J. R. Statist. Soc. B (2004)
66, *Part* 1, *pp.* 3–16

Hypothesis testing in mixture regression models

Hong-Tu Zhu and Heping Zhang

Yale University, New Haven, USA

[Received July 2002. Final revision June 2003]

Summary. We establish asymptotic theory for both the maximum likelihood and the maximum modified likelihood estimators in mixture regression models. Moreover, under specific and reasonable conditions, we show that the optimal convergence rate of $n^{-1/4}$ for estimating the mixing distribution is achievable for both the maximum likelihood and the maximum modified likelihood estimators. We also derive the asymptotic distributions of two log-likelihood ratio test statistics for testing homogeneity and we propose a resampling procedure for approximating the p-value. Simulation studies are conducted to investigate the empirical performance of the two test statistics. Finally, two real data sets are analysed to illustrate the application of our theoretical results.

Keywords: Hypothesis testing; Logistic regression; Mixture regression; Poisson regression; Power

1. Introduction

Finite mixture models arise in many applications, particularly in biology, psychology and genetics. Statistical inference and computation based on these models pose a serious challenge; see Titterington *et al*. (1985), Lindsay (1995) and McLachlan and Peel (2000) for systematic reviews. The purpose of this work is to establish asymptotic theory for mixture regression models originating from those applications.

Suppose that we observe data from n units and within each unit, say unit i, we have n_i measurements, $i = 1, \ldots, n$. This is a typical data structure in longitudinal and family studies. Before introducing the additional notation, let us examine a few examples.

1.1. Example 1: a finite mixture logistic regression model

To study the genetic inheritance pattern of a binary trait such as alcoholism, Zhang and Merikangas (2000) proposed a frailty model in which the data consist of a binary vector response $\mathbf{y}_i = (Y_{i1}, \ldots, Y_{i,n_i})^{\mathrm{T}}$ and covariates X_i from the ith family, $i = 1, \ldots, n$. To model the potential familial correlation, they introduced a Bernoulli latent variable U_i for each family. Conditionally on all latent variables $\{U_i\}$, the Y_{ij}s are assumed to be independent and to follow the logistic regression model

$$\mathrm{logit}\{P(Y_{ij} = 1|U_i)\} = \mathbf{x}_{ij}\beta + \mathbf{z}_{ij}\{U_i\mu_1 + (1 - U_i)\mu_2\}, \tag{1}$$

where $\mathbf{x}_{ij}$ is a covariate vector in X_i from the jth member in the ith family, $\mathbf{z}_{ij}$ is a part of $\mathbf{x}_{ij}$ that interacts with the latent variable and the β and the μs are parameters. The interaction terms

Address for correspondence: Heping Zhang, Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College Street, New Haven, CT 06520-8034, USA.
E-mail: Heping.Zhang@Yale.EDU

 1369–7412/04/66003

in model (1) are very important in genetic studies for assessing potential gene–environment interactions.

Beyond this example, finite mixtures of Bernoulli distributions such as model (1) have received much attention in the last five decades. See Teicher (1963) for an early example. More recently, Wang and Puterman (1998) among others generalized binomial finite mixtures to mixture logistic regression models, and Zhang *et al*. (2003) applied mixture cumulative logistic models to analyse correlated ordinal responses.

1.2. Example 2: mixture of non-linear hierarchical models

Longitudinal and genetic studies commonly involve a continuous response $\{Y_{ij}\}$, also referred to as a quantitative trait. See, for example, Diggle *et al*. (2002), Haseman and Elston (1972) and Risch and Zhang (1995). Pauler and Laird (2000) used general finite mixture non-linear hierarchical models to analyse longitudinal data from heterogeneous subpopulations. Specifically, when there are only two subgroups, the model is of the form

$$Y_{ij} = g\{\mathbf{x}_{ij}, \beta, U_i \mathbf{z}_{ij}\mu_1 + (1 - U_i)\mathbf{z}_{ij}\mu_2\} + \varepsilon_{i,j},$$

where the $\varepsilon_{i,j}$s are independent and identically distributed according to $N(0, \sigma^2)$ and $g(\cdot)$ is a prespecified function. Here, the known covariates $\mathbf{x}_{ij}$ may contain observed time points to reflect a time course in longitudinal data.

1.3. Example 3: a finite mixture of Poisson regression models

Poisson distribution and Poisson regression have been widely used to analyse count data (McCullagh and Nelder, 1989), but observed count data often exhibit overdispersion relative to this. Finite mixture Poisson regression models (Wang *et al.*, 1996) provide a plausible explanation for overdispersion. Specifically, conditionally on all U_is, the Y_{ij}s are independent and follow the Poisson regression model

$$p(Y_{ij} = y_{ij}|\mathbf{x}_{ij}, U_i) = \frac{1}{y_{ij}!}\lambda_{ij}^{y_{ij}} \exp(-\lambda_{ij}), \tag{2}$$

where $\lambda_{ij} = \exp\{\mathbf{x}_{ij}\beta + U_i \mathbf{z}_{ij}\mu_1 + (1 - U_i)\mathbf{z}_{ij}\mu_2\}$.

To summarize the models presented above, we consider a random sample of n independent observations $\{\mathbf{y}_i, X_i\}_1^n$ with the density function

$$p_i(y_i, \mathbf{x}_i; \omega) = \{(1 - \alpha)\ f_i(y_i, \mathbf{x}_i; \beta, \mu_1) + \alpha\ f_i(y_i, \mathbf{x}_i; \beta, \mu_2)\}\ g_i(\mathbf{x}_i), \tag{3}$$

where $g_i(\mathbf{x}_i)$ is the distribution function of X_i. Further, $\omega = (\alpha, \beta, \mu_1, \mu_2)$ is the unknown parameter vector, in which β ($q_1 \times 1$) measures the strength of association that is contributed by the covariate terms and the two $q_2 \times 1$ vectors, μ_1 and μ_2, represent the different contributions from two different groups.

Equivalently, if we consider $P(U_i = 0) = 1 - P(U_i = 1) = \alpha$, and assume that the conditional density of $\mathbf{y}_i$ given U_i is $p_i(y_i|U_i) = f_i\{y_i, \mathbf{x}_i; \beta, \mu_2(1 - U_i) + \mu_1 U_i\}$, then model (3) is the marginal density of $\mathbf{y}_i$. In fact, McCullagh and Nelder (1989) considered a special case in which f_i is from an exponential family distribution, i.e.

$$f_i\{y_i, \mathbf{x}_i; \beta, \mu_2(1 - U_i) + \mu_1 U_i\} = \prod_{j=1}^{n_i} \exp[\phi\{y_{ij}\theta_{ij} - a(\theta_{ij})\} + c(y_{ij}, \phi)], \tag{4}$$

where $\theta_{ij} = h\{\mathbf{x}_{ij}, \beta, U_i\mu_1 + (1 - U_i)\mu_2\}$, $h(\cdot)$ is a link function and ϕ is a dispersion parameter. This family of mixture regression models is very useful in practice.

