Statistica Sinica Preprint No: SS-2016-0026R3	
Title	Using differential variability to increase the power of the homogeneity test in a two-sample problem
Manuscript ID	SS-2016-0026R3
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0026
Complete List of Authors	Guanfu Liu Yuejiao Fu Pengfei Li and Xiaolong Pu
Corresponding Author	Yuejiao Fu
E-mail	yuejiao@mathstat.yorku.ca

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

Submitted to Statistica Sinica

Using differential variability to increase the power of the homogeneity test in a two-sample problem

Guanfu Liu[1], Yuejiao Fu[2], Pengfei Li[3], and Xiaolong Pu[1]

[1]*East China Normal University,* [2]*York University,* [3]*University of Waterloo*

Abstract: We consider a two-sample homogeneity testing problem often encountered in case-control studies with contaminated controls, or in detecting a treatment effect when some subjects are not affected by the treatment in biological experiments. We propose an EM-test designed to simultaneously detect mean difference and differential variability in the two samples. We show that the EM-test statistic has a chi-squared null limiting distribution. The asymptotic properties of the EM-test under local alternatives are also investigated, and sample-size calculation is given. The main results are established for general location-scale family of distributions. Simulation results show that the EM-test outperforms existing methods, and two data examples are used to illustrate the application of the proposed method.

Key words and phrases: Differential variability, EM-test, Homogeneity test, Limiting distribution, Local power, Mixture model, Two-sample problem.

1. Introduction

Recently there has been increasing interest in the use of mixture models

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

in medical research. Motivated by case-control studies with contaminated controls, or in the detection of a treatment effect in biological experiments when some subjects are not affected by the treatment, a number of researchers have studied the problem of testing homogeneity in a two-sample problem in which one of the samples has a mixture structure (see Good (1979); Boos and Brownie (1991); Fu, Chen, and Kalbfleisch (2009); Qin and Liang (2011); Liu, Li, and Fu (2012)). Specifically, consider the independent samples

$$x_{11}, \ldots, x_{1n_1} \sim f_1(x), \quad x_{21}, \ldots, x_{2n_2} \sim (1-\lambda) f_1(x) + \lambda f_2(x). \tag{1.1}$$

The goal is to test the homogeneity of the two samples against the specified mixture alternative.

A number of testing methods designed for this situation have been proposed in the literature. They can be classified as nonparametric or semiparametric tests, and parametric tests. The semiparametric tests include Qin and Liang (2011)'s score test and Liu, Li, and Fu (2012)'s EM-EL test. These methods are built on the assumption that the logarithm of the component density ratio is linear in the observations and the component densities are otherwise unspecified. Fu, Chen, and Kalbfleisch (2009) proposed a modified likelihood ratio test (MLRT), which falls into the second category. They showed that the MLRT statistic has a simple chi-squared null

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

limiting distribution in mixtures of general one-parameter kernels , and in a situation where the kernels have an additional structural parameter (e.g., normal kernels with different means and common unknown variance). The assumption of homogeneity in the variance is the fundamental prerequisite for using the MLRT for this two-sample problem, but this assumption is often violated, for example, in DNA methylation studies. Indeed, some recent DNA methylation studies use differential variability as a tool for cancer risk marker selection; see Jaffe et al. (2012) and the references therein. When we seek to understand disease phenotypes, identifying markers that differ in terms of variability may be as important as identifying markers that differ in terms of mean. Teschendorff and Widschwendter (2012) have shown the merit of combining differential variability with differential mean when selecting cancer risk markers in DNA methylation studies.

Motivated by such studies, we aim to design an effective test for the situation where the component densities differ in both mean and variance. Specifically, suppose the component densities come from the same location-scale family of distributions, $f_1(x) = f(x; \mu_1, \sigma_1)$ and $f_2(x) = f(x; \mu_2, \sigma_2)$ with $f(x; \mu, \sigma) = \sigma^{-1} f\big((x-\mu)/\sigma; 0, 1\big)$. We wish to test

$$H_0 : \lambda \begin{pmatrix} \mu_1 - \mu_2 \\ \\ \sigma_1 - \sigma_2 \end{pmatrix} = \mathbf{0} \tag{1.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

under model (1.1). A two-component mixture of normal distributions, in both mean and variance parameters, is a typical example here.

Designing an effective method for testing (1.2) under (1.1) with component densities from a general location-scale distribution family is challenging, and there are two issues worth highlighting. First, the likelihood function is unbounded (Chen, Tan, and Zhang (2008); Tanaka (2009)). Second, the Fisher information on the mixing proportion can be infinite (Chen and Li (2009)). Therefore, we cannot directly apply the asymptotic results for such existing methods as the likelihood ratio test (Dacunha-Castelle and Gassiat (1999); Liu and Shao (2003)) and the D-test (Charnigo and Sun (2004, 2010)). Li, Chen, and Marriott (2009) and Chen and Li (2009) proposed a class of EM-tests for testing homogeneity in the two-component mixture model with a one-dimensional mixing parameter, in normal mixtures on the mean, and in normal mixtures on both mean and variance. Some of the ideas of the EM-test can be traced back to the MLRT, but it has several additional advantages. For example, its null limiting distribution does not depend on the finiteness of the Fisher information on the mixing proportion. Recently, the EM-test has been applied to test homogeneity in multivariate mixture models (Niu, Li, and Zhang (2011)), in linear switching autoregressive models (Ketterer and Holzmann (2012)),

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

and in subgroup analysis (Shen and He (2015)).

