Scandinavian Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics, Vol. 42: 504–517, 2015
doi: 10.1111/sjos.12119

A Class of Pseudolikelihood Ratio Tests for Homogeneity in Exponential Tilt Mixture Models

YANG NING
Department of Statistics and Actuarial Science, University of Waterloo
YONG CHEN
Division of Biostatistics, The University of Texas School of Public Health

ABSTRACT. Mixture models are commonly used in biomedical research to account for possible heterogeneity in population. In this paper, we consider tests for homogeneity between two groups in the exponential tilt mixture models. A novel pairwise pseudolikelihood approach is proposed to eliminate the unknown nuisance function. We show that the corresponding pseudolikelihood ratio test has an asymptotic distribution as a supremum of two squared Gaussian processes under the null hypothesis. To maintain the appeal of simplicity for conventional likelihood ratio tests, we propose two alternative tests, both shown to have a simple asymptotic distribution of χ_1^2 under the null. Simulation studies show that the proposed class of pseudolikelihood ratio tests performs well in controlling type I errors and having competitive powers compared with the current tests. The proposed tests are illustrated by an example of partial differential expression detection using microarray data from prostate cancer patients.

Key words: composite likelihood, density ratio model, non-regular problem, pairwise conditional likelihood

1. Introduction

Mixture models have been widely used in biomedical research to account for possible heterogeneity in population. An important problem in mixture models is testing for homogeneity. A variety of tests for homogeneity in parametric mixture models have been proposed (Lemdani & Pons, 1999; Liang & Rathouz, 1999). It has been long recognized that testing for homogeneity in mixture models is a non-regular problem because the mixture proportion parameter lies on the boundary of its parameter space and some parameters may not be identifiable under the null (Mclachlan & Basford, 1988). Thus, the asymptotic distributions of tests of homogeneity are often not chi-square distributions and are usually rather complicated and possibly dependent on the parametric distributions assumed. An excellent review of inference on parametric mixture models can be found in Titterington *et al.* (1985) and Lindsay (1995), among many others.

In this paper, we limit our attention to test of homogeneity between two groups. Important applications under this setting include genetic linkage analysis (Di & Liang, 2011; Fu *et al.*, 2006; Lemdani & Pons, 1995; 1997; Liang & Rathouz, 1999; Zhu & Zhang, 2004), case–control studies with contaminated controls (Lancaster & Imbens, 1996; Lemdani & Pons, 1999; Steinberg & Cardell, 1992) and testing for partial differential gene expression in microarray study (Ghosh & Chinnaiyan, 2009; Van Wieringen & Van de Viel, 2009). However, unlike the fully parametric mixture models, we allow the density function from one group to be completely unspecified. Specifically, consider an exponential tilt mixture model (ETMM) where two groups of $\{x_1, \ldots, x_{n_0}\}$ and $\{y_1, \ldots, y_{n_1}\}$ are drawn from the density functions $f(x)$ and $h(y)$, respectively. The density function $h(y)$ is a mixture of two distributions, defined as

$$h(y) = (1-\lambda) f(y) + \lambda g(y), \tag{1.1}$$

where $0 \leq \lambda \leq 1$ is an unknown mixture proportion, and

$$\log\{g(y)/f(y)\} = \alpha + \beta y, \tag{1.2}$$

where $f(\cdot)$ is an unknown baseline density function, β is an unknown parameter and $\alpha = -\{\int \exp(\beta y) f(y)\, \mathrm{d}y\}$ is a normalizing constant for density function $g(y)$. We note that $\beta = 0$ implies $\alpha = 0$. Under the ETMM, testing for homogeneity between two groups is equivalent to testing

$$H_0 : \lambda = 0 \quad \text{or} \quad \beta = 0.$$

Because only parameters λ and β are tested, we treat the unknown function $f(\cdot)$ as a nuisance function. The exponential tilt model (1.2), also known as the density ratio model, has been considered by many authors (Anderson, 1979; Qin, 1998; Rathouz & Gao, 2009). The exponential tilt model is flexible in the sense that many parametric models such as normal, Gamma, Weibull and -normal for continuous variables and binomial and Poisson for discrete variables are included as special cases. One attractive feature of the ETMM, compared with the parametric mixture models, is that the distributions are modelled non-parametrically except for a parametric 'exponential tilt' that is used to relate one distribution to the other.

The ETMM has been studied under the assumption of known mixture proportion λ (Tan, 2009; Zou *et al.*, 2002). Even with known λ, standard likelihood-based inference cannot be applied because of the singularity of the information matrix. To overcome this difficulty, a partial empirical likelihood method (Zou *et al.*, 2002) and a profile likelihood method (Tan, 2009) have been proposed. However, to the best of our knowledge, the work on testing for homogeneity with unknown mixture proportion is sparse. Important contributions to this topic are recently made by Qin & Liang (2011) and Liu *et al.* (2012), where a convenient score test (ST) and a novel modified empirical likelihood test are proposed, respectively.

In this paper, we propose a class of pseudolikelihood ratio tests for homogeneity in the ETMM. Unlike the tests proposed by Qin & Liang (2011) and Liu *et al.* (2012) where the nuisance function $f(\cdot)$ is eliminated by profiling in the empirical likelihood method, the proposed tests are based on a novel pairwise pseudolikelihood. The key idea is to eliminate $f(\cdot)$ by conditioning a pair of observations from two groups on the order statistics. Such a conditioning technique was first proposed by Kalbfleisch (1978) in the setting of non-parametric tests and later applied by Liang & Qin (2000) in regression analysis of cross-sectional data. The corresponding pseudolikelihood ratio test avoids the estimation of nuisance function $f(\cdot)$. However, the pseudolikelihood ratio test still faces the boundary, non-identifiability and singular sensitivity matrix problems. Its asymptotic distribution has to be considered separately in different regions of the parameter space. We show that the asymptotic distribution of the pseudolikelihood ratio test is a supremum of two squared Gaussian processes. To maintain the appeal of simplicity for conventional likelihood ratio tests, we derive two alternative tests, both shown to have a simple asymptotic distribution of χ^2_1, namely a modified pseudolikelihood ratio test (MPLRT) and a restricted pseudolikelihood ratio test (RPLRT). The asymptotic local powers of MPLRT and RPLRT are derived and found to be equivalent to the tests proposed by Qin & Liang (2011) and Liu *et al.* (2012). Because the impact of unknown nuisance function $f(\cdot)$ is minimized by the conditioning procedure, we expect that the pairwise pseudolikelihood-based tests outperform the empirical likelihood-based tests in finite samples.