Asymptotic theory is critical to the understanding of the behaviour of any model. Existing asymptotic results require, however, that f_i in equation (3) is identical among the observation units. This requirement is too restrictive in longitudinal and family studies. Thus, our aim is to eliminate this restriction by allowing f_i to vary between study subjects or families as a result of study designs or missing data.

Let P_* denote the true model from which the data are generated. The well-known identifiability problem in mixture models implies that there may be a set of parameters that yield P_*. We use $\Omega_* = \{\omega_* \in \Omega : P_{\omega_*} = P_*\}$ to represent this set of true model parameters. Here, Ω is the entire parameter space. The use of an asterisk in the subscript reminds us that the parameter belongs to the true parameter set. In the light of the symmetry for α, without loss of generality, we consider only $\alpha \in [0, 0.5]$. The parameter space Ω is defined as $\Omega = \{\omega : \alpha \in [0, 0.5], \beta \in \mathcal{B}, ||\mu_1|| \leqslant M, ||\mu_2|| \leqslant M\}$, where M is a large positive scalar such that $||\mu_*|| < M$, β_* is an interior point of $\mathcal{B}$ and $\mathcal{B}$ is a subset of R^{q_1}.

One of the key hypotheses involving mixture models is that the mixture is warranted. In family studies, it delineates whether the data are familial or not. Formally, this hypothesis can be stated as follows:

$$H_0 : \alpha_*(1-\alpha_*)||\mu_{1*} - \mu_{2*}|| = 0, \quad \textit{versus} \quad H_1 : \alpha_*(1-\alpha_*)||\mu_{1*} - \mu_{2*}|| \neq 0, \tag{5}$$

where $||\cdot||$ is the Euclidean norm of a vector. Whether the null hypothesis or its alternative is true has a critical bearing on asymptotic theory and statistical inference. When the alternative hypothesis is true, the existing asymptotic theory established by Andrews (1999) is applicable. Challenges arise under the null hypothesis, e.g. the derivation of the significance level (p-value).

As discussed by Lemdani and Pons (1999) and Lindsay (1995), there are at least three main difficulties. First, Ω_* is on the boundary of Ω. The lack of identifiability between α and the μs is the second difficulty. Finally, the Fisher information matrix for ω is singular. Recently, major progress has been made for finite mixture models by Bickel and Chernoff (1993), Chen *et al.* (2001), Cheng and Liu (2001), Dacunha-Castelle and Gassiat (1999), Lemdani and Pons (1999), Lindsay (1995) and references therein. It is noteworthy that covariates are absent from their work, and the μs are scalar; moreover, previous results cannot accommodate non-independent or non-identically distributed data. As our examples demonstrate, however, there is a need from practical applications to consider the general model (3). For this, we establish general asymptotic theory in the presence of covariates, nuisance parameters and high dimensional μ-parameters.

The paper is organized as follows. In Section 2, we introduce two likelihood-ratio-based test statistics and present related asymptotic theory. On the basis of these results, we propose a resampling procedure for approximating the p-value. In Section 3, we perform some simulation studies to demonstrate the power of the two test statistics. Two real data sets are also analysed. It should be noted that our assumptions in this paper are not optimal. Some extensions are still possible and warrant future research.

2. Test statistics

The log-likelihood function $L_n(\omega)$ for data $\{y_i, \mathbf{x}_i\}_1^n$ under model (3) is given by

$$L_n(\omega) = \sum_{i=1}^{n} \log\{(1-\alpha)\ f_i(\beta, \mu_1)/f_{i*} + \alpha\ f_i(\beta, \mu_2)/f_{i*}\},$$

where $f_{i*} = f_i(y_i, \mathbf{x}_i; \beta_*, \mu_*)$ and $f_i(\beta, \mu) = f_i(y_i, \mathbf{x}_i; \beta, \mu)$. To test hypothesis (5), we usually start from the log-likelihood-ratio statistic:

$$\mathrm{LR}_n = \sup_{\omega\in\Omega}\{L_n(\omega)\} - \sup_{\omega\in\Omega_0}\{L_n(\omega)\},$$

where $\Omega_0 = \{\omega \in \Omega : \alpha = 0.5, \mu_1 = \mu_2\}$. A challenging issue is to derive the asymptotic distribution of LR_n. To do so, we need to define some notation.

For a generic symmetric $q_2 \times q_2$ matrix B, $\mathrm{Vecs}(B)$ and $\mathrm{dvecs}(B)$ are defined as $\mathrm{Vecs}(B) = (b_{11}, b_{21}, b_{22}, \ldots, b_{q_2 1}, \ldots, b_{q_2 q_2})^{\mathrm{T}}$ and

$$\mathrm{dvecs}(B) = (b_{11}, 2b_{21}, b_{22}, 2b_{31}, 2b_{32}, b_{33}, \ldots, 2b_{q_2 1}, \ldots, 2b_{q_2,\, q_2-1}, b_{q_2 q_2})^{\mathrm{T}}$$

respectively. Furthermore, we define $F_{i,1}(\beta,\mu) = \partial_\beta f_i(\beta,\mu)/f_{i*}$, $F_{i,2}(\mu) = \partial_\mu f_i(\beta_*,\mu)/f_{i*}$ and $F_{i,6}(\mu) = \partial^2_{\mu\mu} f_i(\beta_*,\mu)/f_{i*}$. Let

$$w_i(\mu) = (F_{i,1}(\beta_*,\mu_*)^{\mathrm{T}}, F_{i,2}(\mu_*)^{\mathrm{T}}, \mathrm{dvecs}(F_{i,6}(\mu))^{\mathrm{T}})^{\mathrm{T}},$$

$$W_n(\mu) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i(\mu),$$

$$J_n(\mu) = \frac{1}{n}\sum_{i=1}^{n} w_i(\mu) w_i(\mu)^{\mathrm{T}},$$

where $W_n(\mu)$ is an $r \times 1$ vector and $J_n(\mu)$ is an $r \times r$ matrix. Finally, let $k_1(\omega) = (1-\alpha)\,\Delta\mu_1 + \alpha\,\Delta\mu_2$, $k_2(\omega) = \mathrm{Vecs}\{(1-\alpha)\,\Delta\mu_1^{\otimes 2} + \alpha\,\Delta\mu_2^{\otimes 2}\}$ and $K(\omega) = (\Delta\beta, k_1(\omega), k_2(\omega))$, in which $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^{\mathrm{T}}$, $\Delta\mu_1 = \mu_1 - \mu_*$ and $\Delta\mu_2 = \mu_2 - \mu_*$.