To the best of our knowledge, the EM-test has not been extended to test homogeneity in general location-scale mixture models in either one-sample or two-sample problems. By taking advantage of the special two-sample structure, we extend the EM-test to mixtures of the general location-scale family of distributions. We propose an EM-test that simultaneously detects differential variability and mean difference in the two-sample problem. It outperforms the MLRT in simulation studies and in data analyses. It also has the advantage that the asymptotic results hold without the need for a bounded parameter space.

Establishing the asymptotic properties for the proposed EM-test is technically challenging. A key step here is to show that any estimator with mixing proportion λ bounded away from 0 and with a large penalized log-likelihood value is consistent for μ_h and σ_h, $h = 1, 2$, under the null model. Chen, Tan, and Zhang (2008) showed the consistency of penalized maximum likelihood estimators of unknown parameters under normal mixtures in the one-sample case. Their results are not directly applicable to our problem since we are dealing with mixtures whose component density is from a general location-scale distribution family. Tanaka (2009) showed the consistency of penalized maximum likelihood estimators of unknown

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

parameters under a mixture of location-scale distributions with one-sample data and correctly specified the order of the mixture. However, his results cannot be used in our current setup since we overfit the number of components in the second sample. Despite the rather complicated derivations, we establish consistency for the component parameters (Lemma 1 in the supplementary material). Building on this, we prove that the EM-test statistic has a chi-squared null limiting distribution. We also investigate the asymptotic properties under local alternatives.

The rest of this paper is organized as follows. In Section 2, we provide a description of the EM-test procedure and its asymptotic results, including a local power analysis. We also discuss sample-size calculation. In Section 3, we use simulations to illustrate the empirical performance of the method. Section 4 presents data analyses, and Section 5 provides some discussion and a brief conclusion.

2. Main results

We propose an EM-test for the homogeneity problem (1.2), and establish its theoretical foundations and asymptotic properties.

2.1 EM-test

Let $n = n_1 + n_2$ be the total sample size. Based on the observed data,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

the log-likelihood function for $(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2)$ is

$$\begin{aligned} & l_n(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) \\ = & \sum_{i=1}^{n_1} \log f(x_{1i}; \mu_1, \sigma_1) + \sum_{i=1}^{n_2} \log\{(1-\lambda) f(x_{2i}; \mu_1, \sigma_1) + \lambda f(x_{2i}; \mu_2, \sigma_2)\}. \end{aligned}$$

Note that because of the nonregularity of the mixture in the second sample, the null hypothesis H_0 in (1.2) is on the boundary of the parameter space ($\lambda = 0$), and the parameters are not identifiable under the null hypothesis. In addition, the log-likelihood $l_n(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2)$ is unbounded, because $l_n(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) \to \infty$ when we set μ_2 equal to one of the data points in the second sample and $\sigma_2 \to 0$. To deal with these nonregularity problems, we propose the penalized log-likelihood function

$$pl_n(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) = l_n(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) + p(\lambda) + p_n(\sigma_2), \tag{2.1}$$

where $p(\lambda)$ is chosen such that it is maximized at $\lambda = 1$ and goes to negative infinity as λ goes to 0, and $p_n(\sigma_2)$ is selected such that it is bounded when σ_2 is large but goes to negative infinity as σ_2 goes to 0. The presence of the first sample automatically prevents the fitting of σ_1 close to 0, so there is no need for a penalty function on σ_1. Recommendations for $p(\lambda)$ and $p_n(\sigma_2)$ will be discussed in Section 3.

The motivation for the EM-test comes from a constrained likelihood ratio test in which the mixing proportion λ is set to a fixed value in

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

$(0, 1]$. The null hypothesis (1.2) with a fixed mixing proportion reduces to $(\mu_1, \sigma_1) = (\mu_2, \sigma_2)$ and the parameters are identifiable. Hence, we expect that the constrained likelihood ratio test has a χ^2-type limiting distribution. To improve the efficiency lost by setting λ to a fixed value, we suggest choosing multiple values for λ and then using the EM-algorithm (Dempster, Laird, and Rubin (1997)) to update the mixing proportion λ and the other parameters. The EM-test statistic is taken as the maximum of the likelihood ratio test statistics from the multiple values of λ. Our experience indicates that a few, say three, iterations suffice.

We provide the complete-data penalized log-likelihood function and the monotonicity of the EM algorithm in the supplementary material. Building on (2.1) and the complete-data penalized log-likelihood function, we develop an EM-test for the two-sample problem. The EM-test statistic can be constructed as follows. We first choose a set $\{\lambda_1, \ldots, \lambda_J\} \subset (0, 1]$, for example $\{0.1, 0.4, 0.7, 1.0\}$, as the initial values for λ and a positive integer K as the number of iterations.

