbol{x}))$ are given in Figure 6 (Left). We have also computed and displayed in Figure 6 (Right) the quantiles for $\lambda_n(\boldsymbol{x}, \mu_0)$ with μ_0 known (for more replications since the code runs faster: there are no EM algorithms involved). In both cases, the empirical quantiles are slowly converging to the asymptotic ones from Garel (2013) (as in **M2**), but the effect of adding a plug-in estimate increases the difference even more. All this suggests that the empirical quantiles should be prefered in practice in any actual situation.

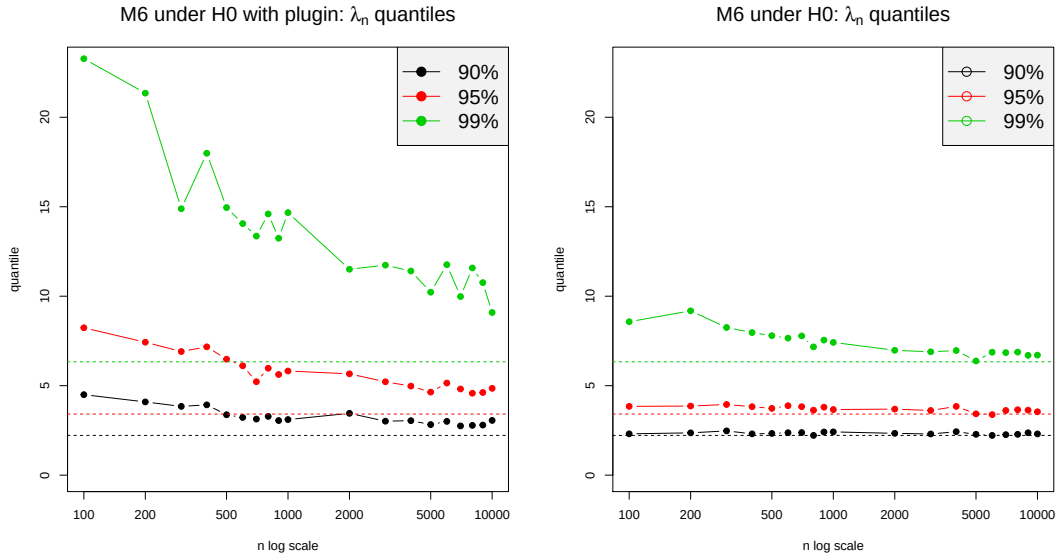


Figure 6: **M6** Empirical quantiles of the LRT statistic for three percentiles (90%, 95% et 99%), $a = 1$. Left: for $\lambda_n(\boldsymbol{x}, \hat{\mu}_0(\boldsymbol{x}))$, 2000 replications; Right: for $\lambda_n(\boldsymbol{x}, \mu_0)$, 10,000 replications. Dotted lines are asymptotic values from Garel (2013). The n axis is in log scale.

An EM-test for **M6** is provided in Li et al. (2009), since it is a special case of their general result for a scalar parameter θ, and a general density f satisfying regularity assumptions. Their theorem 2 says that the limiting distribution (in n) of the test statistic is $EM_n^{(K)} \to 0.5\,\delta_0 + 0.5\,\chi^2(1)$ for any K. This EM-test has been implemented in the package MixtureInf (Li et al., 2016), in the function `emtest.norm0` in the last available version we tried (1.0-1). We have checked numerically that the statistic actually converges to the claimed distribution for realistic sample sizes. However, we noticed that the discrete part $0.5\,\delta_0$ comes in their code from the fact that any negative value of the statistic are simply replaced by 0, which is a procedure not in accordance with the theory, and for which we have no explanations.

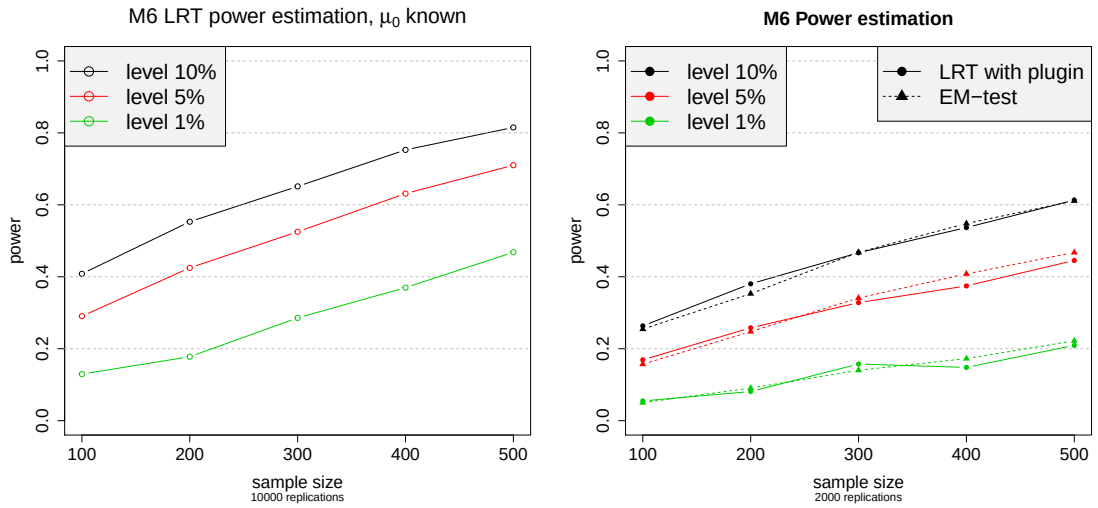


Figure 7: Monte-Carlo estimation of the power for **M6**, H_1 as in Equation (5). *Left:* LRT using asymptotic definition of the statistic, i.e. μ_0 known. *Right:* LRT with plug-in estimate based on the same sample, compared with the EM-test.