Under Ω_0, $L_n(\omega)$ simplifies to $\Sigma_{i=1}^n \log\{f_i(\beta,\mu)/f_{i*}\}$. Standard asymptotic theory yields that

$$\sup_{\omega\in\Omega_0}\{2\,L_n(\omega)\} = W_n(\mu_*)^{\mathrm{T}} H\{H^{\mathrm{T}}\,J_n(\mu_*)H\}^{-1} H^{\mathrm{T}}\,W_n(\mu_*) + o_p(1),$$

where $H = (I_{q_1+q_2}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ is a $\{q_1 + q_2 + q_2(q_2-1)/2\} \times (q_1+q_2)$ matrix. Furthermore, if we consider the maximum likelihood estimator $\hat{\omega}_M = (\hat{\alpha}_M, \hat{\beta}_M, \hat{\mu}_{1M}, \hat{\mu}_{2M}) = \arg\max_{\omega\in\Omega}\{L_n(\omega)\}$ in Ω, we need to address several fundamental issues such as the consistency and rate of convergence of $K(\hat{\omega}_M)$ and the asymptotic expansion of $L_n(\hat{\omega}_M)$. After some tedious manipulations (see Zhu and Zhang (2003) for details), we find that

$$2\,L_n(\hat{\omega}_M) = \sup_{||\mu_2||\leqslant M} (W_n(\mu_2)^{\mathrm{T}}\,J_n(\mu_2)^{-1}\,W_n(\mu_2) - \inf_{\omega\in\Omega_{\mu_2}}[Q_n\{\sqrt{n}\,K(\omega),\mu_2\}]) + o_p(1)$$

holds under assumptions 1–3 that are given in Appendix A, where $\Omega_{\mu_2} = \{\omega \in \Omega : \mu_2 \text{ is fixed}\}$ and

$$Q_n(\lambda,\mu_2) = (\lambda - J_n(\mu_2)^{-1}\,W_n(\mu_2))^{\mathrm{T}}\,J_n(\mu_2)(\lambda - J_n(\mu_2)^{-1}\,W_n(\mu_2)).$$

Define the convex cone Λ_{μ_2} as

$$\Lambda_{\mu_2} = \left\{\eta : \eta = \begin{pmatrix} \Delta\mu_2 & I_{q_2} \\ \mathrm{Vecs}(\Delta\mu_2^{\otimes 2}) & \mathbf{0} \end{pmatrix}\mathbf{x}^{\mathrm{T}}, \mathbf{x} = (x_1,\ldots,x_{q_2+1}) \in [0,\infty)\times R^{q_2}\right\}.$$

Then, the log-likelihood-ratio statistic becomes

$$\mathrm{LR}_n = \sup_{||\mu_2||\leqslant M}\{\hat{V}(\mu_2)^{\mathrm{T}}\,J_n(\mu_2)\,\hat{V}(\mu_2)\} - W_n(\mu_*)^{\mathrm{T}} H\{H^{\mathrm{T}}\,J_n(\mu_*)H\}^{-1}H^{\mathrm{T}}\,W_n(\mu_*) + o_p(1), \quad (6)$$

where $Q_n\{\hat{V}(\mu_2),\mu_*\} = \inf_{\lambda\in R^{q_1}\times\Lambda_{\mu_2}}\{Q_n(\lambda,\mu_*)\}$.

We obtain the following theorems, whose detailed proofs are given in Zhu and Zhang (2003).

Theorem 1. Under assumptions 1–3 in Appendix A, the following results hold.

(a) $K(\hat{\omega}_M) = O_p(n^{-1/2})$.
(b) The log-likelihood-ratio statistic converges to the maxima of a stochastic process that is indexed by μ_2 in distribution as $n \to \infty$.
(c) If $q_2 = 1$, $\int_\mu |\hat{G}_M(\mu) - G(\mu)|\, \mathrm{d}\mu = O_p(n^{-1/4})$, where $G(\mu) = I(\mu \geqslant \mu_*)$ is the true mixing distribution and $\hat{G}_M(\mu) = G(\hat{\mu}_{1M}, \hat{\mu}_{2M}, \hat{\alpha}_M) = (1 - \hat{\alpha}_M)\, I(\mu \geqslant \hat{\mu}_{1M}) + \hat{\alpha}_M\, I(\mu \geqslant \hat{\mu}_{2M})$.

To our knowledge, theorem 1, part (a), provides the convergence rate of $K(\hat{\omega}_M)$ for the first time under model (3). Our result has several implications. First, we confirm that the convergence rate of $\hat{\beta}$ (the maximum likelihood estimator) is $n^{-1/2}$ under the conditions defined. van der Vaart (1996) proved a similar result under semiparametric mixture models. Second, under hypothesis H_0, we prove that only $k_1(\hat{\omega}_M)$ and $k_2(\hat{\omega}_M)$ can reach the rate $n^{-1/2}$, which is useful for determining the asymptotic distributions of the estimators. Third, theorem 1, part (c), implies that $\hat{G}_M(\mu)$ converges to $G(\mu)$ in L_1-norm at the optimal rate $n^{-1/4}$ under a more general model than that considered by Chen (1995). It is important to note that theorem 1 of Chen (1995) gives the lower bound for the optimal rate of convergence for estimating $G(\mu)$ by using $\hat{G}_M(\mu)$ in L_1-norm is at most $n^{-1/4}$.