Step 1. Set $k = 1$. For each $j = 1, 2, \ldots, J$, set $\lambda_j^{(1)} = \lambda_j$ and compute

$$(\mu_{1j}^{(1)}, \mu_{2j}^{(1)}, \sigma_{1j}^{(1)}, \sigma_{2j}^{(1)}) = \arg \max_{\{\mu_1, \mu_2, \sigma_1, \sigma_2\}} pl_n(\lambda_j^{(1)}, \mu_1, \mu_2, \sigma_1, \sigma_2).$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

Step 2. For $i = 1, \ldots, n_2$ and the current k, use an E-step to compute

$$w_{ij}^{(k)} = \frac{\lambda_j^{(k)} f(x_{2i}; \mu_{2j}^{(k)}, \sigma_{2j}^{(k)})}{(1-\lambda_j^{(k)}) f(x_{2i}; \mu_{1j}^{(k)}, \sigma_{1j}^{(k)}) + \lambda_j^{(k)} f(x_{2i}; \mu_{2j}^{(k)}, \sigma_{2j}^{(k)})}.$$

Update λ and the other parameters via an M-step such that

$$\lambda_j^{(k+1)} = \arg\max_{\lambda} \left\{ \left(n_2 - \sum_{i=1}^{n_2} w_{ij}^{(k)} \right) \log(1-\lambda) + \sum_{i=1}^{n_2} w_{ij}^{(k)} \log \lambda + p(\lambda) \right\},$$

$$(\mu_{1j}^{(k+1)}, \sigma_{1j}^{(k+1)}) = \arg \max_{\{\mu_1, \sigma_1\}} \left\{ \sum_{i=1}^{n_1} \log f_1(x_{1i}) + \sum_{i=1}^{n_2} (1 - w_{ij}^{(k)}) \log f_1(x_{2i}) \right\},$$

$$(\mu_{2j}^{(k+1)}, \sigma_{2j}^{(k+1)}) = \arg \max_{\{\mu_2, \sigma_2\}} \left\{ \sum_{i=1}^{n_2} w_{ij}^{(k)} \log f_2(x_{2i}) + p_n(\sigma_2) \right\}.$$

Iterate the E-step and M-step $K-1$ times.

Step 3. For each k and j, take

$$M_n^{(k)}(\lambda_j) = 2\{pl_n(\lambda_j^{(k)}, \mu_{1j}^{(k)}, \mu_{2j}^{(k)}, \sigma_{1j}^{(k)}, \sigma_{2j}^{(k)}) - pl_n(1, \hat{\mu}_0, \hat{\mu}_0, \hat{\sigma}_0, \hat{\sigma}_0)\},$$

where $(\hat{\mu}_0, \hat{\sigma}_0) = \arg\max_{\{\mu,\sigma\}} pl_n(1, \mu, \mu, \sigma, \sigma)$. The EM-test statistic is

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

Finally, reject the null hypothesis H_0 if $EM_n^{(K)}$ exceeds a prespecified critical value.

As indicated in Chen and Li (2009), the choices of the initial values λ_j and the integer K are not crucial; our simulation studies in Section 3 verify

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

this. Recommendations for these values, and for $p(\lambda)$ and $p_n(\sigma_2)$, will be discussed in Section 3.

2.2 Asymptotic properties

In this section, we present the asymptotic properties of the EM-test under regularity conditions given in the Appendix. The proofs are in the supplementary material. The following theorem sheds light on the iterative process.

Theorem 1. *Suppose Conditions* B1–B7 *and* C1–C5 *in the Appendix hold,* $n_h/n = \rho_h > 0$ $(h = 1, 2)$ *are constant, and* $\lambda_j \in (0, 1]$ *for* $j = 1, \ldots, J$. *Under the null distribution* $f(x; \mu_0, \sigma_0)$, *we have, for* $j = 1, \ldots, J$ *and any* $k \leq K$,

$$\lambda_j^{(k)} - \lambda_j = o_p(1), \ \ \mu_{hj}^{(k)} - \mu_0 = O_p(n^{-1/2}), \ \ \textit{and} \ \ \sigma_{hj}^{(k)} - \sigma_0 = O_p(n^{-1/2}), \ h = 1, 2.$$

The iteration changes the value of λ by only an $o_p(1)$ quantity. This is crucial, and results in a simple null limiting distribution for the EM-test.

Theorem 2. *Assume the conditions of Theorem 1, and that* $\lambda_1 = 1$. *Under the null distribution* $f(x; \mu_0, \sigma_0)$, *for any finite* K, *as* $n \to \infty$,

$$EM_n^{(K)} \xrightarrow{d} \chi_2^2.$$

Here $\xrightarrow{d}$ *stands for convergence in distribution, and* χ_m^2 *denotes the chi-squared distribution with* m *degrees of freedom.*

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

We observe that the regularity conditions on $f(x;\mu,\sigma)$ are not restrictive. The commonly used location-scale distributions such as the normal, t, logistic, and extreme value distributions all satisfy Conditions B1–B7. Examples of functions satisfying Conditions C1–C5 are given in Section 3. Since the user has the freedom to choose the penalty functions, these conditions are not restrictive as long as such functions exist.