The paper is organized as follows. Section 2 describes the pairwise pseudolikelihood method. The asymptotic results for the pseudolikelihood ratio test and its variants are provided in

Section 3. Section 4 provides the estimation procedure of the two distribution functions and a goodness-of-fit test for ETMM. Simulation studies comparing type I errors and power functions of the proposed tests with current tests are summarized in Section 5. A real data application is given in Section 6.

2. Proposed method

Under the ETMM described in (1.1) and (1.2), the estimation of α, β and λ is complicated because of the presence of the unknown function $f(\cdot)$. In this section, we propose a novel pseudolikelihood method to eliminate the unknown function $f(\cdot)$. Our method is motivated by the proportional structure of $g(y)$ and $f(y)$ specified via (1.2).

Consider a pair of observations: x_i from the first group and y_j from the second group. The conditional density of (x_i, y_j), given their order statistics $(t^{(1)}, t^{(2)})$, is

$$\begin{aligned}
&\mathrm{P}\left(x_i, y_j | t^{(1)}, t^{(2)}\right) \\
&= \frac{f(x_i)h(y_j)}{f(x_i)h(y_j) + f(y_j)h(x_i)} \\
&= \frac{f(x_i)\left\{(1-\lambda)f(y_j) + \lambda f(y_j)\exp(\alpha+\beta y_j)\right\}}{f(x_i)\left\{(1-\lambda)f(y_j) + \lambda f(y_j)\exp(\alpha+\beta y_j)\right\} + f(y_j)\left\{(1-\lambda)f(x_i) + \lambda f(x_i)\exp(\alpha+\beta x_i)\right\}} \\
&= \frac{(1-\lambda) + \lambda\exp(\alpha+\beta y_j)}{\left\{(1-\lambda) + \lambda\exp(\alpha+\beta y_j)\right\} + \left\{(1-\lambda) + \lambda\exp(\alpha+\beta x_i)\right\}} \\
&= \left\{1 + R(x_i, y_j; \lambda, \alpha, \beta)\right\}^{-1},
\end{aligned} \tag{2.1}$$

where

$$R(x_i, y_j; \lambda, \alpha, \beta) = \frac{(1-\lambda) + \lambda\exp(\alpha+\beta x_i)}{(1-\lambda) + \lambda\exp(\alpha+\beta y_j)}.$$

Note that $f(\cdot)$ is cancelled in (2.1), because $g(y)$ is proportional to $f(y)$ in (1.2). Thus, $\mathrm{P}(x_i, y_j | t^{(1)}, t^{(2)})$ does not involve $f(\cdot)$ explicitly. However, it depends on $f(\cdot)$ implicitly through $\alpha = -\int f(x)\exp(\beta x)\,\mathrm{d}x$. Because $f(\cdot)$ is unknown, α contains little information about β when β is not close to 0. Similar to Qin & Liang (2011) and Liu *et al.* (2012), we therefore treat α as a free parameter when $\beta \neq 0$.

For each possible pair of observations (x_i, y_j), we can calculate the pairwise conditional density (2.1). Although the pairs with common observations are no longer independent, following the spirit of composite likelihoods (Besag, 1974; Cox & Reid, 2004; Lindsay, 1988; Varin *et al.*, 2011), one can still multiply these conditional densities together. The pseudolikelihood function for all observations is given by

$$l_p(\lambda, \alpha, \beta) = \frac{1}{n}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} -\log\left\{1 + R(x_i, y_j; \lambda, \alpha, \beta)\right\},$$

where $n = n_0 + n_1$. Note that $l_p(\lambda, \alpha, \beta)$ is not a true likelihood function. In fact, the proposed pairwise pseudolikelihood belongs to the family of composite likelihoods. We refer the readers to a recent review paper of composite likelihood methods by Varin *et al.* (2011). Although the composite likelihood has been widely used in many applications, the major motivation for the use of composite likelihood is for computational simplicity or model robustness. However, in our proposed method, the major motivation is to eliminate the nuisance function $f(\cdot)$.

Under H_0, the pseudolikelihood reduces to $l_p(0,\alpha,\beta) = -n_0 n_1 n^{-1}$ 2. We denote $\psi = (\lambda,\alpha,\beta)$, and the parameter space for ψ by Ω. We assume $\Omega = [0,1]\times\Omega_\alpha\times\Omega_\beta$, where Ω_α and $\Omega_\beta \subset R$ are compact intervals. The pseudolikelihood ratio test statistic (PLRT) for H_0 based on $l_p(\psi)$ can be constructed as

$$\text{PLRT} = -2\left\{l_p(0,\alpha,\beta) - l_p\left(\hat{\psi}\right)\right\},$$

where $\hat{\psi} = (\hat{\lambda},\hat{\alpha},\hat{\beta})$ maximizes $l_p(\lambda,\alpha,\beta)$ over the parameter space Ω. We note that the estimation of $f(\cdot)$ is avoided in this testing procedure.

3. Large-sample properties

3.1. Irregularity issues

Although $f(\cdot)$ is eliminated in the pseudolikelihood, there still exist the following three irregularity problems in the pseudolikelihood-based test:

(N1). *(Boundary problem)*. The mixture proportion $\lambda = 0$ lies on the boundary of the parameter space. As a consequence, the asymptotic distribution of the PLRT is usually not a chi-square distribution. The general results on asymptotic distributions of the PLRT under boundary conditions for parametric models were developed by Self & Liang (1987) and Chen & Liang (2010).

(N2). *(Non-identifiability problem)*.] The parameters α and β are non-identifiable when $\lambda = 0$, and similarly, λ is non-identifiable when $\beta = 0$. Because a subset of parameters disappear under the null, the test statistic is often not chi-square distributed asymptotically. Such a non-identifiability issue has been considered in parametric mixture models by Chen & Chen (2001) and Zhu & Zhang (2004), among others.