3.2 Model M6 power evaluation

We have evaluated the power of the EM-test and the LRT for **M6**, on the same model under H_1 as for **M2**, Equation (5), and same sample sizes and test levels. For the EM-test we have used the function `emtest.norm0()` from the last available update of the R package MixtureInf. We have then compared these results with the two LRT options, the asymptotic strategy i.e. with μ_0 known and statistic $\lambda_n(\boldsymbol{x}, \mu_0)$, and the plugin strategy with the statistic $\lambda_n(\boldsymbol{x}, \hat{\mu}_0(\boldsymbol{x}))$ detailed above. In each case we used the quantiles obtained by our Monte-Carlo experiment for the corresponding strategy. The results in Figure 7 (Right) show that our strategy for building a plug-in estimate of μ_0 in the LRT statistic preserves most of the power of the test: the LRT using the n-sample twice, statistic $\lambda_n(\boldsymbol{x}, \hat{\mu}_0(\boldsymbol{x}))$, is very comparable to the EM-test. Figure 7 (Left) shows that the power of the LRT when μ_0 is known, is superior to that of the EM-test.

Our advice is then that the LRT is a better option if the mean under homogeneity (H_0) is known from previous experiments, expert prior information, or as a standard from some area of expertise. If μ_0 has to be estimated from the data, then LRT and EM-test are comparable. In all cases, the LRT should be applied using the corresponding Monte-Carlo quantiles. Note also that the comparison with model **M2**, Figure 4, is not meaningful since **M2** is an even simpler model with all parameters known under H_0.

4 Model M8

This model is actually the general one, in particular more general than **M9** that will be considered later, since all the parameters are unknown and unconstrained under both H_0 and H_1.

$$H_0 : X \sim \mathcal{N}(\mu, \sigma^2) \quad \text{vs.} \quad H_1 : X \sim \lambda\mathcal{N}(\mu_1, \sigma_1^2) + (1-\lambda)\mathcal{N}(\mu_2, \sigma_2^2),$$

i.e. the test for H_0 : "Gaussian distribution" vs. H_1 : "Two-component Gaussian mixture". For this model, the LRT-based strategy only proposed a test statistic as a conjecture (Garel, 2001).

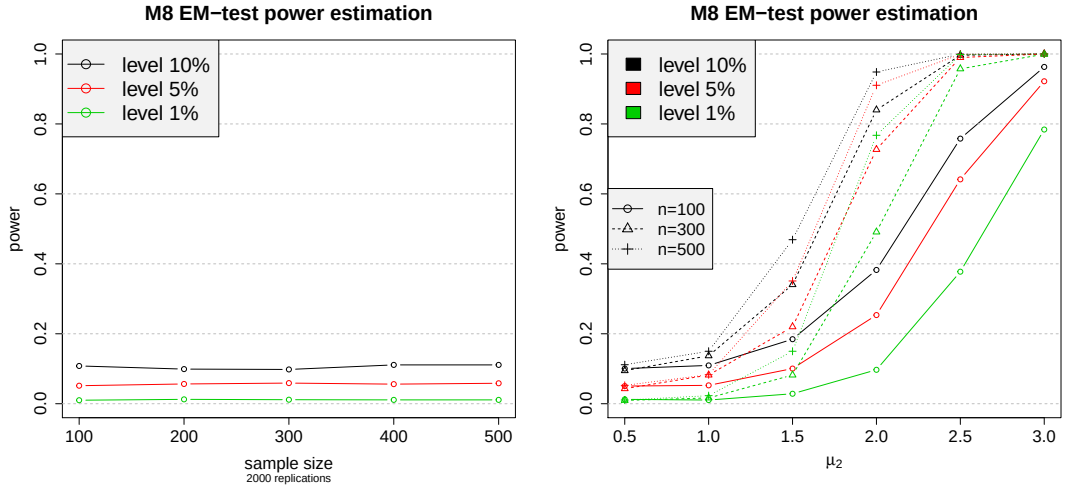


Figure 8: Monte-Carlo estimation of the power for **M8** based on 2000 replications, using the MixtureInf package: *Left*, parameters as in (5) with $\mu_2 = 0.7$, *Right*: power as a function of μ_2, the mean of the second component.

Chen and Li (2009) Section 3 handles this model with the EM-test, and claims that the limiting distribution is $\chi^2(2)$. The code for the EM-test in **M8** is provided by the function `emtest.norm` in the last available update of the R package MixtureInf (Li et al., 2016). This code also handles these authors recent extensions to higher order tests, namely a mixture with m_0 components for the null hypothesis versus a mixture with $m > m_0$ components, see Chen et al. (2012).

We have experimented this EM-test strategy under H_0 first. The announced limiting distribution of the M_n statistic as a $\chi^2(2)$ is partially verified in practice, in the sense that, under $H_0 : \mathcal{N}(0, 1)$, we always observed a small percentage (between 3% and 5%) of negative values. If we remove these negative values, then the empirical distribution is reasonably fitted by a $\chi^2(2)$.

We have also evaluated the power of the EM-test for **M8** using a Monte-Carlo experiment as before, for the same mixture model as in **M2** and **M6**, i.e. with parameters as in Equation (5). The results are rather disappointing, in comparison with, e.g., **M6**, even if the statistical problem is obviously more difficult here. Figure 8 (Left) shows the empirical power using the same settings as before. It is clear that a model with means 0 and 0.7 is too difficult for this test to capture the (severely overlapping) mixture structure. Actually, the EM-test power here is approximately equal to its level. We have thus conducted a second experiment, where the power is estimated when the mean of the second component increases, i.e. the mixture becomes more and more easy to detect, for the same range of sample sizes. Results are displayed in Figure 8 (Right).

Our advice to users for this general model is that a good power can only be obtained for "separated enough" models. Figure 9 illustrates for instance the type of true and empirical distributions one can expect, to achieve a power $\geq 90\%$ with a sample of size $n = 300$. It shows in particular that in this case the empirical distribution can be at least skewed, if not bimodal, so that H_0 is often not reasonable.