Asymptotic local power analysis has become an important and increasingly used tool in statistical inference. To investigate the asymptotic local power of the EM-test, we consider the local alternative

$$H_a^n : \lambda = \lambda_0,\ (\mu_1,\sigma_1) = (\mu_0,\sigma_0),\ (\mu_2,\sigma_2) = (\mu_0 + n_2^{-1/2}\Delta_1, \sigma_0 + n_2^{-1/2}\Delta_2), \tag{2.2}$$

where $0 < \lambda_0 \leq 1$, $(\Delta_1,\Delta_2) \neq (0,0)$, and $\Delta_2 > -\sigma_0$. Let $\chi^2_m(c)$ denote the noncentral chi-squared distribution with noncentrality parameter c and m degrees of freedom.

Theorem 3. *Assume the conditions of Theorem 2. Under the local alternative H_a^n in (2.2),*

$$EM_n^{(K)} \xrightarrow{d} \chi^2_2(c_0^2),$$

where $c_0^2 = \lambda_0^2\rho_1\{\Delta_1^2 E(U^2) + 2\Delta_1\Delta_2 E(UV) + \Delta_2^2 E(V^2)\}$ *with*

$$U = \frac{\partial f(x_{11};\mu_0,\sigma_0)/\partial\mu}{f(x_{11};\mu_0,\sigma_0)} \quad \textit{and} \quad V = \frac{\partial f(x_{11};\mu_0,\sigma_0)/\partial\sigma}{f(x_{11};\mu_0,\sigma_0)},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

and the expectation is taken under $f(x;\mu_0,\sigma_0)$.

From the above theorem, we could perform asymptotic local power analysis to gain more insight into the testing problem. Specifically, we could compare the asymptotic local power of the EM-test and the MLRT theoretically. Let M_n be the MLRT statistic. Let f be the normal distribution and $(\mu_0,\sigma_0)=(0,1)$. In this setup, M_n and $EM_n^{(K)}$, respectively, converge to $\chi_1^2(\lambda_0^2\rho_1\Delta_1^2)$ and $\chi_2^2\{\lambda_0^2\rho_1(\Delta_1^2+2\Delta_2^2)\}$ in distribution under the local alternative H_a^n. The theoretical power functions for the MLRT and the EM-test at the 5% significance level are respectively $P(M_n > \chi^2_{1,0.95})$ and $P(EM_n^{(K)} > \chi^2_{2,0.95})$, where $\chi^2_{m,1-\alpha}$ is the $1-\alpha$ upper quantile of the χ^2_m distribution. Figure 1 presents the asymptotic power curves of the two methods at the 5% significance level. Panel (a) shows that when $\Delta_2 = 0$ the asymptotic local power of the EM-test is always lower than that of the MLRT under the same Δ_1. If $\Delta_2 > 0$, the EM-test can perform much better than the MLRT in terms of power. Panel (b) shows that the asymptotic local power of the EM-test increases as Δ_2 increases, while that of the MLRT stays the same.

Based on Theorem 3, we can calculate the sample size required to obtain a target power $1-\beta$. For the significance level α and the given ρ_1, the sample

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

Figure 1: Dashed curve: MLRT; Solid curve: EM-test.

size n_2 can be calculated by solving

$$\begin{cases} \Delta_1 = n_2^{1/2}(\mu_2 - \mu_1) \\ \Delta_2 = n_2^{1/2}(\sigma_2 - \sigma_1) \\ P(\chi_2^2(c_0^2) > \chi_{2,1-\alpha}^2) = 1 - \beta. \end{cases} \tag{2.3}$$

The size of the first sample can be calculated via $n_1 = \rho_1 n_2/(1-\rho_1)$. Because of the complicated form of $P(\chi_2^2(c_0^2) > \chi_{2,1-\alpha}^2)$, we have no explicit formula for n_1 or n_2. In the supplementary material, we provide R code to solve (2.3) for the normal and logistic kernels.

3. Simulation studies

We conducted simulations to compare the finite-sample performance of the EM-test with that of the MLRT in Fu, Chen, and Kalbfleisch (2009) under the normal and logistic kernels. We present empirical type I errors and the power of the two methods for different combinations of $(n_1, n_2, \lambda, \mu_2, \sigma_2)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

To implement the MLRT, we must specify a penalty function on λ: for the normal kernel, we took $2\log(\lambda)$ as suggested by Fu, Chen, and Kalbfleisch (2009); for the logistic kernel, we chose $2.5\log(\lambda)$ based on our empirical study. To calculate the EM-test statistics, we must specify K, $\{\lambda_1, \ldots, \lambda_J\}$, $p(\lambda)$, and $p_n(\sigma)$. We used $K = 3$ as recommended by Chen and Li (2009), and chose $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.4, 0.7, 1.0\}$ or $\{0.1, 0.2, \ldots, 1.0\}$. Following Fu, Chen, and Kalbfleisch (2009), we chose $p(\lambda) = a_1 \log(\lambda)$, and following Chen and Li (2009), we chose $p_n(\sigma_2) = -a_2\{\hat{\sigma}^2/\sigma_2^2 + \log(\sigma_2^2/\hat{\sigma}^2)\}$, where $\hat{\sigma}$ is the maximum likelihood estimator of σ_0 under the null. These penalties satisfy the regularity conditions in the Appendix. Furthermore, under the normal and logistic kernels, our simulation studies suggest that $a_1 = 1$ and $a_2 = 1.5$ work well in terms of accurate type I error and reasonable power. In the simulation, all the empirical type I errors were based on 10,000 repetitions, and the power values were based on 1,000 repetitions. The simulation results for the normal and logistic kernels are quite similar. The simulation results for the logistic kernel are given in the supplementary material.