(N3). *(Singular sensitivity matrix problem)*.] The negative expected Hessian matrix of the pseudolikelihood is often called sensitivity matrix in the composite likelihood literature (Varin *et al.*, 2011). We find that the sub-sensitivity matrix with respect to (α,β) under $\beta = 0$,

$$-E\left\{\frac{\partial^2 l_p(\lambda,\alpha,0)}{\partial(\alpha,\beta)^2};\lambda,\alpha,0\right\} = \begin{pmatrix} 0 & 0 \\ 0 & \frac{n_1 n_0 \sigma^2 \lambda^2 \exp(2\alpha)}{2n\{1-\lambda+\lambda\exp(\alpha)\}^2} \end{pmatrix},$$

is singular, where $\sigma^2 = \text{var}(x_i)$. Thus, the whole sensitivity matrix is singular. Equivalently, the score function for α under $\beta = 0$ is always 0; that is, $\partial l_p(\lambda,\alpha,0)/\partial\alpha = 0$, for any α and λ. To the best of our knowledge, this singularity problem in pseudolikelihood-based inference has not been studied.

To handle problems (N1) and (N2), we partition the parameter spaces into four separate regions where asymptotic expansion in each region can be studied separately. For (N3), we treat α as a nuisance parameter and estimate the remaining parameters (i.e. λ and β) based on the profile pseudolikelihood.

3.2. Asymptotic properties under the null

To establish the consistency and statistical rate of convergence, we introduce $K(\lambda,\beta) = \lambda\beta$ as an index of non-homogeneity. Specifically, $K(\lambda,\beta) = 0$ is equivalent to $\lambda = 0$ or $\beta = 0$, which corresponds to the null hypothesis of homogeneity. On the other hand, a non-zero value of $K(\lambda,\beta)$ suggests non-homogeneity. A similar measure of non-homogeneity is also proposed by

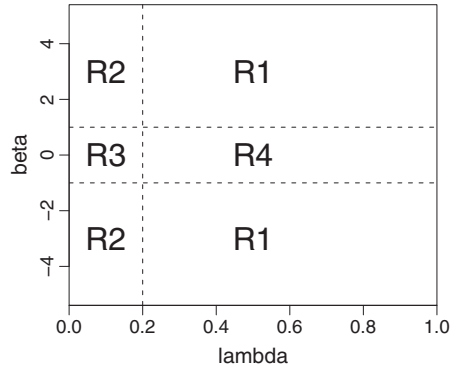

Fig. 1. Illustration of regions R1, R2, R3 and R4, where the vertical dashed line represents $\lambda = \delta_1$ and the horizontal dashed lines represent $\beta = \delta_2$ and $\beta = -\delta_2$.

Zhu & Zhang (2004). Although there can be other indices of non-homogeneity (e.g. $\lambda\beta^2$) that also correspond to the null hypothesis, we define $K(\lambda, \beta) = \lambda\beta$ for simplicity and transparency.

Lemma 1. *Under the regularity conditions (C1)–(C3) in the Appendix and the null hypothesis* H_0, $K(\hat{\lambda}, \hat{\beta}) \to 0$ *in probability.*

The proof is given in the supporting information. To establish the rate of convergence for the maximum pseudolikelihood estimator $\hat{\psi}$ and the asymptotic distribution of PLRT, we must partition the parameter space Ω into four regions: R1, $[\delta_1, 1] \times \Omega_\alpha \times \{\Omega_\beta / [-\delta_2, \delta_2]\}$; R2, $[0, \delta_1] \times \Omega_\alpha \times \{\Omega_\beta / [-\delta_2, \delta_2]\}$; R3, $[0, \delta_1] \times \Omega_\alpha \times [-\delta_2, \delta_2]$; and R4, $[\delta_1, 1] \times \Omega_\alpha \times [-\delta_2, \delta_2]$, where δ_1 and δ_2 are some small positive numbers. Such a partition of the parameter space Ω is to allow different asymptotic expansions in different regions. An illustration of this partition is shown in Figure 1, where for simplicity, we ignore the axis for α. The following theorems establish the rate of convergence for $K(\hat{\lambda}, \hat{\beta})$ and the asymptotic distribution of PLRT.

Theorem 1. *Under the regularity conditions (C1)–(C3) in the Appendix and the null hypothesis* H_0, *we have* $K(\hat{\lambda}, \hat{\beta}) = O_p(n^{-1/2})$.

Theorem 2. *Under the regularity conditions (C1)–(C3) in the Appendix and the null hypothesis* H_0, *the PLRT converges in distribution to*

$$\max\left[\sup_{\beta}\left\{W^+(\beta)\right\}^2, W^2(0)\right],$$

where $W^+(\beta) = W(\beta) I(W(\beta) > 0)$ *and* $W(\beta)$ *is the limiting process of* $W_n(\beta)$ *and*

$$W_n(\beta) = \frac{\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}\{\exp(\beta y_j) - \exp(\beta x_i)\}}{\left[n\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}\{\exp(2\beta x_i) + 2\exp\{\beta(y_j + x_i)\} - 3\exp(2\beta y_j)\}\right]^{1/2}},$$

if $\beta \neq 0$ *and* $W_n(0) = \lim_{\beta\to 0} W_n(\beta)$. *The process* $W(\beta)$ *is a Gaussian process with mean 0, variance* $\lim_{n\to\infty} var\{W_n(\beta)\}$ *and correlation* $\lim_{n\to\infty} cor\{W_n(\beta_1), W_n(\beta_2)\}$.

The proofs of theorems 1 and 2 are given in the supporting information.

Remark 3.1. Because of the non-identifiability problems described in (N2), in general, the asymptotic distribution of PLRT is a functional of Gaussian processes possibly depending on all three parameters α, β and λ. Interestingly, theorem 2 shows that the limiting process only depends on β rather than λ and α.

Remark 3.2. Although the asymptotic distribution of PLRT depends on β only, it is still rather complicated. Therefore, we propose the following bootstrap procedure to obtain p-values.

(i) Denote by B the number of bootstrap replicates. Generate the b-th bootstrap samples by randomly drawing $\{x_1^{(b)}, \ldots, x_{n_0}^{(b)}\}$ and $\{y_1^{(b)}, \ldots, y_{n_1}^{(b)}\}$ with replacement from the pooled sample $\{x_1, \ldots, x_{n_0}, y_1, \ldots, y_{n_1}\}$.
(ii) For each bootstrap sample $\{x_1^{(b)}, \ldots, x_{n_0}^{(b)}\}$ and $\{y_1^{(b)}, \ldots, y_{n_1}^{(b)}\}$, calculate the corresponding $\mathrm{PLRT}^{(b)}$.
(iii) The p-value can be approximated by $\sum_{b=1}^{B} I(\mathrm{PLRT}^{(b)} > \mathrm{PLRT})/B$, where $I(\cdot)$ is an indicator function.