As in Chen *et al.* (2001), we consider an alternative approach to testing hypothesis (5) by using a modified log-likelihood function

$$\mathrm{ML}_n(\omega) = L_n(\omega) + \log(M) \log\{4\alpha(1-\alpha)\},$$

where M is as defined above. Compared with the log-likelihood function, the extra term $\log(M)\log\{4\alpha(1-\alpha)\}$ in $\mathrm{ML}_n(\omega)$ can keep α away from both 0 and 1, which partially solves the identifiability problem. Let $\hat{\omega}_P$ be the resulting estimator as $\hat{\omega}_P = \arg\max_{\omega\in\Omega}\{\mathrm{ML}_n(\omega)\}$. The modified log-likelihood-ratio statistic is defined as

$$\mathrm{MLR}_n = 2\,\mathrm{ML}_n(\hat{\omega}_P) - \sup_{\omega\in\Omega_0}\{2\,\mathrm{ML}_n(\omega)\}.$$

Similarly to the log-likelihood-ratio statistic, we find that

$$\mathrm{MLR}_n = W_n(\mu_*)^{\mathrm{T}}[J_n(\mu_*)^{-1} - H\{H^{\mathrm{T}} J_n(\mu_*) H\}^{-1} H^{\mathrm{T}}]\, W_n(\mu_*) - \inf_{\lambda\in R^{q_1}\times\Lambda_0}\{Q_n(\lambda, \mu_*)\} + o_p(1), \tag{7}$$

where Λ_0 is a closed cone defined by

$$\Lambda_0 = \{\eta : \eta = (\mathbf{x}^{\mathrm{T}}, \mathrm{Vecs}(\mathbf{y}^{\otimes 2}))^{\mathrm{T}},\ \text{both } \mathbf{x} \text{ and } \mathbf{y} \in R^{q_2}\}.$$

We have the following asymptotic result for both $\hat{\omega}_P$ and MLR_n.

Theorem 2. Under the assumptions of theorem 1, we have the following results.

(a) $\hat{\alpha}_P = O_p(1), \Delta\hat{\beta}_P = O_p(n^{-1/2}), \Delta\hat{\mu}_{1P} = O_p(n^{-1/4})$ and $\Delta\hat{\mu}_{2P} = O_p(n^{-1/4})$.
(b) The modified log-likelihood-ratio statistic converges to a random variable U_2 in distribution as $n \to \infty$.
(c) If $q_2 = 1$, $\int_\mu |\hat{G}_P(\mu) - G(\mu)|\, \mathrm{d}\mu = O_p(n^{-1/4})$, where $\hat{G}_P(\mu) = (1 - \hat{\alpha}_P)\, I(\mu \geqslant \hat{\mu}_{1P}) + \hat{\alpha}_P\, I(\mu \geqslant \hat{\mu}_{2P})$.

Theorem 2, part (a), gives the exact convergence rate of $\hat{\omega}_P$. Theorem 2, part (b), determines the asymptotic distribution of MLR_n. Whereas the explicit form of this distribution is generally complicated, our result gives rise to a simple asymptotic distribution, $0.5\chi_1^2 + 0.5\chi_0^2$, for

MLR_n when $q_2 = 1$. This coincides with theorem 1 of Chen *et al.* (2001) when there are no covariates, i.e. $q_1 = 0$. Theorem 2, part (c), shows that the $n^{-1/4}$ consistent rate for estimating the mixing distribution $G(\mu)$ is reachable by using $\hat{\omega}_P$.

Until now, we have systematically investigated the order of convergence rate for both $K(\hat{\omega}_P)$ and $K(\hat{\omega}_M)$, and the asymptotic distributions of both LR_n and MLR_n. The next two issues are also very important. The first is how to compute the critical values for these potentially complicated distributions of test statistics. The second is to compare the power of LR_n and MLR_n. In what follows, we study the empirical and asymptotic behaviour of these two statistics under departures from hypothesis H_0.

2.1. A resampling method

Although we have obtained the asymptotic distributions of the likelihood-based statistics, the limiting distributions usually have complicated analytic forms. To alleviate this difficulty, we use a resampling technique to calculate the critical values of the testing statistics. Although the bootstrapping method is an obvious approach, it requires repeated maximizations of the likelihood and modified likelihood functions. The maximizations are computationally intensive for the finite mixture models. Thus, we prefer a computationally more efficient method, as proposed and used by Hansen (1996), Kosorok (2003) and others.

On the basis of equations (6) and (7), we only need to focus on $W_n(\mu_2)$ and $J_n(\mu_2)$. If hypothesis H_0 is true, $(\hat{\beta}^0, \hat{\mu}^0) = \arg\max_{\omega \in \Omega_0}\{L_n(\omega)\}$ provides consistent estimators of β_* and μ_*. By substituting (β_*, μ_*) with $(\hat{\beta}^0, \hat{\mu}^0)$ in the definitions of $W_n(\mu_2)$ and $J_n(\mu_2)$, we obtain $\hat{w}_i(\mu_2)$, $\hat{W}_n(\mu_2)$ and $\hat{J}_n(\mu_2)$ accordingly, i.e.

$$\hat{w}_i(\mu_2) = (F_{i,1}(\hat{\beta}^0, \hat{\mu}^0)^{\mathrm{T}}, F_{i,2}(\hat{\beta}^0, \hat{\mu}^0), \mathrm{dvecs}(F_{i,6}(\hat{\beta}^0, \mu_2))^{\mathrm{T}})^{\mathrm{T}},$$

$$\hat{W}_n(\mu_2) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{w}_i(\mu_2),$$

$$\hat{J}_n(\mu_2) = \frac{1}{n} \sum_{i=1}^{n} \hat{w}_i(\mu_2)\, \hat{w}_i(\mu_2)^{\mathrm{T}}.$$

To assess the significance level and the power of the two test statistics, we need to obtain empirical distributions for these statistics in lieu of their theoretical distributions. What follow are the four key steps in generating the stochastic processes that have the same asymptotic distributions as the test statistics.

Step 1: we generate independent and identically distributed random samples, $\{v_{i,m} : i = 1, \ldots, n\}$, from the standard normal distribution $N(0, 1)$. Here, m represents a replication number.

Step 2: we calculate

$$\hat{W}_n^m(\mu_2) = \sum_{i=1}^{n} \hat{w}_i(\mu_2) v_{i,m} / \sqrt{n}$$

and

$$Q_n^m(\lambda, \mu_2) = (\lambda - \hat{J}_n^{-1}(\mu_2)\, W_n^m(\mu_2))^{\mathrm{T}}\, \hat{J}_n(\mu_2)(\lambda - \hat{J}_n^{-1}(\mu_2)\, W_n^m(\mu_2)).$$

It is important to note that $\hat{W}_n^m(\mu_2)$ converges weakly to $W(\mu_2)$ as $n \to \infty$. This claim can be proved by using the conditional central limit theorem; see theorem (10.2) of Pollard (1990) and theorem 2 of Hansen (1996).