Table 1 gives the empirical type I errors of the MLRT and EM-test for the nominal type I errors $\alpha = 0.05$ and $\alpha = 0.01$. The data were generated from $N(0, 1)$. The critical values for the two test statistics were based

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

on their asymptotic distributions. The EM-test performs better than the MLRT in terms of simulated type I error rates. In the supplementary material, we give the quantile-quantile plot of $EM^{(3)}$, showing that its limiting distribution provides an accurate approximation to its finite-sample distribution.

Table 1: Type I error comparison for $f_1 = f_2 = N(0, 1)$.

	$\alpha = 0.05$			$\alpha = 0.01$		
	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$
$n_1 = 50, n_2 = 50$	0.0612	0.0523	0.0531	0.0143	0.0111	0.0111
$n_1 = 50, n_2 = 100$	0.0596	0.0542	0.0550	0.0129	0.0120	0.0121
$n_1 = 100, n_2 = 50$	0.0557	0.0526	0.0529	0.0128	0.0122	0.0124
$n_1 = 100, n_2 = 100$	0.0555	0.0519	0.0525	0.0119	0.0105	0.0106

Results in columns 3 and 6 used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.4, 0.7, 1.0\}$. Results in columns 4 and 7 used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.2, \ldots, 1.0\}$.

For the normal kernel, Tables 2–4 compare the power of the MLRT and the EM-test. For a fair comparison, the critical values of the two methods were obtained by 10,000 simulations from the homogenous model $N(0, 1)$. The data were generated from f_1 and $(1-\lambda)f_1+\lambda f_2$, where $f_1 = N(0, 1)$ and $f_2 = N(\mu_2, \sigma_2^2)$. From Table 2, we see that when f_1 and f_2 have different

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

means and the same variance the MLRT performs better than the EM-test. In contrast, from Table 3, when f_1 and f_2 have the same means but different variances, the EM-test is superior to the MLRT. From Table 4, when f_1 and f_2 differ in both mean and variance, the EM-test is again more powerful. For both methods, the power of the test increases with sample size and with λ closer to 1.

As suggested by the associate editor, we also compared the performance of the EM-test with that of the two-sample t-test with unequal variance and the Wilcoxon rank sum test, in the situation where there is no mixture. We conducted simulations for the normal, logistic, and Gumbel distributions. They are discussed in the supplementary material.

From Tables 1–4 and the simulation results in the supplementary material, we see that for the EM-test the number of λ_j values has little effect on the empirical type I error or the power, so we recommend using $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.4, 0.7, 1.0\}$ in practice.

4. Data examples

We compared the two-sample t-test, the MLRT, and the EM-test by analyzing two data examples. The MLRT and EM-test are calculated as in Section 3. The p-values of the MLRT and EM-test were based on simulated distributions.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

Table 2: Power comparison for $f_1 = N(0,1)$ and $f_2 = N(0.5,1)$.

	$\alpha = 0.05$			$\alpha = 0.01$		
$\lambda = 0.9$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$
$n_1 = 50, n_2 = 50$	0.577	0.494	0.492	0.298	0.233	0.233
$n_1 = 50, n_2 = 100$	0.722	0.657	0.658	0.472	0.385	0.382
$n_1 = 100, n_2 = 50$	0.728	0.625	0.625	0.458	0.371	0.371
$n_1 = 100, n_2 = 100$	0.881	0.825	0.825	0.706	0.638	0.635
	$\alpha = 0.05$			$\alpha = 0.01$		
$\lambda = 0.7$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$
$n_1 = 50, n_2 = 50$	0.394	0.335	0.330	0.191	0.154	0.154
$n_1 = 50, n_2 = 100$	0.470	0.409	0.406	0.239	0.188	0.190
$n_1 = 100, n_2 = 50$	0.503	0.417	0.416	0.250	0.205	0.205
$n_1 = 100, n_2 = 100$	0.677	0.592	0.595	0.441	0.371	0.369

Results in columns 3 and 6 used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.4, 0.7, 1.0\}$. Results in columns 4 and 7 used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.2, \ldots, 1.0\}$.

The first example concerns morphine addiction in rats; see Weeks and Collins (1971), Good (1979), Boos and Brownie (1991), and Fu, Chen and Kalbfleisch (2009). In an experiment, rats could obtain morphine by pressing a lever. The frequency of lever presses (self-injection rates) after six

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

Table 3: Power comparison for $f_1 = N(0,1)$ and $f_2 = N(0, 1.5^2)$.