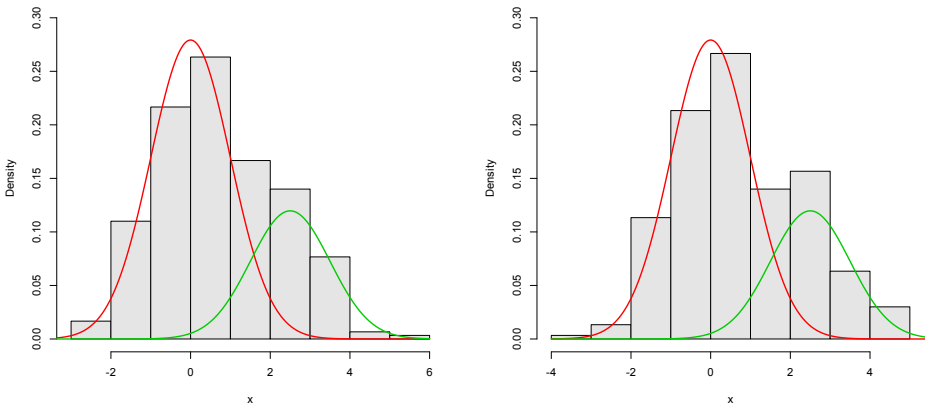


Figure 9: Mixture **M8** under H_1 for $\mu_2 = 2.5$, and two examples for samples of size $n = 300$.

5 Model M9

This model is as before a mixture on the mean, with an unknown but equal variance, sometimes called the structural parameter. Thus **M9** is more general than **M6**, but less than **M8**:

$$H_0 : \mathcal{N}(\mu_0, \sigma^2) \quad \text{vs.} \quad H_1 : (1-\pi)\mathcal{N}(\mu_1, \sigma^2) + \pi\mathcal{N}(\mu_2, \sigma^2). \tag{7}$$

The LRT statistic for this model has been proved under a separation condition and conjectured, without this separation condition, by Garel (2001) section 3.3. The statistic λ_n is expressed as in Equation (6), where here for $\mu \neq \mu_0$ we have:

$$\begin{aligned} T_n(\mu) &= \frac{n^{-1/2}}{D(\mu)} \times \\ &\sum_{i=1}^{n} \left\{ \exp\left[\frac{1}{2} d_0^2(X_i) - \frac{1}{2}\left(\frac{X_i - \mu}{\sigma_0} \right)^2 \right] - 1 - d_0(\mu)\, d_0(X_i) - \frac{1}{2} d_0^2(\mu) \left[d_0^2(X_i) - 1 \right] \right\}, \end{aligned}$$

$$\begin{aligned} D^2(\mu) &= \exp\left[d_0^2(\mu) \right] - 1 - d_0^2(\mu) - \frac{1}{2} d_0^4(\mu), \\ d_0(u) &= \frac{u - \mu_0}{\sigma_0}, \end{aligned}$$

where μ_0 and σ_0 are respectively the true values of μ and σ under H_0. We have also:

$$\lim_{\mu \longrightarrow \mu_0} T_n^2(\mu) = (6n)^{-1} \left\{ \sum_{i=1}^{n} \left[d_0^3(X_i) - 3 d_0(X_i) \right] \right\}^2.$$

LRT statistics are given up to a remainder term R_n which is $o_p(1)$ under H_0. This statistic includes the knowledge of the mean μ_0 under H_0 (as for **M6**) but also of the structural parameter σ^2. Of course, when one works with real data, these values are unknown and have to be estimated from the data. Again, as for **M6**, applying here the MLE's under H_0 results in a test with very low power.

5.1 Model M9: LRT with plug-in estimate

We thus develop here also a LRT with plugged-in estimates of (μ_0, σ^2), in the spirit of what we did for **M6**. We use the whole sample both for estimation and computation of the LRT statistic, that is consequently denoted $\lambda_n(\boldsymbol{x}, \hat{\mu}_0(\boldsymbol{x}), \hat{\sigma}^2(\boldsymbol{x}))$. Estimation of (μ_0, σ^2) in this model requires a different version of the EM algorithm, precisely for a two-component Gaussian mixture and equal but unknown variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$, so that the model parameter is now $\boldsymbol{\theta} = (\pi, \mu_1, \mu_2, \sigma) \in \boldsymbol{\Theta}$. This specific EM is also available in the mixtools package (Benaglia et al., 2009). As for **M6**, μ_0 is estimated by the mean associated to the largest component weight of the fitted mixture. For initialization and optimization of the EM algorithm, we use a more general scheme than for **M6**, since the model is more complicated (four parameters instead of three, and more flexibility due to the unknown variance). This initialization, called *exhaustive exploration* of the parameter's space $\boldsymbol{\Theta}$ in Bordes and Chauveau (2016), consists in starting several (say k) EM algorithms from k values of $\boldsymbol{\theta}^0$'s uniformly drawn from a compact inside $\boldsymbol{\Theta}$, and retaining as previously the estimate with the largest log-likelihood. Precisely, the EM algorithms were started with initial $\pi^0 \sim \mathcal{U}_{[0.05,0.5]}$ and μ_j^0's draws uniformly over $[q_{\alpha/2}(\boldsymbol{x}), q_{1-\alpha/2}(\boldsymbol{x})]$, where $q_\alpha(\boldsymbol{x})$ is the empirical quantile of order α obtained from the data, with $\alpha = 0.2$.

We obtained empirical quantiles of this LRT, by Monte-Carlo simulation under H_0, for the simplest situation, i.e. (μ_0, σ^2) known, and for the plug-in approach. We do not detail the results here for brevity, and since these are behaving as for **M6**: the empirical quantiles are close to the asymptotic ((μ_0, σ^2) known and $n = \infty$) values from Garel (2013), and are much larger when using the plug-in approach. As a consequence, it is important in practice to use the empirical quantiles corresponding to the actual situation (and a value corresponding to the plausible range for μ_2).