Step 3: the third step is to calculate the likelihood ratio

$$\mathrm{LR}_n^m = \sup_{||\mu_2|| \leqslant M} \{\hat{V}^m(\mu_2)^{\mathrm{T}} \hat{J}_n(\mu_2)\, \hat{V}^m(\mu_2)\} - \hat{W}_n^m(\hat{\mu}^0)^{\mathrm{T}} H\{H^{\mathrm{T}} \hat{J}_n(\hat{\mu}^0) H\}^{-1} H^{\mathrm{T}}\, \hat{W}_n^m(\hat{\mu}^0),$$

and

$$\mathrm{MLR}_n^m = \hat{W}_n^m(\hat{\mu}^0)^{\mathrm{T}}[J_n(\mu^0)^{-1} - H\{H^{\mathrm{T}} J_n(\mu^0) H\}^{-1} H^{\mathrm{T}}]\, \hat{W}_n^m(\hat{\mu}^0) - \inf_{\lambda \in R^{q_1} \times \Lambda_0} \{Q_n^m(\lambda, \hat{\mu}^0)\},$$

where $Q_n^m\{\hat{V}^m(\mu_2), \mu_2\} = \inf_{\lambda \in R^{q_1} \times \Lambda_{\mu_2}} \{Q_n^m(\lambda, \mu_2)\}$.
Step 4: we repeat the above three steps J times and obtain two realizations—$\{\mathrm{LR}_n^m : m = 1, \ldots, J\}$ and $\{\mathrm{MLR}_n^m : m = 1, \ldots, J\}$. It can be shown that the empirical distribution of LR_n^m converges to the asymptotic distribution of LR_n, and likewise for MLR_n^m. Therefore, the empirical distributions of these two realizations form the basis for calculation of the critical values in hypothesis testing as well as the power analysis.

We approximate $p_1 = \Pr(U_1 > \mathrm{LR}_n)$ and $p_2 = \Pr(U_2 > \mathrm{MLR}_n)$ by using

$$p_1 \approx \hat{p}_1^J = \sum_{m=1}^{J} I(\mathrm{LR}_n^m > \mathrm{LR}_n)/J,$$

$$p_2 \approx \hat{p}_2^J = \sum_{m=1}^{J} I(\mathrm{MLR}_n^m > \mathrm{MLR}_n)/J.$$

Since J is under our control, $\hat{p}_1^J$ can be made arbitrarily close to p_1 by using a sufficiently large J. To guide the selection of J in practice, it is useful to note that $(\hat{p}_1^J - p_1)\sqrt{J}$ converges to a normal distribution with mean 0 and variance $p_1(1 - p_1)$ in distribution as $J \to \infty$ and that $\hat{p}_1^J$ almost lies in $(p_1 - 2.5\sqrt{\{p_1(1 - p_1)/J\}}, p_1 + 2.5\sqrt{\{p_1(1 - p_1)/J\}})$; see also Hansen (1996). Thus, if we want to set an error control δ_0 for $|\hat{p}_1^J - p_1|$, we can show that $2.5\sqrt{\{p_1(1 - p_1)/J\}} \leqslant \delta_0$. For example, when $p_1 = 0.01$, choosing $J = 10000$ yields an error of about 0.0025. Exactly the same formula applies to p_2 as to p_1.

2.2. Asymptotic local power

An assessment of power is necessary for choosing a reasonable sample size as well as appropriate and powerful tests of significance (Cox and Hinkley (1974), page 103). Often, there is no closed form for the power calculation. Many researchers have considered alternatives by exploiting the asymptotic local power. In our case, the distribution of $J_n^{-1}(\mu_2)\, W_n(\mu_2)$ plays a critical role in determining the asymptotic local power of LR_n and MLR_n; see equations (6) and (7). Thus, we explore its properties under a sequence of local alternatives.

Consider a sequence of local alternatives ω^n consisting of $\alpha^n = \alpha_0, \beta^n = \beta_* + n^{-1/2}\mathbf{h}_1, \mu_1^n = \mu_* - n^{-1/4}\{\alpha_0/(1 - \alpha_0)\}^{0.5}\mathbf{h}_2$ and $\mu_2^n = \mu_* + n^{-1/4}\{(1 - \alpha_0)/\alpha_0\}^{0.5}\mathbf{h}_2$, where α_0 is a constant between 0 and 1, and $\mathbf{h}_1$ and $\mathbf{h}_2$ are $q_1 \times 1$ and $q_2 \times 1$ vectors respectively. At ω^n, $K(\omega^n) = n^{-1/2}\mathbf{h}$, where $\mathbf{h}^{\mathrm{T}} = (\mathbf{h}_1^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}}, \mathrm{Vecs}(\mathbf{h}_2^{\otimes 2})^{\mathrm{T}})$. We have the following theorem.

Theorem 3. Under assumptions 1–3 in Appendix A and the alternatives ω^n,

$$J_n(\mu_2)^{-1}\, W_n(\mu_2) \xrightarrow{\mathrm{d}} N\{J(\mu_2)^{-1}\, J(\mu_2, \mu_*)\mathbf{h}, J(\mu_2)^{-1}\}.$$

Theorem 3 can be used to compare the local power of our testing statistics. In particular, $J_n(\mu_*)^{-1}\, W_n(\mu_*)$ converges to $N\{\mathbf{h}, J(\mu_*)^{-1}\}$ in distribution. For instance, the asymptotic power function that is associated with the modified log-likelihood-ratio statistic is

$$p_\alpha(\mathbf{h}) = \lim_{n\to\infty}[P\{F^0(\mathrm{MLR}_n) \geqslant 1-\alpha|\mathbf{h}\}],$$

where $F^0(\cdot)$ is the asymptotic distribution of MLR_n under hypothesis H_0 as $n \to \infty$. As $||\mathbf{h}|| \to \infty$, the $p_\alpha(\mathbf{h})$ converges to 1, since MLR_n becomes large as $||\mathbf{h}||$ increases.

3. Simulation study and real examples

Two computational issues are related to our test procedures. First, we must calculate three different estimators: $(\hat\beta^0, \hat\mu^0)$, $\hat\omega_M$ and $\hat\omega_P$. It is relatively straightforward to compute $(\hat\beta^0, \hat\mu^0)$ and $\hat\omega_P$ by using Newton–Raphson and/or EM algorithms. Since finding a global maximizer is still an open problem, we employ the EM algorithm using 200 different starting-points. During the computation, we use $M = 10$ to obtain $\hat\omega_P$.