In general, a large number of bootstrap replicates are required to obtain accurate approximated p-values. Such computational burden can be a practical limitation for PLRT in applications such as differential expression detection for microarray data where tens of thousands of genes are tested. In addition, the complicated asymptotic distribution of PLRT loses the appeal of simplicity for conventional likelihood ratio tests.

3.3. Modified pseudolikelihood ratio test and restricted pseudolikelihood ratio test

The complicated limiting distribution of PLRT in theorem 2 makes the use of PLRT difficult in practice. In this section, we further propose two variants of the pseudolikelihood ratio test, both later shown to have a simple asymptotic distribution of χ_1^2. The first test is the MPLRT, defined as

$$\mathrm{MPLRT} = \sup_{\psi \in \Omega} 4\{l_p(\psi) - l_p(0, \alpha, \beta) + C \log \lambda\},$$

where C is a positive constant. The second test is the RPLRT, defined as

$$\mathrm{RPLRT} = \sup_{\psi \in R1 \cup R4} 4\{l_p(\psi) - l_p(0, \alpha, \beta)\}.$$

We note that the likelihood ratios in MPLRT and RPLRT are multiplied by 4 rather than 2 in the conventional likelihood ratio tests. As shown in corollary 1, such a modification is to adjust both MPLRT and RPLRT such that they follow a conventional χ_1^2 distribution asymptotically. The motivation behind the constructions of MPLRT and RPLRT is to avoid the non-identifiability problem (N2). Specifically, the construction of MPLRT adds a smooth penalty function $C\ \lambda$ to force λ away from 0 in a similar fashion to that in Chen *et al.* (2001,2004) for parametric mixture models. The construction of RPLRT directly excludes the region $R2 \cup R3$ (i.e. $\lambda < \delta_1$) in the maximization of pseudolikelihood, similar to that in Lemdani & Pons (1995). The RPLRT can be thought as adding an infinitely large penalty on the pseudolikelihood whenever λ is close enough to 0. For these two tests, the tuning parameter C and the region cut-off δ_1 need to be specified. We now establish the asymptotic properties of MPLRT and RPLRT.

For any fixed C and δ_1, denote $\hat{\psi}_m = (\hat{\lambda}_m, \hat{\alpha}_m, \hat{\beta}_m) = \mathrm{argmax}_{\psi \in \Omega}\{l_p(\psi) + C\ \lambda\}$ and $\hat{\psi}_r = (\hat{\lambda}_r, \hat{\alpha}_r, \hat{\beta}_r) = \mathrm{argmax}_{\psi \in R1 \cup R4} l_p(\psi)$. The following corollary establishes results on the rate of convergence of $(\hat{\lambda}_r, \hat{\beta}_r)$ and $(\hat{\lambda}_m, \hat{\beta}_m)$, and the asymptotic distributions of RPLRT and MPLRT.

Corollary 1. *Under the regularity conditions (C1)–(C3) in the Appendix and the null hypothesis H_0, we have*

(i) $\hat{\beta}_r = O_p(n^{-1/2}), \hat{\lambda}_r = O_p(1), \hat{\beta}_m = O_p(n^{-1/2})$ *and* $\hat{\lambda}_m = O_p(1)$.
(ii) The RPLRT and MPLRT are asymptotically equivalent; both converge weakly to χ^2_1.
(iii) The RPLRT and MPLRT are asymptotically equivalent to the ST proposed by Qin & Liang (2011) and the modified empirical likelihood ratio test (MELRT) proposed by Liu et al. (2012).

The proof is given in the supporting information.

Remark 3.3. We note that the asymptotic distributions of RPLRT and MPLRT are independent of C and δ_1.

3.4. Asymptotic power

The evaluation for power of tests is crucial in the sample size calculation. To this end, the asymptotic local power is often studied. In this section, we consider the power of RPLRT and MPLRT under a sequence of local alternatives. Specifically, for any $0 < \lambda_0 < 1$, density function $f_0(x)$ and $\tau_0 \neq 0$, let

$$H_a : \lambda = \lambda_0, f(x) = f_0(x), \beta = n^{-1/2}\tau_0.$$

Note that under H_a, the local alternative for α can be determined by β and $f(x)$ approximately. Indeed, in the supporting information, we show that $\alpha = -\frac{1}{2}\beta^2\sigma^2 + o(n^{-1})$ under H_a, where $\sigma^2 = \mathrm{var}(x_i)$. Note that α converges to 0 at the rate of $O(n^{-1})$. By LeCam's third lemma (Van der Vaart, 1998), we can establish the following results.

Theorem 3. *Under the regularity conditions (C1)–(C3) in the Appendix, the probability measures of the mixture model under H_0 and H_a are mutually contiguous. Moreover, under the alternatives H_a, the asymptotic distributions of RPLRT and MPLRT are $\chi^2_1\left(\lambda_0^2\tau_0^2(1-\rho)\rho\sigma^2\right)$, where $n_1/n \to \rho$, as $n \to \infty$, and $\chi^2_1(c)$ is a chi-square distribution with one degree of freedom and non-centrality parameter c.*

The proof is given in the supporting information. Because $\lambda_0^2\tau_0^2(1-\rho)\rho\sigma^2 > 0$ under the regularity conditions, RPLRT and MPLRT are asymptotically locally unbiased, which implies that asymptotically, the power of tests is no smaller than the permitted type I error. Moreover, RPLRT and MPLRT have the same asymptotic local power as the ST based on the empirical likelihood (Qin & Liang, 2011) and the MELRT (Liu *et al.*, 2012). While composite likelihood methods may suffer from efficiency loss (Varin *et al.*, 2011), theorem 3 demonstrates that our pairwise pseudolikelihood approach loses no power for testing H_0 locally. From theorem 3, we observe the following intuitive facts: (i) when two groups become further separated (i.e. $\lambda_0^2\tau_0^2$ increases) or (ii) when the design becomes more balanced (i.e. ρ is close to 0.5), RPLRT and MPLRT become more powerful. All these theoretical findings will be verified by simulation studies in Section 5.