We have then estimated the power of the LRT in the two settings: (μ_0, σ^2) known, or using the plug-in approach, and for the same alternatives as for **M8** (see Figure 8, Right), i.e. for several values of μ_2 given easier mixtures. Results are displayed in Figure 10 and discussed below.

5.2 Model M9: EM-test

This model is also studied in Chen and Li (2009), Section 2, where the EM-test limiting distribution is given by their Theorem 2: for any fixed number of iterations K of the EM algorithm using a penalized log-likelihood,

$$\mathbb{P}(EM_n^{(K)} \leq x) \to F(x - \Delta)[0.5 + 0.5F(x)] \quad \text{as } n \to \infty,$$

where $F(\cdot)$ is the cdf of the $\chi^2(1)$ distribution, and Δ is a negative but fixed constant that depends only on the penalty $p(\pi)$ and the π_j's chosen in the initialization procedure (Chen and Li, 2009). This distribution has its support in (Δ, ∞) and

$$\mathbb{P}(EM_n^{(K)} \leq 0) \to 0.5F(-\Delta) \quad \text{for } \Delta < 0.$$

In particular with the settings from Chen and Li (2009),

$$p(\pi) = \log(1 - |1 - 2\pi|), \quad \Delta = 2 \max_{\pi \in \{0.1, \ldots, 0.4\}} (p(\pi) - p(0.5)) \approx -0.446,$$

so that it gives $\mathbb{P}(EM_n^{(K)} \leq 0) \approx 0.248$.

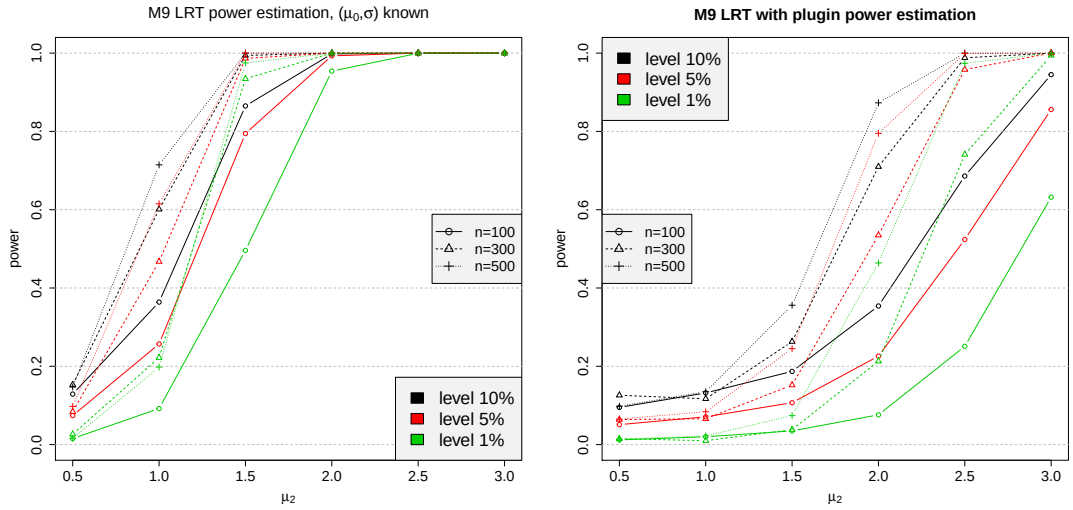


Figure 10: Monte-Carlo estimation of the power of the LRT for **M9**, $\lambda_n(\boldsymbol{x}, \mu_0, \sigma^2)$, i.e. when (μ_0, σ^2) are known (Left), and of $\lambda_n(\boldsymbol{x}, \hat{\mu}_0(\boldsymbol{x}), \hat{\sigma}^2(\boldsymbol{x}))$, i.e. LRT with plug-in (Right), as a function of the mean μ_2 of the second component; $a = 3$, 2000 replications.

We did not find any implementation of the EM-test in this case in the MixtureInf package, even though the algorithm and the associated penalty functions (for $p(\pi)$ and $p(\sigma)$) are fully described in Chen and Li (2009) section 2. Hence we develop our implementation of this EM-test to compare it with **M8**. We were unable to clearly validate the asymptotic distribution under H_0, essentially because we observed a higher estimated value for $\mathbb{P}(EM_n^{(K)} \leq 0)$, even for large samples up to $n = 2000$. Our estimated type I error (Table 3) also did not exactly match the simulations provided by Chen and Li (2009) Table 1. We finally also apply the EM-test for **M9** with the same settings already used for **M8**, as in Figure 8 (Right). Results are in Figure 11 and show a slightly better power, which is expected since the model is simpler, but again good power is associated to somehow "easy to detect" mixtures.

All these results show that (as for **M6**), if (μ_0, σ^2) are available from prior experiments or external (expert) information, then the LRT should be preferred, using the corresponding empirical quantiles for the decision. If these parameters are unknown, then the EM-test achieves a slightly better power and shoud be preferred.

Table 3: Estimated type I error (%) for **M9**, $n = 200$, $K = 3$ and $\pi \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$; 20,000 replications.

Level	10%	5%	1%
Estimates	7.81	4.16	0.99

A note about CPU time and computing efficiency: our code of the EM-test for **M9** uses C calls for computing the E-step, borrowed from the mixtools package (Benaglia et al., 2009). We noticed that for the Monte-Carlo simulations used in power estimations showed in Figures 8 (Right) and 11, our code is approximately 30 times faster than the EM-test for **M8** provided by the MixtureInf package.