	$\alpha = 0.05$			$\alpha = 0.01$		
$\lambda = 0.9$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$
$n_1 = 50, n_2 = 50$	0.305	0.643	0.639	0.137	0.390	0.392
$n_1 = 50, n_2 = 100$	0.304	0.767	0.765	0.127	0.511	0.507
$n_1 = 100, n_2 = 50$	0.501	0.805	0.805	0.273	0.606	0.606
$n_1 = 100, n_2 = 100$	0.551	0.914	0.914	0.300	0.783	0.783
	$\alpha = 0.05$			$\alpha = 0.01$		
$\lambda = 0.7$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$
$n_1 = 50, n_2 = 50$	0.222	0.514	0.508	0.088	0.270	0.270
$n_1 = 50, n_2 = 100$	0.218	0.632	0.630	0.094	0.369	0.368
$n_1 = 100, n_2 = 50$	0.367	0.615	0.614	0.173	0.399	0.399
$n_1 = 100, n_2 = 100$	0.436	0.809	0.809	0.212	0.622	0.619

Results in columns 3 and 6 used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.4, 0.7, 1.0\}$. Results in columns 4 and 7 used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.2, \ldots, 1.0\}$.

days' treatment with morphine was recorded as the response variable. The self-injection rates for five groups of rats corresponding to four different dose levels and one saline control are presented in Figure 1 of Good (1979). Following Fu, Chen, and Kalbfleisch (2009), we are interested in compar-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

Table 4: Power comparison for $f_1 = N(0,1)$ and $f_2 = N(0.5, 1.5^2)$.

	$\alpha = 0.05$			$\alpha = 0.01$		
$\lambda = 0.9$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$
$n_1 = 50, n_2 = 50$	0.694	0.833	0.831	0.443	0.617	0.618
$n_1 = 50, n_2 = 100$	0.778	0.918	0.918	0.572	0.755	0.755
$n_1 = 100, n_2 = 50$	0.834	0.917	0.917	0.690	0.784	0.785
$n_1 = 100, n_2 = 100$	0.944	0.989	0.989	0.836	0.949	0.949
	$\alpha = 0.05$			$\alpha = 0.01$		
$\lambda = 0.7$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$
$n_1 = 50, n_2 = 50$	0.543	0.668	0.666	0.328	0.442	0.443
$n_1 = 50, n_2 = 100$	0.645	0.799	0.797	0.432	0.585	0.587
$n_1 = 100, n_2 = 50$	0.682	0.792	0.792	0.493	0.600	0.600
$n_1 = 100, n_2 = 100$	0.840	0.936	0.936	0.703	0.824	0.824

Results in columns 3 and 6 used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.4, 0.7, 1.0\}$. Results in columns 4 and 7 used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.2, \ldots, 1.0\}$.

ing the self-injection rates of the treatment group (at dose level 1.0) and the control group. We used the same data transformation as in Fu, Chen, and Kalbfleisch (2009). The p-value of the Kolmogorov–Smirnov (K-S) test for the control data was 0.525, which shows that the transformed control

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

data can be assumed to be drawn from a normal distribution. We applied the t-test, the MLRT, and the EM-test to the transformed data, and the p-values were respectively 0.0249, 0.0344, and 0.0163. The EM-test has the smallest p-value and thus the highest power.

The second example is a DNA methylation study that measured the methylation levels of 27,578 CpG sites for 152 women. Cytologically normal cells from the uterine cervix of the 152 women were used. The study found that 75 women developed a cervical intraepithelial neoplasia of grade 2 or higher (CIN2+) within a three year follow-up period (cases), whereas 77 women did not develop any abnormal cytology (controls). This dataset is available from the Gene Expression Ominbus (GEO; www.ncbi.nlm.nih.gov/geo/) with accession number GSE30760. After data quality checking and cleaning (excluding CpG sites associated with single-nucleotide polymorphisms and regressing out batch effects), we kept 22,399 CpG sites for our analysis. For each CpG site, we applied the K-S test to the 77 control samples. The proportions of K-S p-values that were greater than 0.05, 0.10, 0.25, and 0.5 were respectively 0.778, 0.689, 0.518, and 0.309. Let d_α be those CpG sites corresponding to K-S p-values greater than α.

With f_1 and f_2 as normal kernels, we applied the t-test, MLRT, and EM-test to the data in $d_{0.5}$, which has $M = 6920$ CpG sites. Table 5 shows

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

the proportions of p-values that were less than or equal to α and α/M with $\alpha = 0.10, 0.05, 0.01$ for the t-test, MLRT, and EM-test. The EM-test has the highest proportion and thus the greatest power. For illustrative purposes, we chose 16 CpG sites for further analysis; the details are given in the supplementary material.

Table 5: Proportion of p-values less than or equal to α and α/M for the t-test, MLRT, and EM-test. Here M=6920 is the number of CpG sites in $d_{0.5}$.

α	0.10	0.05	0.01	0.10/M	0.05/M	0.01/M
t-test	0.103	0.054	0.011	0	0	0
MLRT	0.152	0.104	0.060	0.031	0.031	0.031
$EM_n^{(3)}$	0.239	0.176	0.107	0.042	0.042	0.042

Results in row 4 used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.4, 0.7, 1.0\}$.