4. Goodness-of-fit statistics

Although $f(x)$ is a nuisance function for testing H_0, the estimation of $f(x)$ is sometimes of interest when, for example, we are interested in examining the validity of the exponential

tilt assumption described in (1.2) (Qin 1998, 1999). Let $F(x)$ be the cumulative distribution function of x_i. An estimate of $F(u)$ can be constructed in a similar fashion as in Qin (1998),

$$\tilde{F}(u) = \frac{1}{n}\sum_{h=1}^{n} \frac{1}{1+\rho\tilde{\lambda}\left\{\exp\left(\tilde{\alpha}+\tilde{\beta}t_h\right)-1\right\}} I(t_h \le u),$$

where $(t_1,\ldots,t_n) = (x_1,\ldots,x_{n_0},y_1,\ldots,y_{n_1})$ and $\tilde{\psi} = (\tilde{\lambda},\tilde{\alpha},\tilde{\beta})$ can be $\hat{\psi}$, $\hat{\psi}_r$ or $\hat{\psi}_m$. Similarly, $H(u)$, the cumulative distribution function of y_j, can be estimated by

$$\tilde{H}(u) = \frac{1}{n}\sum_{h=1}^{n} \frac{1-\tilde{\lambda}+\tilde{\lambda}\exp\left(\tilde{\alpha}+\tilde{\beta}t_h\right)}{1+\rho\tilde{\lambda}\left\{\exp\left(\tilde{\alpha}+\tilde{\beta}t_h\right)-1\right\}} I(t_h \le u).$$

To test the exponential tilt assumption, we suggest the following test statistics

$$\Delta_F = \sup_{u\in(-\infty,\infty)} \sqrt{n}\left|\tilde{F}(u)-F_E(u)\right| \quad \text{and} \quad \Delta_H = \sup_{u\in(-\infty,\infty)} \sqrt{n}\left|\tilde{H}(u)-H_E(u)\right|,$$

where F_E and H_E are empirical distribution functions of x and y, defined as

$$F_E(u) = \frac{1}{n_0}\sum_{i=1}^{n_0} I(x_i \le u) \quad \text{and} \quad H_E(u) = \frac{1}{n_1}\sum_{j=1}^{n_1} I(y_j \le u).$$

The asymptotic distributions of Δ_F and Δ_H can be approximated by the following bootstrap procedure.

(i) For $b = 1,\ldots,B$, generate the b-th bootstrap samples $\{x_1^{(b)},\ldots,x_{n_0}^{(b)}\}$ and $\{y_1^{(b)},\ldots,y_{n_1}^{(b)}\}$ drawn from $\tilde{F}(u)$ and $\tilde{H}(u)$, respectively.
(ii) For bootstrap samples $\{x_1^{(b)},\ldots,x_{n_0}^{(b)}\}$ and $\{y_1^{(b)},\ldots,y_{n_1}^{(b)}\}$, we can calculate the test statistics $\Delta_F^{(b)}$ and $\Delta_H^{(b)}$.
(iii) The p-values are approximated by $\sum_{b=1}^{B} I(\Delta_F^{(b)} > \Delta_F)/B$ and $\sum_{b=1}^{B} I(\Delta_H^{(b)} > \Delta_H)/B$, respectively.

A simple Bonferroni approach can be used to combine the p-values from the tests Δ_F and Δ_H.

5. Simulations

To evaluate the finite-sample performance of the proposed tests and compare it with that of current tests, we consider a variety of parametric models. Specifically, for each simulation setting, the data are generated from the mixture model (1.1) with one of the following choices of density functions $f(\cdot)$ and $g(\cdot)$.

Model 1 (Gamma distribution). Let $g(x)$ and $f(x)$ be the density functions of Gamma(m_2,θ) and Gamma(m_1,θ) with shape parameters m_1 and m_2 and scale parameter $\theta > 0$. Then

$$\log\frac{g(x)}{f(x)} = \frac{\Gamma(m_1)}{\Gamma(m_2)} + (m_1-m_2)\log\theta + (m_2-m_1)\log x.$$

Model 2 (Log-normal distribution). Let $g(x)$ and $f(x)$ be the density functions of log-normal distributions $N(\mu_2,\sigma^2)$ and $N(\mu_1,\sigma^2)$. Then

$$\log\frac{g(x)}{f(x)} = \frac{\mu_1^2-\mu_2^2}{2\sigma^2} + \frac{\mu_2-\mu_1}{\sigma^2}\log x.$$

Model 3 (Negative binomial distribution). Let $g(x)$ and $f(x)$ be the density functions of negative binomial distributions NB(r, p_2) and NB(r, p_1), where r is the number of failures until the experiment is stopped and p_1 and p_2 are success probabilities in each experiment. Then

$$\log \frac{g(x)}{f(x)} = r\frac{1-p_2}{1-p_1} + x \log \frac{p_2}{p_1}.$$

Model 4 (t-distribution). Let $g(x)$ and $f(x)$ be the density functions of t-distributions with three degrees of freedom and non-centrality parameters c_1 and c_2.

To generate data under H_0, we take $m_1 = m_2 = 2$ and $\theta = 1$ in model 1, $\mu_1 = \mu_2 = 0$ and $\sigma^2 = 1$ in model 2, $r = 5$ and $p_1 = p_2 = 0.3$ in model 3 and $c_1 = c_2 = 0$ in model 4. Note that models 1–3 belong to the semiparametric exponential tilt model with x replaced by x in models 1 and 2. On the other hand, model 4 does not satisfy the exponential tilt model assumption. It is expected that such a misspecified model may lead to incorrect type I errors for tests based on the ETMM assumption.