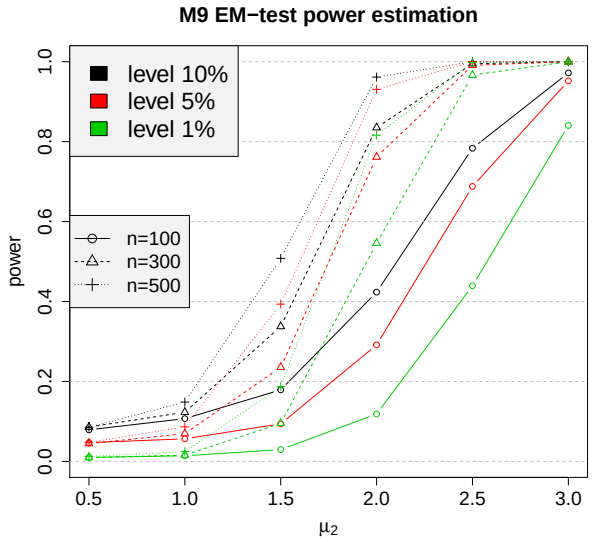


Figure 11: Monte-Carlo estimation of the power of the EM-test for **M9** based on our implementation and 2000 replications, as a function of the mean μ_2 of the second component.

6 An application to real data

To illustrate the application of some of the previous models, quantiles and power to actual data of moderate sample sizes, we have applied it to data collected for a research project from the French National Institute for Agricultural Research (INRA)[3]. Briefly, these quantitative data correspond to the number of days between two events concerning ovarian response and lambing for ewes of several kinds (races), coming from actual farms or experimental situations, and years. The purpose is the study of the so-called "ram effect" (or male effect), see, e.g., Fabre-Nys et al. (2016). A two-component mixture is sometimes suspected for the distributions of these datasets, and the empirical distributions are not always giving evidence of it. Hence the test for homogeneity comes as a natural technique, and the conclusions of the tests have important consequences for the biological application.

We focus here on a dataset of $n = 660$ observations for a selected race (*Romane*), location (*Sapinière*) and year (2009). Figure 12 shows the empirical distribution of the data, which does not look very well bell-shaped, but is not obviously bimodal in the sense expected from the biological context (a mode around 150 days and another mode about 6 days later). We have first added to this plot a single normal fit (i.e., assuming homogeneity), and a two-component Gaussian mixture fit using the plain `normalmixEM()` function of the Benaglia et al. (2009) mixtools package for the R statistical software R Core Team (2016). From the empirical distribution of the data and the biological context, equal variance among the components can reasonably be assumed. Since this variance is unknown in this real data case without expert information, **M9** can be used. The general model **M8** can also be used if we do not assume variance equality. In addition, we also tried to use **M6** with a prior estimation of this common variance, to illustrate and compare the various available procedures, even though this is not theoretically correct.

[3] *Projet de Recherche d'Intérêt Régional* DURAREP2.

6.1 Choosing the parameter a in practice

As mentionned in the previous sections, computing the LRT statistic requires the selection of a parameter a, since a supremum over the mean μ of the second component in a suitable compact set $[\mu_0; \mu_0 + a]$ or $[\mu_0 - a; \mu_0 + a]$ around the null or first component mean μ_0 is involved in the computation of λ_n in all the models we studied. As seen before, if the interval to which the parameter is assumed to belong is unbounded, the likelihood ratio test statistic converges in probability to $+\infty$ as n tends to $+\infty$. This phenomenon has been highlighted, for the first time, by Hartigan (1985). For a contaminated model on the mean, the value of a determines the interval where the mean of the second component is likely to be. Our experience indicates that this value has little effect on the result of the test, i.e. it will generally not modify the detection of a mixture.

Of course, from a methodological point of view, it is more satisfying to choose a large enough, in a "reasonable" range deducted from the empirical distribution (histogram) of the data and the selected model, to compute the observed value of the statistic $\lambda_n(\boldsymbol{x})$ for a given dataset $\boldsymbol{x}$. However, a parameter a is involved in the computation of Monte-Carlo quantiles of λ_n under H_0 as well, and this could be an issue if one would have to compute these quantiles specifically for each new experiment, data set and a value. Fortunately, in the Gaussian case, the LRT statistic is invariant by a linear transformation of the data, so that ultimately what is needed is an approximation of quantiles under a null model with (e.g.) $\mu_0 = 0$, $\sigma_0 = 1$, and some values of a such that $a \geq |\mu_2 - \mu_1|/\sigma$ where μ_2 is the larger likely value obtained from an examination of the data. This is why we generally need only values such as $a = 1$, $a = 2.5$ or $a = 5$ in view of the null and normalized model. We illustrate that below for this dataset and **M6** and **M9**.

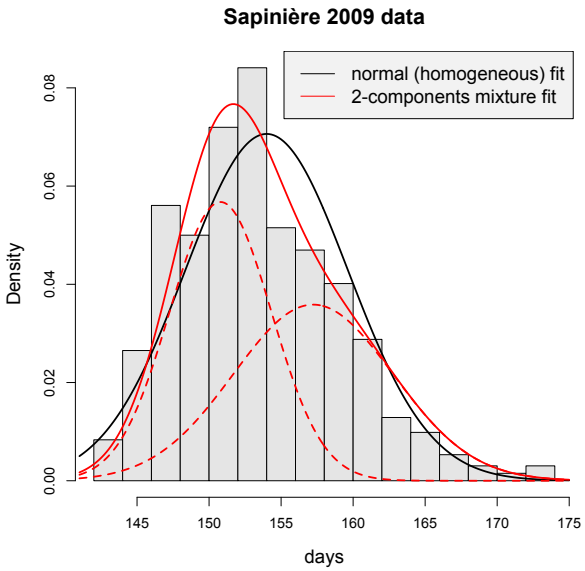


Figure 12: Homogeneous and 2-components mixture fits for the data *Romane Sapinière 2009*.