5. Discussion and conclusion

Motivated by many problems in medical research, we considered a special two-sample homogeneity testing problem where one of the two samples has a mixture structure. The presence of mixture in one of the samples, the unboundedness of the likelihood function, and the possibly infinite Fisher information in the direction of the mixing proportion make the test partic-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

ularly challenging. We proposed an EM-test that simultaneously tests for difference in the mean and variance of the component densities. We also extended the literature of the EM-test by considering the homogeneity test in mixtures of general location-scale distributions. For future research, we may consider the one-sample problem of homogeneity testing, or testing the number of components in a mixture of general location-scale distributions.

As suggested by the associate editor, we give the guidelines for the use of our method. In the analysis of two-sample data, the proposed EM-test should always be considered if there is some possibility of a mixture in one of the two samples; see the examples in Qin and Liang (2011). The preliminary data analysis and the EM-test should be used jointly. If the preliminary analysis suggests the presence of a mixture, the EM-test can be used for confirmation.

Appendix

Notation and regularity conditions

The asymptotic properties of the EM-test rely on regularity conditions on $f(x;\mu,\sigma)$, $p(\lambda)$, and $p_n(\sigma)$. We impose mild regularity conditions on $f(x;\mu,\sigma)$ in which the expectations are taken under the distribution $f(x;\mu_0,\sigma_0)$.

B1. (Wald's integrability conditions) (i) $E\{|\log f(x;\mu_0,\sigma_0)|\} < \infty$;

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

(ii) for sufficiently small ρ and sufficiently large r, $E[\log\{1+f(x;\mu,\sigma,\rho)\}] < \infty$ for $(\mu,\sigma) \in \Theta$ and $E[\log\{1+\phi(x;r)\}] < \infty$, where Θ is the parameter space of (μ,σ), $f(x;\mu,\sigma,\rho) = \sup_{|\mu'-\mu|^2+|\sigma'-\sigma|^2\le\rho} f(x;\mu',\sigma')$, and $\phi(x;r) = \sup_{\mu^2+\sigma^2\ge r} f(x;\mu,\sigma)$; (iii) $f(x;\mu,\sigma) \to 0$ in probability as $\mu^2+\sigma^2 \to \infty$.

B2. (Smoothness) The kernel $f(x;\mu,\sigma)$ has common support and is three times continuously differentiable with respect to μ and σ.

B3. (Identifiability) For any two mixing distribution functions Ψ_1 and Ψ_2 with two supporting points such that $\int f(x;\mu,\sigma)\, d\Psi_1(\mu,\sigma) = \int f(x;\mu,\sigma)\, d\Psi_2(\mu,\sigma)$ for all x, we must have $\Psi_1 = \Psi_2$.

B4. (Uniform boundedness) There exists a function g with finite expectation such that

$$\Big|\frac{\partial^{(h+l)} f(x;\mu_0,\sigma_0)/\partial\mu^h\partial\sigma^l}{f(x;\mu_0,\sigma_0)}\Big|^3 \le g(x), for\ h+l\le 2,$$

where h and l are two nonnegative integers. Moreover, there exists a positive ϵ such that

$$\sup_{|\mu-\mu_0|^2+|\sigma-\sigma_0|^2\le\epsilon}\Big|\frac{\partial^{(h+l)} f(x;\mu,\sigma)/\partial\mu^h\partial\sigma^l}{f(x;\mu_0,\sigma_0)}\Big|^3 \le g(x), for\ h+l=3.$$

B5. (Positive definiteness) The covariance matrix of (U,V) is positive

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

definite, where

$$U = \frac{\partial f(x_{11};\mu_0,\sigma_0)/\partial\mu}{f(x_{11};\mu_0,\sigma_0)} \quad \text{and} \quad V = \frac{\partial f(x_{11};\mu_0,\sigma_0)/\partial\sigma}{f(x_{11};\mu_0,\sigma_0)}.$$

B6. (Tail condition) There exist positive constants v_0, v_1, and β_0 with $\beta_0 > 1$ such that $f(x;0,1) \le \min\{v_0, v_1|x|^{-\beta_0}\}$.

B7. (Upper bound function) There exist a nonnegative function $s(x;\mu,\sigma)$ that satisfies Condition B1 and is continuous in (μ,σ), a positive number a with $1/\beta_0 < a < 1$, a positive number b, and a positive number ϵ^* with $\epsilon^* < 1$ such that for $\sigma \in (0, \epsilon^*\sigma_0)$, $s(x;\mu,\sigma)$ is uniformly bounded, $\int s(x;\mu,\sigma)dx < 1$, and

$$f(x;\mu,\sigma) \le \begin{cases} \sigma^{-1}s(x;\mu,\sigma), & \text{if } |x-\mu| \le \sigma^{1-a} \\ \sigma^b s(x;\mu,\sigma), & \text{if } |x-\mu| > \sigma^{1-a} \end{cases}.$$

We list regularity conditions for $p(\lambda)$ and $p_n(\sigma)$.

C1. $p(\lambda)$ is a continuous function that is maximized at $\lambda = 1$ and goes to negative infinity as $\lambda \to 0$.