Now we briefly describe the empirical likelihood-based tests by Qin & Liang (2011) and Liu *et al.* (2012). In a recent paper, Qin & Liang (2011) suggested a convenient ST defined as $\mathrm{ST} = S(\hat{\alpha}_1, \hat{\beta}_1)/(1 + n_1/n_0)$, where $S(\alpha, \beta) = \sum_{j=1}^{n_1} \{\exp(\alpha + \beta y_j) - 1\}$, $(\hat{\alpha}_1, \hat{\beta}_1) = \mathrm{argmax}_{\alpha,\beta} H(1, \alpha, \beta)$ and

$$H(\psi) = -\sum_{h=1}^{n} \log \left[1 + \{\exp(\alpha + \beta t_h)\} \frac{1}{n} \sum_{j=1}^{n_1} \frac{\lambda \exp(\alpha + \beta y_j)}{1 - \lambda + \lambda \exp(\alpha + \beta y_j)} \right] + \sum_{j=1}^{n_1} \log \{1 - \lambda + \lambda \exp(\alpha + \beta y_j)\},$$

where $\{t_1, \ldots, t_{n_0}, t_{n_0+1}, \ldots, t_n\} = \{x_1, \ldots, x_{n_0}, y_1, \ldots, y_{n_1}\}$ is the pooled sample. Liu *et al.* (2012) proposed a novel MELRT,

$$\mathrm{MELRT} = \sup_{\psi \in \Omega} \{2H(\psi) + 2 \log \lambda\},$$

which is calculated by an expectation–maximization-based method. Both MELRT and ST are convenient to implement and are shown to have a simple asymptotic distribution of χ_1^2 under H_0.

To calculate p-values for PLRT, we use bootstrap to approximate the asymptotic distribution of PLRT. The asymptotic distribution χ_1^2 is used to calculate p-values of MPLRT and RPLRT under H_0. The number of bootstrap replicates is 2000. For MPLRT, a commonly used choice of C is $C = 1$. For RPLRT, we choose $\delta_1 = 0.5$. A sensitivity analysis for the choices of C and δ_1 is provided in the supporting information. In summary, we find that the MPLRT is more robust with respect to the choice of C.

The number of simulations is 4000. Tables 1 and 2 report the type I errors of all tests for models 1 and 2 under various sample sizes. The simulation results under models 3 and 4 are shown in the supporting information. PLRT based on 2000 bootstrap replicates produces type I errors close to the nominal levels. However, the large number of bootstrap replicates requires much more computational time than other asymptotic tests. Among the four asymptotic tests (MPLRT, RPLRT, MELRT and ST), we see that MPLRT is the only test that is reliable under the whole spectrum of models, nominal levels and sample sizes. RPLRT has a similar performance as MELRT at 10% and 5% nominal levels but outperforms MELRT and ST at 1% and 0.5% nominal levels. Even if the sample size is $n_0 = n_1 = 50$, which is a similar sample size setting for the prostate cancer data in Section 6, MELRT and ST still tend to have grossly inflated

Table 1. *Empirical rejection rates (%) of the pseudolikelihood ratio test (PLRT), modified pseudolikelihood ratio test (MPLRT), restricted pseudolikelihood ratio test (RPLRT), modified empirical likelihood ratio test (MELRT) and score test (ST) at 0.1, 0.05, 0.01 and 0.005 significance levels for the Gamma model (model 1)*

(n_0, n_1)	Level (%)	PLRT	MPLRT	RPLRT	MELRT	ST
(20, 20)	10	10.7	10.5	13.5	14.9	13.1
	5	5.4	5.5	7.0	8.1	8.0
	1	1.2	1.0	1.3	2.2	3.1
	0.5	0.6	0.4	0.5	1.3	2.4
(50, 50)	10	10.6	9.4	11.3	11.2	11.0
	5	5.7	4.8	5.9	6.1	5.6
	1	1.2	1.4	1.5	1.8	2.0
	0.5	0.7	0.6	0.9	1.1	1.4
(20, 40)	10	10.6	12.2	14.1	14.8	13.9
	5	5.2	6.1	7.8	8.1	8.5
	1	0.9	0.9	1.4	1.6	2.9
	0.5	0.4	0.4	0.8	0.8	2.0
(20, 80)	10	9.7	12.0	14.9	14.5	13.0
	5	4.7	6.0	8.8	8.6	7.6
	1	0.8	1.3	2.3	2.6	3.4
	0.5	0.5	0.7	1.6	1.8	2.3

Table 2. *Empirical rejection rates (%) of the pseudolikelihood ratio test (PLRT), modified pseudolikelihood ratio test (MPLRT), restricted pseudolikelihood ratio test (RPLRT), modified empirical likelihood ratio test (MELRT) and score test (ST) at 0.1, 0.05, 0.01 and 0.005 significance levels for the -normal model (model 2)*

(n_0, n_1)	Level (%)	PLRT	MPLRT	RPLRT	MELRT	ST
(20, 20)	10	9.7	11.8	13.8	16.7	14.6
	5	4.9	6.1	7.4	9.0	8.8
	1	1.5	1.1	1.7	2.9	4.0
	0.5	0.8	0.7	0.8	1.6	2.9
(50, 50)	10	10.2	10.2	12.5	12.3	12.0
	5	5.4	5.7	6.7	7.4	7.1
	1	1.3	1.1	1.3	1.7	1.8
	0.5	0.4	0.5	0.7	0.9	1.3
(20, 40)	10	10.4	10.4	13.4	13.9	12.9
	5	4.8	5.3	6.8	7.7	7.0
	1	0.7	1.0	1.6	1.8	3.1
	0.5	0.4	0.4	0.6	1.0	2.1
(20, 80)	10	10.3	11.2	14.2	13.9	12.5
	5	5.2	5.7	6.7	8.0	7.2
	1	1.1	1.4	1.8	2.2	2.8
	0.5	0.7	0.7	0.9	1.2	1.9

type I errors, especially at 1% and 0.5% nominal levels. To examine the performance of the tests under model misspecification, we consider the type I errors for the t-distribution model (model 4) in the supporting information. We find that among the four asymptotic tests, MPLRT is the most robust test in the sense that it produces the type I errors closest to the nominal levels