6.2 Results and comparisons

Using M6. The prior estimation of σ required for applying **M6** with the present dataset is done under the general model using an EM algorithm with the constraint $\sigma_1^2 = \sigma_2^2$ (actually similarly to what is done in **M9** for the plug-in approach), resulting in the estimate $\hat{\sigma} = 4.43$. The data are then normalized using this $\hat{\sigma}$. Note that normalizing the data using simply its empirical standard deviation (over-estimated value of 5.65) is not appropriate: clearly, normalizing with $\hat{\sigma} = 4.43$ better retains the features of the distribution depicted in Figure 12. Then the LRT for **M6** with plug-in strategy for μ_0 (Section 3.1) and providing $\hat{\sigma}$ as "known" is applied on the normalized data $\boldsymbol{y} = \boldsymbol{x}/\hat{\sigma}$, with a supremum over a compact $[\mu_0 - a; \mu_0 + a]$ with $a = 5$, in view of the range of the normalized data and the discussion above. We obtain an observed LRT statistic $\lambda_n(\boldsymbol{y}, \hat{\mu}_0(\boldsymbol{y})) = 3248$. We then use the Monte-Carlo quantiles with the same setting ($n = 660$ and $a = 5$) which leads to clear rejection of H_0 at level 1% (the 99% quantile for λ_n in this case being approximately 26). The approximate p-value has been computed from the same Monte-Carlo experiment for a large number of replication, and gives $p \approx 10^{-4}$. The EM-test for **M6** with the same normalization (using $\boldsymbol{y}$) returns a p-value $\approx 10^{-28}$, leading to clear rejection as well. Note that the code provided by the authors of the EM-test returns a p-value based on the asymptotic χ^2 distribution without control of the validity of this asymptotic, whereas our approach uses a p-value based on the actual non asymptotic distribution of the LRT statistic. This may explain the difference.

Using M9. The LRT for this model can be applied straightforwardly to these data, since the two unknown parameters are handled by the plug-in approach, as explained in Section 5.1. We also do not have to normalize the data. Looking at the possible modes in Figure 12 we choose $a = 10$. We then found $\lambda_n(\boldsymbol{x}, \hat{\mu}_0(\boldsymbol{x}), \hat{\sigma}^2(\boldsymbol{x})) = 3166$, and the plugged-in estimates $\hat{\mu}_0(\boldsymbol{x}) = 152.4$ and $\hat{\sigma}(\boldsymbol{x}) = 4.428$, consistent with the data. We obtain 2470 for the 99% quantile of $\lambda_n(\boldsymbol{x}, \hat{\mu}_0(\boldsymbol{x}), \hat{\sigma}(\boldsymbol{x}))$ under H_0 for $n = 660$ and $a = 3$ (since $10/\hat{\sigma} < 3$, using the invariance discussed in 6.1). This leads again to clear rejection. The estimated p-value is $p \approx 0.0064$.

Using M8. Finally, **M8** can also be used straightforwardly in this case, but only (among the various approaches we have detailed) using the EM-test approach. The function `emtest.norm2` of the MixtureInf package returns a p-value of 5.10^{-9} indicating clear rejection, in accordance with the conclusions using the other models and tests. Again this code returns a p-value based on the asymptotic χ^2 distribution. A Monte-Carlo quantile could give a slightly different result.

7 Discussion

7.1 A brief summary of our results and observations

M2 Results on the LRT statistic distribution is available (Garel, 2001). We have compared the non asymptotic quantiles with the asymptotic ones, and showed the good power of the test, even for a severely overlapping and non obvious mixture. To our knowledge, there is no EM-test available in the literature for this case. Our derivation of the EM-test and its implementation indicate a limiting distribution $EM_n^{(K)} \to \chi^2(1)$ under the null hypothesis, and a power similar to the results from the LRT test. Note that the convergence to the $\chi^2(1)$ distribution seems very slow.

M6 The LRT approach for **M6** gives asymptotic results on the statistic distribution, but the expression of the test statistic includes the value of the true μ_0 under H_0. Quantiles and power are evaluated in our work, in this set-up. We have also proposed a novel approach allowing this LRT to be used in practice, by estimating μ_0 under the general model with a constrained EM algorithm, and plugging-in this estimate in the test statistic. Note that this strategy using the same sample twice preserves the theoretical properties, but decreases the rate of convergence of the empirical quantiles. For the EM-test approach, Li et al. (2009) Theorem 2, claim the limit distribution $EM_n^{(K)} \to 0.5\,\delta_0 + 0.5\,\chi^2(1)$. A code is available in the MixtureInf package. We however noticed that values corresponding to the discrete part δ_0 are forced in the code, where negative values for $EM_n^{(K)}$ are checked and simply replaces by 0. Note also that the p-value of this EM-test is computed from its asymptotic distribution. The limit distribution under H_0 is recovered from our Monte-Carlo investigation, except the convergence to the weight 0.5 even for large n. We observe that the power of the EM-test is weaker than the LRT when μ_0 is known. Our advice is then that the LRT is a better option if the mean under homogeneity (H_0) is known from external sources. If μ_0 is unknown, then the LRT with plug-in estimation achieves a power comparable to that of the EM-test. The LRT should be applied using the corresponding Monte Carlo quantiles.

M8 In that model, a conjecture is proposed for the LRT statistic distribution but it has not been investigated. In Chen and Li (2009) Section 3, the limiting distribution $EM_n^{(K)} \to \chi^2(2)$ is given for the EM-test statistic. Code for this case is available in the MixtureInf package. Following our investigations, the message we deliver to potential users is that the power can be rather weak for datasets from non-obvious mixture, i.e. non obviously bimodal (or multi-modal) empirical distributions. **M8** clearly deserves more study on actual datasets.