Secondly, to approximate the critical values of the asymptotic distributions, we need to obtain two realizations: $\{\mathrm{MLR}_n^m : m = 1,\ldots,J\}$ and $\{\mathrm{LR}_n^m : m = 1,\ldots,J\}$. Obtaining the latter is actually a computationally intensive burden even if q_2 is as small as 2. Hence, it is sensible to consider replacing $B(\mathbf{0}, M) = \{\mu; ||\mu|| \leqslant M\}$ with a discrete approximation $B(\mathbf{0}, M)_A$. Then, LR_n^m is approximated by

$$\max_{\mu_2\in B(\mathbf{0},M)_A}\{\hat V^m(\mu_2)^{\mathrm{T}}\,\hat J_n(\mu_2)\,\hat V^m(\mu_2)\} - \hat W_n^{m\mathrm{T}}(\hat\mu^0)H\{H^{\mathrm{T}}\,\hat J_n(\hat\mu^0)H\}^{-1}H^{\mathrm{T}}\,\hat W_n^m(\hat\mu^0).$$

Owing to the computational complexity, we only consider LR_n when $q_2 = 1$ and $M = 10$, i.e. $B(\mathbf{0}, 10) = [-10, 10]$. We take $B(0,10)_A = \{-10, -10 + 0.02, \ldots, 10 - 0.02, 10\}$.

3.1. Simulated data

In this subsection, we conduct some simulation studies to assess the performance of the two tests that were introduced in Section 2. Data are drawn from the mixture linear regression model

$$y_{ij} = z_{ij}\{U_i\tilde\mu_1 + (1-U_i)\tilde\mu_2\} + \varepsilon_{ij},$$

where $\tilde\mu_1 = \beta + \mu_1$, $\tilde\mu_2 = \beta + \mu_2$ and $\varepsilon_{ij} \sim N(0,1)$. From now on, we use $(\tilde\mu_1, \tilde\mu_2)$ instead of (β, μ_1, μ_2) to avoid the identifiability problem.

We consider two cases: $q_2 = 1$ and $q_2 = 2$. For $q_2 = 1$, all the z_{ij}s are equal to 1. For $q_2 = 2$, $z_{ij} = (1, u_{ij})$, where u_{ij} comes from the uniform [0, 2] generator. Therefore, we have parameters $\tilde\mu_1$, $\tilde\mu_2$, α and σ. The true value of σ is 1 and the number of observations in each cluster is set equal to 3. It should be noted that σ is the true nuisance parameter in the general model (3), implying that $q_1 = 1$.

For $q_2 = 1$, four different settings of $(\alpha, \tilde\mu_1, \tilde\mu_2)$, denoted A1, B1, C1 and D1, are considered. Similarly, for $q_2 = 2$, four other different settings of $(\alpha, \tilde\mu_1, \tilde\mu_2)$, denoted A2, B2, C2 and D2, are considered. See Table 1 for all eight designs. We choose sample sizes $n = 50$ and $n = 100$.

Table 1. Eight designs for the simulation study

Unknown parameter	*Results for the following designs:*							
	A1	*B1*	*C1*	*D1*	*A2*	*B2*	*C2*	*D2*
α	0.5	0.5	0.5	0.2	0.5	0.5	0.5	0.2
μ_1	1	1	1	1	$\mathbf{1}_2$	$\mathbf{1}_2$	$\mathbf{1}_2$	$\mathbf{1}_2$
μ_2	1	1.4	2	2	$\mathbf{1}_2$	$1.4 \times \mathbf{1}_2$	$1.707 \times \mathbf{1}_2$	$1.707 \times \mathbf{1}_2$

For each simulation, the three significance levels 10%, 5% and 1% are considered, and 50000 replications are used to estimate nominal significance levels (or rejection rates). To calculate standard errors for the rejection rate of each case, we use 50000 replications to form 50 consecutive 1000 replications, which give 50 estimates of the rejection rate and the standard error of these 50 estimates.

Table 2 presents the estimates and standard errors of the rejection rates for the log-likelihood-ratio testing statistic. As expected, the power increases as does the distance between $\tilde{\mu}_1$ and $\tilde{\mu}_2$. Moreover, the simulated critical values are again precise and the null rejection level is close to the true rejection level.

Tables 3 and 4 report the estimates and standard errors of the rejection rates for the modified log-likelihood-ratio statistic. The proposed testing procedure based on MLR_n performs remarkably well. First, the power increases fast as the distance between $\tilde{\mu}_1$ and $\tilde{\mu}_2$ becomes large. Second, the simulated critical values are quite accurate. In particular, the null rejection level is quite close to the true rejection level. Third, as expected, a larger sample size produces better results.

Table 2. Estimates and standard errors of rejection rates for the log-likelihood-ratio statistic of a one-component mixture against a two-component mixture

True rate		*Results ($q_2 = 1$) for the following designs and values of n:*							
		A1		*B1*		*C1*		*D1*	
		n = 50	*n = 100*	*n = 50*	*n = 100*	*n = 50*	*n = 100*	*n = 50*	*n = 100*
0.010	Estimate	0.0151	0.0146	0.0399	0.0528	0.4943	0.8177	0.2886	0.5422
	Standard error	0.0046	0.0044	0.0070	0.0065	0.0153	0.0139	0.0148	0.0182
0.050	Estimate	0.0596	0.0576	0.1253	0.1592	0.7218	0.9352	0.5077	0.7535
	Standard error	0.0079	0.0075	0.0102	0.0105	0.0112	0.0071	0.0168	0.0145
0.100	Estimate	0.1072	0.1069	0.2031	0.2540	0.8192	0.9669	0.6281	0.8404
	Standard error	0.0094	0.0094	0.0141	0.0132	0.0114	0.0054	0.0139	0.0128

Table 3. Estimates and standard errors of rejection rates for the modified log-likelihood-ratio statistic of a one-component mixture against a two-component mixture

True rate		*Results ($q_2 = 1$) for the following designs and values of n:*							
		A1		*B1*		*C1*		*D1*	
		n = 50	*n = 100*	*n = 50*	*n = 100*	*n = 50*	*n = 100*	*n = 50*	*n = 100*
0.0100	Estimate	0.0151	0.0121	0.0404	0.0572	0.5601	0.8732	0.2886	0.5439
	Standard error	0.0041	0.0038	0.0068	0.0082	0.0139	0.0086	0.0122	0.0165
0.0500	Estimate	0.0539	0.0520	0.1264	0.1659	0.7779	0.9607	0.5053	0.7525
	Standard error	0.0079	0.0063	0.0101	0.0141	0.0131	0.0060	0.0132	0.0145
0.1000	Estimate	0.0981	0.0980	0.2075	0.2628	0.8641	0.9821	0.6270	0.8385
	Standard error	0.0102	0.0094	0.0096	0.0163	0.0099	0.0040	0.0150	0.0115