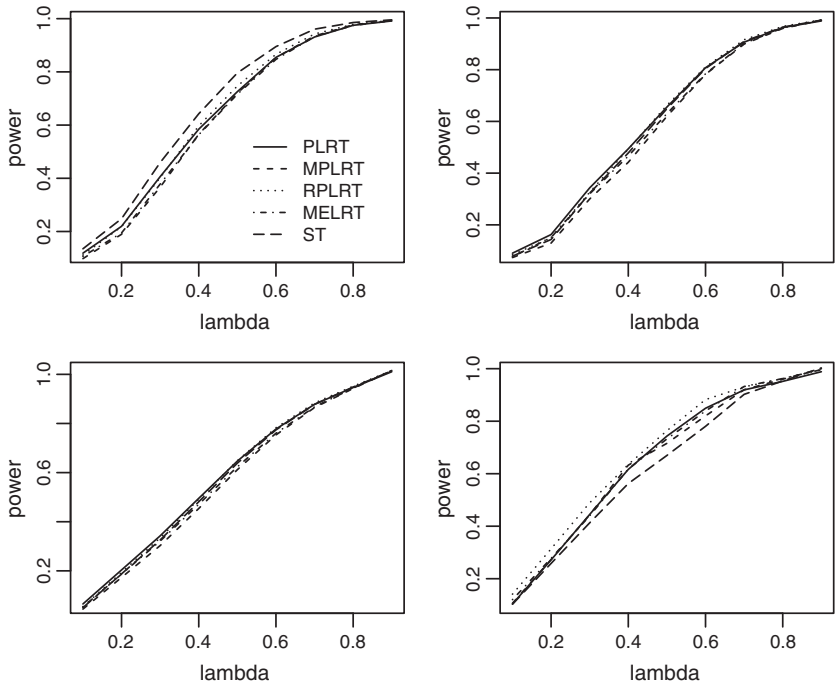

Fig. 2. Powers (level $= 0.05$) of the pseudolikelihood ratio test (PLRT), modified pseudolikelihood ratio test (MPLRT), restricted pseudolikelihood ratio test (RPLRT), modified empirical likelihood ratio test (MELRT) and score test (ST) for models 1–4 as a function of mixture proportion λ when the numbers of observations in two groups are $n_0 = n_1 = 50$. Top left: model 1. Top right: model 2. Bottom left: model 3. Bottom right: model 4.

under the misspecified model. In contrast, the type I errors of MELRT and ST are seriously inflated, especially at small nominal levels.

To compare the powers of the tests, we consider the alternatives where λ takes values from 0 to 1, $m_1 = 1$ and $m_2 = 2$ in model 1, $\mu_1 = 0$ and $\mu_2 = 1$ in model 2, $p_1 = 0.3$ and $p_2 = 0.5$ in model 3 and $c_1 = 0$ and $c_2 = 1.5$ in model 4. To ensure a fair comparison, the critical regions of all tests are adjusted such that the corresponding type I errors are equal to their nominal levels. Figure 2 plots the power curves of all tests for models 1–4 when the design is balanced. The comparison of powers under the unbalanced design is shown in the supporting information. We find that all asymptotic tests share similar powers under all settings, which agrees with corollary 1 and theorem 3. In addition, the tests are more powerful in the balanced design than in the unbalanced design given the same total sample size, which is also consistent with our theoretical findings in theorem 3. Similar results hold for the t-distribution model, while it does not belong to the semiparametric exponential tilt model.

As is common for many statistical methods, the pseudolikelihood is not strictly concave. In practice, the pseudolikelihood may have multiple local maximizers. Based on our simulations, we find that the situations that standard optimization algorithms such as Newton's method fail to converge are very rare. Usually, the algorithm converges to points that are close to the true values.

In summary, our simulation studies comparing type I errors and powers suggest that the class of pseudolikelihood ratio tests perform well in finite samples. Based on our limited investigation of the model not belonging to ETMM, MPLRT seems to be more robust than all other tests.

Given the performance in controlling type I errors, competitive powers and the simplicity of the asymptotic distribution, we recommend the use of MPLRT in practice.

6. Application to prostate cancer data

In many analyses of genomic datasets, differential expression analysis is typically assessed by testing for differences in the mean expression or testing difference in the entire distribution function under two experimental conditions, such as tumour and non-tumour samples (Ghosh & Chinnaiyan, 2009). Although microarray studies of cancers have identified genes differentially expressed in tumour and non-tumour samples, tumour subtypes based on gene expression have not been well appreciated (Lapointe *et al.*, 2004). An important pattern was found by Lapointe *et al.* (2004) and Tomlins *et al.* (2005) that for some genes, only a fraction of samples in one group were overexpressed relative to those in the other group and the remaining samples showed no evidence of differential expression. To model this pattern, Van Wieringen *et al.* (2008) and Qin & Liang (2011) considered a two-sample mixture model. Specifically, at each gene, let x_i be the expression level of the gene of non-tumour group i ($i = 1, 2, \ldots, n_0$) and y_k be the expression level of the same gene of tumour group k ($k = 1, 2, \ldots, n_1$). The data are assumed to follow the ETMM. We apply our tests to the prostate cancer data reported by Lapointe *et al.* (2004). The data consist of the expression level of 5153 genes from 103 tissue samples. Among those 103 samples, 41 were normal, and 62 were cancerous. Under the ETMM assumption, testing for partial differential gene is equivalent to testing for homogeneity in expression levels for normal and cancerous tissues.

We first applied the goodness-of-fit tests Δ_F and Δ_H at each gene. The asymptotic distributions of these tests were approximated by 2000 bootstrap replicates. About 90% of genes in the original data satisfied the exponential tilt model assumption at the significance level of 0.05. Hence, the ETMM may provide an adequate fit to most of the genes in the data. The asymptotic tests, namely, MPLRT, RPLRT, MELRT and ST, were then applied to these genes. The numbers of differentially expressed genes identified by these tests are summarized in Table 3. Interestingly, the conditional tests, MPLRT and RPLRT, identified a similar number of genes, whereas the empirical likelihood-based tests, MELRT and ST, identified a similar number of genes. The empirical likelihood-based tests claimed 29–59% more differentially expressed genes than the conditional tests. Such discrepancy may be explained by what we observed in our simulation studies where the empirical likelihood-based tests tend to be liberal and the type I errors can be grossly larger than the nominal levels, especially when the nominal level is low (e.g. 1% and 0.5%). Thus, the genes identified by the conditional tests may be more reliable than those by the empirical likelihood-based tests. The discussion of the biological meanings of the differentially expressed genes claimed by MPLRT is beyond the scope of the paper. We refer to Lapointe *et al.* (2004) for more details.