M9 Results on the LRT are available but as for M6, only asymptotically in the sense that the computation of the test statistic requires the knowledge of both the mean μ_0 under H_0 and the structural parameter σ^2. As for **M6**, we have proposed here a plug-in approach, with an estimation under the general (more complex) model and another type of constrained EM. These two LRT have been compared. The EM-test for **M9** has been studied in Chen and Li (2009) section 2 and their Theorem 2 gives an asymptotic behaviour of the cdf of the statistic test $EM_n^{(K)}$. One difficulty is that there is no code available online (in the MixtureInf package or another). We thus develop our own code considering the algorithm proposed in Chen and Li (2009). We did not recover the asymptotic distribution under H_0 claimed by the author; in particular, we noticed a higher estimated value for $P(EM_n^{(K)} \leq 0)$ and slightly different estimated type I errors. Using a similar setting used for M8, we noticed that good power is reached for obvious mixtures. Our conclusion is that for **M9**, if the parameters (μ_0, σ^2) are known then the LRT achieves a better power than the EM test. In the general case, our implementation of the EM-test achieves a slightly better power than the LRT with plug-in estimation. Here again, the LRT shoud be applied using the Monte Carlo quantiles.

A note about higher order models In view of our results for the power of the EM-test in the case of a homogeneous null model, and since this model has been extended in Chen et al. (2012) and in the MixtureInf package (Li et al., 2016), to higher order models i.e. for $H_0 : m_0$-component mixture vs. $H_1 : m > m_0$-component mixture, we tried to re-run some of the experiments provided in Chen et al. (2012), Section 4 (simulation study). We reproduced the

Table 4: Powers of EM test at the 5% level for two higher order models and four alternatives for each, compared with results from Chen et al. (2012) showed in (·) for the EM-test with 3 iterations; estimates based on 5000 replications.

	H_1	$n = 200$	$n = 400$
$m_0 = 2$ vs. $m = 4$ (Table 4)	1	16.2 (20.0)	43.9 (44.1)
	2	33.2 (33.5)	67.6 (70.2)
	3	99.1 (40.5)	100 (60.2)
	4	100 (100)	100 (100)
$m_0 = 3$ vs. $m = 5$ (Table 6)	1	13.0 (10.4)	25.9 (28.4)
	2	41.8 (40.7)	79.7 (84.6)
	3	42.7 (44.8)	78.6 (83)
	4	80.1 (82.3)	99.2 (99.5)

experiments given by these authors in their Table 4, bottom panel, corresponding to $H_0 : m_0 = 2$ vs. four instances of $H_1 : m = 4$, and their Table 6, bottom panel, corresponding to $H_0 : m_0 = 3$ vs. four instances of $H_1 : m = 5$. Results are summarized in our Table 4. We estimate the powers based on 5000 replications (instead of 1000 in Chen et al. (2012)), to achieve more precise estimates. The results are often comparable, except for $m_0 = 2$ vs. $m = 4$, case 3.

7.2 Conclusion

This paper brought a new insight towards the problem of testing a two-component Gaussian mixture vs. the null hypothesis of homogeneity (no mixture). We compare two approaches, one based on the Likelihood-Ratio Test, and the other on the EM-test; guidelines for practionners are summarized in section 7.1, and an illustration of both approaches using different models is proposed. Classical results obtained in the frame of Likelihood Ratio Test (Ghosh and Sen (1985), Garel (2001)) rely on the true value of some parameters under H_0. But this value is unknown and may be difficult to obtain in a general framework. That is why we proposed new plug-in methods for two models we consider, in order to use the corresponding statistics with real data.

One interesting question is related to the proximity of the mixture model to homogeneity; but it is a tough problem. One aspect of this question can be described by: for a fixed π, the mean μ_2 tends to μ_1. This case has been thorougly addressed by the *Removing separation conditions in mixture problems* papers. The LRT statistic admits a limit and the completed statistic can be used for detection.

A second aspect can be described by $\pi \to 0$. Also allowing $n \to +\infty$, Donoho and Jin (2004) found the separation curve between the two hypotheses which is in $\log \log n$. At the practitioner level, the problem consisting in working on samples of small or moderate size, i.e. one hundred to one thousand of individuals, when π is small, has been described as "presence of outlying observations". Specific tests of hypotheses have been developped in this context. Basically, practitionners use to assume $\pi \leq 0.05$. Then a basic question remains : when is it possible or profitable to make a distinction between homogeneity and two-component mixture ? The answer is not unique. For instance, if the goal is to estimate parameters, such problems can be addressed by means of robust techniques. If the problem is to make evidence of two populations inside the

sample, very sensitive tests are needed.

A surprising by-product of our research is the inaccuracy of our simulation results with respect to the EM Test results in the case of **M9**. We have no clear explanation so that further investigations would be needed. Finally, we stress that this work allowed us to add a few contributions as codes implementing the LRT approach, that will be included in a futur update of the mixtools package (Benaglia et al., 2009) for the R statistical software (R Core Team, 2016).

Acknowledgements We gratefully acknowledge the Editor-in-Chief and two Reviewers for their valuable comments helping us improving the original manuscript.

References

Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. (2009). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29.

Bertillon (1874). *Démographie figurée de la France*. Masson, Paris.

Bertillon, L. (1876). *Moyenne. Dictionnaire encyclopédique des sciences médicales*, pages 296–324. Masson, Paris.

Bhattacharya, C. (1967). A simple method for resolution of a distribution into its gaussian components. *Biometrics*, 23:115–135.

Bickel, P. and Chernoff, H. (1993). *Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem*, pages 83–96. Ghosh, J.K. et al. (Eds), Wiley Eastern Limited.

Böhning, D. (2000). *Computer-assisted analysis of mixtures and applications*. Monographs on Statistics and Applied Probability 81. Chapman & Hall.

Bordes, L. and Chauveau, D. (2016). Stochastic EM algorithms for parametric and semiparametric mixture models for right-censored lifetime data. *Computational Statistics*, 31(4):1513–1538.

Chauveau, D. and Hunter, D. R. (2013). ECM and MM algorithm for mixtures with constrained parameters. Technical Report hal-00625285, version 2, HAL.