Table 4. Estimates and standard errors of rejection rates for the modified log-likelihood-ratio statistic of a one-component mixture against a two-component mixture

True rate		*Results ($q_2 = 2$) for the following designs and values of n:*							
		A2		*B2*		*C2*		*D2*	
		n = 50	*n = 100*	*n = 50*	*n = 100*	*n = 50*	*n = 100*	*n = 50*	*n = 100*
0.0100	Estimate	0.0166	0.0148	0.2330	0.4843	0.9613	0.9998	0.7165	0.9611
	Standard error	0.0046	0.0029	0.0121	0.0165	0.0056	0.0005	0.0159	0.0070
0.0500	Estimate	0.0636	0.0585	0.4459	0.7081	0.9905	0.9999	0.8502	0.9866
	Standard error	0.0078	0.0073	0.0141	0.0138	0.0031	0.0002	0.0113	0.0037
0.1000	Estimate	0.1143	0.1078	0.5686	0.8061	0.9960	1.0000	0.9001	0.9923
	Standard error	0.0103	0.0102	0.0143	0.0108	0.0019	0.0001	0.0087	0.0024

By inspecting Tables 2–4, we find that the log-likelihood-ratio statistic does not have any advantage over the modified log-likelihood-ratio statistic. In fact, the opposite seems to be so. Although we do not have a full explanation yet, the computational complexity could be one of the reasons. Thus, on the basis of our simulation studies, we prefer to use the modified log-likelihood-ratio statistic to test hypothesis (5) because of its easier computation and better power.

3.2. Ames salmonella assay data

We reanalyse an assay data set that has been studied by Wang *et al.* (1996) among others. The date set contains the number of revertant colonies of salmonella under different dose levels of quinoline. At each of six dose levels of quinoline d_i, three plates are tested. To fit the data set, Wang *et al.* (1996) chose the two-component Poisson regression (2) with $n = 18$, $n_i = 1$ and Poisson rates

$$\lambda_i = \exp[x_i^{\mathrm{T}}\beta + \{\mu_1 U_i + \mu_2(1 - U_i)\}],$$

where $x_i = (d_i, \log(d_i + 10))$, $i = 1, \ldots, 18$. They used the Akaike information criterion and Bayes information criterion to facilitate the model selection process, but they did not have a testing procedure to test hypothesis (5).

We use the modified log-likelihood-ratio statistic to test the hypothesis as stated in expression (5). First, we use the EM algorithm to obtain $(\hat{\alpha}_P, \hat{\beta}_P, \hat{\mu}_{1P}, \hat{\mu}_{2P})^{\mathrm{T}} = (0.723, -0.0013, 0.378, 1.844, 2.382)^{\mathrm{T}}$. Next, using $\hat{\omega}_P$ and the results in Table 6 of Wang *et al.* (1996), we find that $\mathrm{MLR}_n = 12.447$. Then, the approximation procedure is repeated $J = 10000$ times, leading to $\hat{p}^J = 0.0001$. Since $q_2 = 1$ in this case, the true asymptotic distribution of MLR_n is $0.5\chi_0^2 + 0.5\chi_1^2$, giving a p-value of 0.0002. Therefore, the data support the mixture Poisson regression with two components.

We can also use the log-likelihood-ratio statistic to test the same hypothesis. It follows from Table 6 of Wang *et al.* (1996) that the log-likelihood-ratio statistic $\mathrm{LR}_n = 14.44$. Although it is difficult to compute the true p-value, the approximation procedure gives $\hat{p}^{10000} = 0.0004$. Further analyses such as residual analysis and the goodness of fit have been reported in Wang *et al.* (1996).

3.3. A risk factor analysis of preterm delivery

Although the survival rate has improved for preterm infants that are born with low (less than 2500 g) or even very low birth weights (less than 1000 g), preterm birth remains a major cause of neurodevelopmental disability. Many epidemiological studies have been conducted to examine risk factors for preterm deliveries. Here, we reanalyse a data set that has been collected and extensively analysed by Dr Michael Bracken and his colleagues (see, for example, Zhang and Bracken (1995)). The data for this study consisted of 3858 women whose pregnancies ended in a singleton live-birth at Yale–New Haven Hospital, Connecticut, in 1980–1982. Preterm delivery is defined as less than 37 weeks of gestational age, which is calculated from the first day of the last menstrual period. On the basis of Zhang and Bracken (1995), we examine the following eight putative risk factors: total number of pregnancies (v_1), ethnic background (v_2), use of marijuana (v_3), marital status (v_4), hormones or diethylstilbestrol used by the mother (v_5), parity (v_6), years of education (v_7) and total alcohol use per day (v_8). Among these risk factors, v_2, v_3, v_4 and v_5 are dummy variables with values 0 or 1. For instance, v_2 is an indicator for whether a woman is black or not. Zhang and Bracken (1995) originally considered 15 variables, but seven of them do not contribute to the risk of preterm births and hence are not considered here.

To understand the potential relationship between the preterm delivery and above risk factors, we fit the logistic regression model to the data set and report the output in Table 5. According to the estimators and their corresponding standard errors in Table 5, the most significant risk factors are the number of previous pregnancies, ethnicity and the years of education, as expected from earlier studies (Zhang and Bracken, 1995). However, the remaining risk factors do not appear to be significant at the 0.05 level.

Next, we use the logistic mixture regression

$$\text{logit}\{P(Y_i = 1|U_i)\} = \mathbf{x}_i\beta + \mathbf{z}_i\{U_i\mu_1 + (1 - U_i)\mu_2\},$$

where $\mathbf{x}_i = (1, v_{1i}, v_{2i}, v_{3i})$ and $\mathbf{z}_i = (v_{4i}, v_{5i}, v_{6i}, v_{7i}, v_{8i})$ for $i = 1, \ldots, 3858$. The result becomes intriguing when the two-component logistic regression model is used to analyse this data set; see Table 5 for details. First, all eight putative risk factors under the logistic mixture model become significant. Second, when we use the modified log-likelihood-ratio statistic to test the hypothesis as stated in expression (5), $\text{MLR}_n = 20.810$, giving rise to $\hat{p}^J = 0.0002$ with 10000 (i.e. J) repeated samples. This means that our procedure strongly favours the two-component logistic model, and there is a significant disparity between the null hypothesis H_0 in expression