Table 3. *Number of differentially expressed genes identified by asymptotic tests, that is, modified pseudolikelihood ratio test (MPLRT), restricted pseudolikelihood ratio test (RPLRT), modified empirical likelihood ratio test (MELRT) and score test (ST), at 0.05, 0.01 and 0.005 significance levels in the Lapointe et al. (2004) data*

Level (%)	MPLRT	RPLRT	MELRT	ST
5	2332	2383	3019	2997
1	1700	1730	2536	2569
0.5	1507	1539	2398	2438

Acknowledgements

This work was partially supported by start-up funds from the University of Texas School of Public Health. The authors want to thank Mr Pan Tong for helping with the preparation of the microarray data.

References

Anderson, J. A. (1979). Multivariate logistic compounds. *Biometrika* **66**, 17–26.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **36**, 192–236.

Chen, H. & Chen, J. (2001). The likelihood ratio test for homogeneity in finite mixture models. *Canad. J. Statist.* **29**, 201–215.

Chen, H., Chen, J. & Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63**, (1), 19–29.

Chen, H., Chen, J. & Kalbfleisch, J. D. (2004). Testing for a finite mixture model with two components. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66**, (1), 95–115.

Chen, Y. & Liang, K. Y. (2010). On the asymptotic behavior of the pseudolikelihood ratio test statistic with boundary problems. *Biometrika* **97**, 603–620.

Cox, D. & Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, (3), 729–737.

Di, C. Z. & Liang, K. Y. (2011). Likelihood ratio testing for admixture models with application to genetic linkage analysis. *Biometrics* **67**, 1249–1259.

Fu, Y., Chen, J. & Kalbfleisch, J. D. (2006). Testing for homogeneity in genetic linkage analysis. *Statist. Sinica* **16**, 805–823.

Ghosh, D. & Chinnaiyan, A. M. (2009). Genomic outlier profile analysis: mixture models, null hypotheses and nonparametric estimation. *Biostat.* **10**, 60–69.

Kalbfleisch, J. D. (1978). Likelihood methods and nonparametric tests. *J. Amer. Statist. Assoc.* **73**, 167–170.

Lancaster, T. & Imbens, G. (1996). Case–control studies with contaminated controls. *J. Econometrics* **71**, 145–160.

Lapointe, J., Li, C., Higgins, J. P., van de Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W. & Bergerheim, U. (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl. Acad. Sci. USA* **101**, 811–816.

Lemdani, M. & Pons, O. (1995). Tests for genetic linkage and homogeneity. *Biometrics* **51**, 1033–1041.

Lemdani, M. & Pons, O. (1997). Likelihood ratio tests for genetic linkage. *Stat. Probabil. Lett.* **33**, 15–22.

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

Liang, K. Y. & Qin, J. (2000). Regression analysis under non-standard situations: a pairwise pseudolikelihood approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62**, 773–786.

Liang, K. Y. & Rathouz, P. J. (1999). Hypothesis testing under mixture models: application to genetic linkage analysis. *Biometrics* **55**, 65–74.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemp. Math.* **80**, (1), 221–39.

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

Liu, Y., Li, P. & Fu, Y. (2012). Testing homogeneity in a semiparametric two-sample problem. *J. Probab. Stat.* **2012**, 15.

Mclachlan, G. J. & Basford, K. E. (1988). *Mixture models: inference and applications to clustering*, Marcel Dekker, New York.

Qin, J. (1998). Inferences for case–control and semiparametric two-sample density ratio models. *Biometrika* **85**, 619–630.

Qin, J. (1999). Empirical likelihood ratio based confidence intervals for mixture proportions. *Ann. Statist.* **27**, 1368–1384.

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

Rathouz, P. J. & Gao, L. (2009). Generalized linear models with unspecified reference distribution. *Biostat.* **10**, 205–218.

Self, S. G. & Liang, K. Y. (1987). Large sample properties of the maximum likelihood estimator and the likelihood ratio test on the boundary of the parameter space. *J. Amer. Statist. Assoc.* **82**, 605–611.

Steinberg, D. & Cardell, N. S. (1992). Estimating logistic regression models when the dependent variable has no variance. *Commun. Stat. A-Theor.* **21**, 423–450.

Tan, Z. (2009). A note on profile likelihood for exponential tilt mixture models. *Biometrika* **96**, 229–236.
Titterington, D. M., Smith, A. F. M. & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*, Wiley, New York.
Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X.-W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., Lee, C., Montie, J. E., Shah, R. B., Pienta, K. J., Rubin, M. A. & Chinnaiyan, A. M. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, (5748), 644–648.
Van der Vaart, A. V. (1998). *Asymptotic statistics*, Cambridge University Press, Cambridge, UK.
Van Wieringen, W. & Van de Viel, M. (2009). Nonparametric testing for dna copy number induced differential MRNA gene expression. *Biometrics* **65**, 19–29.
Van Wieringen, W. N., Van De Wiel, M. A. & Van Der Vaart, A. W. (2008). A test for partial differential expression. *J. Amer. Statist. Assoc.* **103**, (483), 1039–1049.
Varin, C., Reid, N. & Firth, D. (2011). An overview of composite likelihood methods. *Statist. Sinica* **21**, (1), 5–42.
Zhu, H. & Zhang, H. (2004). Hypothesis testing in mixture regression models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66**, 3–16.
Zou, F., Fine, J. P. & Yandell, B. S. (2002). On empirical likelihood for a semiparametric mixture model. *Biometrika* **89**, 61–75.

Received April 2013, in final form July 2014

Yang Ning, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada.
E-mail: y4ning@uwaterloo.ca

Appendix

A.1. Regularity conditions

To study the large-sample properties of PLRT, we assume the following regularity conditions, although the results are expected to hold generally.

(C1) The parameter sets Ω_α and Ω_β for α and β are compact.
(C2) The distributions of x_i and y_j have common support and do not degenerate to a point measure.
(C3) The ratio $n_1/n \to \rho$, as $n \to \infty$, where $0 < \rho < 1$. The variance $\sigma^2 = \mathrm{var}(x_i) < \infty$ and for some $t > C'$, $\int x^2 \exp(t|x|) f(x)\,\mathrm{d}x < \infty$, where C' is a small positive constant.

The compactness of the parameter spaces in assumption (C1) is commonly adopted in the statistical literature. Assumption (C2) is to guarantee that the ETMM is identifiable. Assumption (C3) is required in the pursuit of large-sample properties of pseudolikelihood-based tests when applying the uniform law of large number.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.