C2. $\sup_{\sigma>0} \max\{p_n(\sigma), 0\} = o(n)$ and $p_n(\sigma) = o(n)$ for any σ.

C3. $p_n'(\sigma) = o_p(n^{1/2})$ for all $\sigma > 0$, where $p_n'(\sigma)$ is the derivative function with respect to σ.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

C4. $p_n(\sigma) \leq 4(\log n_2)^2 \log(\sigma)$, when $0 < \sigma \leq 8/(n_2 M_0)$ and n_2 is large. Here $M_0 = \max\{\sup_x f(x; \mu_0, \sigma_0), 8\}$.

We allow p_n to depend on the data. To ensure that the EM-test is location-scale invariant, we recommend choosing a p_n that satisfies

C5. $p_n(b_1\sigma; b_1X_1 + b_0, ..., b_1X_n + b_0) = p_n(\sigma; X_1, ..., X_n)$.

Supplementary Material

The online supplementary material includes the definition of the complete-data penalized log-likelihood, the monotonicity property of the EM-algorithm, some additional simulation results, R code for the sample-size calculation, further analysis of the second set of data, regularity conditions, and the proofs of Theorems 1–3.

Acknowledgements

Dr. Pu's research is supported by grants from the National Natural Science Foundation of China (11271135, 11471119, 11501208, 11501209), the outstanding doctoral dissertation cultivation plan of action (PY2015049), the Postdoctoral Science Foundation of China (2015M570348), the Program of Shanghai Subject Chief Scientist (14XD1401600), and the 111 Project (B14019). Dr. Fu's research was partially supported by NSERC Discovery

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

Grant. Dr. Li's research was partially supported by NSERC Grant RGPIN 2015 06592. The authors would like to thank the Editor, an AE, and two referees for their valuable suggestions and comments.

References

Boos, D. D. and Brownie, C. (1991). Mixture models for continuous data in dose-response studies when some animals are unaffected by treatment. *Biometrics* **47**, 1489–1504.

Charnigo, R. and Sun, J. (2004). Testing homogeneity in a mixture distribution via the L^2-distance between competing models. *J. Amer. Statist. Assoc.* **99**, 488–498.

Charnigo, R. and Sun, J. (2010). Asymptotic relationships between the D-test and likelihood ratio-type tests for homogeneity. *Statist. Sinica* **20**, 497–512.

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

Chen, J., Tan, X. and Zhang, R. (2008). Inference for normal mixtures in mean and variance. *Statist. Sinica* **18**, 443–465.

Dacunha-Castelle, D. and Gassiat, E. (1999). Testing the order of a model using locally conic parametrization: Population mixtures and stationary ARMA processes. *Ann. Statist.* **27**, 1178–1209.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

data via EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39**, 1–38.

Fu, Y., Chen, J. and Kalbfleisch, J. D. (2009). Modified likelihood ratio test for homogeneity in a two-sample problem. *Statist. Sinica* **19**, 1603–1619.

Good, P. I. (1979). Detection of a treatment effect when not all experimental subjects will respond to treatment. *Biometrics* **35**, 483–489.

Jaffe, A. E., Feinberg, A. P., Irizarry, I. R. and Leek, J. T. (2012). Significance analysis and statistical dissection of variably methylated regions. *Biostatistics* **13**, 166–178.

Ketterer, F. and Holzmann, H. (2012). Testing for intercept-scale switch in linear autoregression. *Canad. J. Statist.* **40**, 427–446.

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

Liu, Y., Li, P. and Fu, Y. (2012). Testing homogeneity in a semiparametric two-sample problem. *J. Probab. Statist.*, article ID 537474, doi: 10.1155/2012/537474.

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

Niu, X., Li, P. and Zhang, P. (2011). Testing homogeneity in a multivariate mixture model. *Canad. J. Statist.* **39**, 218–238.

Qin, J. and Liang, K.-Y. (2011). Hypothesis testing in a mixture case-control model. *Biometrics* **67**, 182–193.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0026

Shen, J. and He, X. (2015). Inference for subgroup analysis with a structured logistic-normal mixture model. *J. Amer. Statist. Assoc.* **110**, 303–312.

Tanaka, K. (2009). Strong consistency of the maximum likelihood estimator for finite mixtures of location-scale distributions when penalty is imposed on the ratios of the scale parameters. *Scand. J. Statist.* **36**, 171–184.

Teschendorff, A.E. and Widschwendter, M. (2012). Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. *Bioinformatics* **28**, 1487–1494.

Weeks, J. R. and Collins, R. J. (1971). Primary addiction to morphine in rats. *Federation Proceedings* **30**, 277 ABS.

School of Statistics, East China Normal University, Shanghai 200241, China.

E-mail: liuguanfu07@163.com

Department of Mathematics and Statistics, York University, Toronto, ON, Canada M3J 1P3.

E-mail: yuejiao@mathstat.yorku.ca

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada N2L 3G1.

E-mail: pengfei.li@uwaterloo.ca

School of Statistics, East China Normal University, Shanghai 200241, China.

E-mail: xlpu@stat.ecnu.edu.cn