Chen, H. and Chen, J. (2001). Large sample distribution of the likelihood ratio test for normal mixtures. *Statistics and Probability Letters*, 52:125–133.

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

Chen, H., Chen, J., and Kalbfleisch, D. (2004). Testing for a finite mixture model with two components. *Journal Royal Statistical Society*, 66:95–115.

Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: the EM approach. *The Annals of Statistics*, 37:2523–2542.

Chen, J. and Li, P. (2011). Tuning the EM-test for finite mixture models. *The Canadian Journal of Statistics*, 39(3):389–404.

Chen, J., Li, P., and Fu, Y. (2012). Inference on the order of a normal mixture. *Journal of the American Statistical Association*, 107:1096–1115.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics*, 25:573–578.

Dacunha-Castelle, D. and Gassiat, E. (1997). Testing in locally conic models and application to mixture models. *Esaim Prob. Statistics*, 1:285–317.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, pages 662–994.

Everitt, B. and Hand, D. (1981). *Finite mixture distributions.* Chapman and Hall, London.

Fabre-Nys, C., Chanvallon, A., Dupont, J., Lardic, L., Lomet, D., Martinet, S., and Scaramuzzi, R. J. (2016). The "ram effect": A "non-classical"mechanism for inducing lh surges in sheep. *PLoS ONE*, 11(7):e0158530.

Feng, Z. and McCulloch, C. (1996). Using bootstrap likelihood ratio in finite mixture models. *Journal of the Royal Statistical Society B*, pages 609–617.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models.* Springer-Verlag, New-York.

Garel, B. (2001). Likelihood ratio test for univariate gaussian mixture. *Journal of Statistical Planning and Inference*, 96:325–350.

Garel, B. (2005a). Asymptotic theory of the likelihood ratio test for the identification of a mixture. *Journal of Statistical Planning and Inference*, 131:272–296.

Garel, B. (2005b). Percentiles of the supremum of a nonstationary gaussian process. In Ermakov, S., Melas, V., and Pepelyshev, A., editors, *Proceedings of the Fifth Workshop on Simulation*, pages 267–272. Springer.

Garel, B. (2013). *Modèles de mélanges : le nombre de composants*, chapter 3, pages 57–84. Technip.

Garel, B. and Goussanou, F. (2002). Removing separation conditions in a 1 against 3-components gaussian mixture problem. In Jajuga, K., Sokolowski, A., and Bock, H.-H., editors, *Classification, Clustering and Data Analysis*, pages 61–73. Springer.

Ghosh, J. and Sen, P. (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In LeCam, L. and Olshen, R., editors, *Proc. Berkeley Conf. in honor of Jerzy Neyman and Jack Kiefer*, pages 789–806, Monterey, Wadsworth.

Hall, P. and Stewart, M. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Annals of Statistics*, 13:795–800.

Hartigan, J. (1985). A failure of likelihood asymptotics for normal mixtures. In LeCam, L. and Olshen, R., editors, *Proc. Berkeley Conf. in honor of Jerzy Neyman and Jack Kiefer*, pages 807–810, Monterey, Wadsworth.

Lemdani, M. and Pons, O. (1999). Likelihood ratio tests in contamination models. *Bernoulli*, 5:705–719.

Li, P., Chen, J., and Marriot, P. (2009). Non-finite Fisher information and homogeneity: an EM approach. *Biometrika*, 96:411–426.

Li, S., Chen, J., and Li., P. (2016). *MixtureInf: Inference for Finite Mixture Models*. R package version 1.1.

Lindsay, B. (1995). *Mixture models: theory, geometry and applications*, volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, Hayward.

Liu, X. and Shao, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *Annals of Statistics*, 31:807–832.

Liu, X. and Shao, Y. (2004). Asymptotics for the likelihood ratio test in a two-component normal mixture model. *Journal Statistical Planning and Inference*, 123:61–81.

Maciejowska, K. (2013). Assessing the number of components in a normal mixture: an alternative approach. MPRA Paper 50303, University Library of Munich, Germany.

McLachlan, G. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appli. Statist.*, 36:318–324.

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

McLachlan, G. and Krishnan, T. (1997). *The EM algorithm and extensions*. Wiley and Sons, New-York.

McLachlan, G. and Peel, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.

Newcomb, S. (1882). Discussion and results of observations on transits of mercury from 1677 to 1881. *Astr. Papers*, 1:363–487.

Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American Journal of mathematics*, 8:343–366.

Pearson, K. (1894). Testing homogeneity in a multivariate mixture model. *Philosophical Transactions of the Royal Society of London, A*, 185:71–110.

Quetelet, A. (1846). *Lettres à S.A.R. le Duc régnant de Saxe-Cobourg et Gotha, sur la théorie des probabilités appliquée aux sciences morales et politiques*. Hayez, Bruxelles.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Redner, R. (1981). Note on the consistency of the maximum likelihood estimate for non identifiable distributions. *Ann. Statistics*, 9:225–228.

Saint Pierre, G. (2003). Identification du nombre de composants d'un mélange gaussien par maximum de vraisemblance dans le cas univarié. Technical report, Université Toulouse III. http://perso.lcpc.fr/guillaume.saint-pierre/Publis/These-Saint Pierre.pdf.

Schlattmann, P. (2009). *Medical applications of finite mixture models*. Statistics for Biology and Health. Springer-Verlag, Berlin, Heidelberg.

Thode, H., Finch, S., and Mendell, N. (1988). Simulated percentage points for the null distribution of the likelihood ratio for a mixture of two normals. *Biometrics*, 44:1195–1201.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Ltd., Chichester.

Wilks, S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9:60–62.

Wolfe, J. (1971). *A Monte-Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions*, volume 72-2 of *Technical Bulletin STB*. U.S. Nav. Pers. and Train. Res. Lab., San Diego.