Table 5. Modified log-likelihood estimators for the preterm data set†

α	*Intercept*	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
One-component mixture (modified log-likelihood −778.49)									
	−2.465	0.167	0.577	0.232	−0.109	0.131	−0.086	−0.061	−0.039
	(0.463)	(0.077)	(0.210)	(0.223)	(0.218)	(0.170)	(0.110)	(0.033)	(0.073)
Two-component mixture‡ (modified log-likelihood −768.09)									
0.509	−2.949	0.174	0.664	0.241	1.062	0.778	0.276	−0.159	0.218
(0.101)	(0.116)	(0.021)	(0.059)	(0.062)	(0.133)	(0.145)	(0.094)	(0.019)	(0.039)
					−1.361	−0.997	−0.990	0.094	−0.308
					(0.166)	(0.095)	(0.052)	(0.018)	(0.044)

†Standard errors of the estimators are given in parentheses; $\text{MLR}_n = 20.810$; $\hat{p}^{10000} = 0.0002$.
‡α, the intercept and v_1–v_3 are shared, and v_4–v_8 are separated for $U = 0$ (upper values) and $U = 1$ (lower values).

(5) and the observed sample. Third, by looking at the coefficients of v_4–v_8, we find that the estimators in the two latent groups have opposite signs, explaining why they are insignificant in the ordinary logistic model. This underscores the importance of the two-component model. The opposite signs of the estimators indicate reversed relationships between these risk factors and preterm births in the two latent groups.

Acknowledgements

This research is supported in part by National Institutes of Health grants DA12468 and AA12044. We thank the Joint Editor, an Associate Editor and two referees for valuable suggestions which helped to improve our presentation greatly. We also thank Professor Michael Bracken for the use of his data set.

Appendix A

Before we present all the assumptions, we need to define the following derivatives of $f_i(\beta, \mu)$: $F_{i,3}(\beta, \mu) = \partial_\beta^2 f_i(\beta, \mu)/f_{i*}$, $F_{i,4}(\mu) = \partial_\beta \partial_\mu f_i(\beta_*, \mu)/f_{i*}$, $F_{i,5}(\mu) = \partial_\beta \partial_\mu^2 f_i(\beta_*, \mu)/f_{i*}$ and $F_{i,7}(\mu) = \partial_\mu^3 f_i(\beta_*, \mu)/f_{i*}$. The following assumptions are sufficient conditions to derive our asymptotic results. Detailed proofs of theorems 1–3 can be found in Zhu and Zhang (2003).

A.1. Assumption 1 (identifiability)

As $n \to \infty$, $\sup_{\omega \in \Omega}\{n^{-1}|L_n(\omega) - \bar{L}_n(\omega)|\} \to 0$ in probability, where $\bar{L}_n(\omega) = E\{L_n(\omega)\}$. For every $\delta > 0$, we have

$$\liminf_{n\to\infty} (n^{-1}[\bar{L}_n(\bar{\omega}_n) - \sup_{\omega \in \Omega/\Omega_{*,\delta}} \{\bar{L}_n(\omega)\}]) > 0,$$

where $\bar{\omega}_n$ is the maximizer of $\bar{L}_n(\omega)$ and

$$\Omega_{*,\delta} = \{\omega : ||\beta - \beta_*|| \leqslant \delta, ||\mu_1 - \mu_*|| \leqslant \delta, \alpha||\mu_2 - \mu_*|| \leqslant \delta\} \cap \Omega$$

for every $\delta > 0$.

A.2. Assumption 2

For a small $\delta_0 > 0$, let $\mathbf{B}_{\delta_0} = \{(\beta, \mu) : ||\beta - \beta_*|| \leqslant \delta_0 \text{ and } ||\mu|| \leqslant M\} \cap \Omega$,

$$\sup_{(\beta,\mu)\in \mathbf{B}_{\delta_0}} \left\{ \left|\left| \frac{1}{n} \sum_{i=1}^n F_{i,1}(\beta, \mu) \right|\right| + \left|\left| \frac{1}{n} \sum_{i=1}^n F_{i,3}(\beta, \mu) \right|\right| + \left|\left| \frac{1}{n} \sum_{i=1}^n F_{i,2}(\mu) \right|\right| + \sum_{k=4}^6 \left|\left| \frac{1}{n} \sum_{i=1}^n F_{i,k}(\mu) \right|\right| \right\} = o_p(1),$$

$$\sup_{||\mu|| \leqslant M} \left\{ \left|\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n F_{i,k}(\mu) \right|\right| \right\} = O_p(1), \qquad k = 4, 6, 7,$$

$$\sup_{(\beta,\mu)\in \mathbf{B}_{\delta_0}} \left\{ \frac{1}{n} \sum_{i=1}^n ||F_{i,1}(\beta, \mu)||^3 + ||F_{i,3}(\beta, \mu)||^3 + ||F_{i,2}(\mu)||^3 + \sum_{k=4}^7 ||F_{i,k}(\mu)||^3 \right\} = O_p(1).$$

Moreover, $\max_{1\leqslant i\leqslant n} \sup_{(\beta,\mu)\in \mathbf{B}_{\delta_0}} \{||F_{i,1}(\beta, \mu)|| + ||F_{i,2}(\mu)||^2\} = o_p(n^{1/2})$.

A.3. Assumption 3

$(W_n(\cdot), J_n(\cdot)) \Rightarrow (W(\cdot), J(\cdot))$, where these processes are indexed by $||\mu|| \leqslant M$, and the stochastic process $\{(W(\mu), J(\mu)) : ||\mu|| \leqslant M\}$ has bounded continuous sample paths with probability 1. Each $J(\mu)$ is a symmetric matrix and $\infty > \sup_{||\mu|| \leqslant M}[\lambda_{\max}\{J(\mu)\}] \geqslant \inf_{||\mu|| \leqslant M}[\lambda_{\min}\{J(\mu)\}] > 0$ holds almost surely.

The process $W(\mu)$ is a mean vector R^r-valued Gaussian stochastic process $\{W(\mu) : ||\mu|| \leqslant M\}$ such that $E\{W(\mu)\ W(\mu)^{\mathrm{T}}\} = J(\mu)$ and $E\{W(\mu)\ W(\mu')^{\mathrm{T}}\} = J(\mu, \mu')$ for any μ and μ' in $B(\mathbf{0}, M)$.

A.4. Remarks

Assumption 1 is a generalized definition of the identifiable uniqueness in the statistical literature. The assumption $\sup_{\omega\in\Omega}\{n^{-1}|L_n(\omega) - \bar{L}_n(\omega)|\} \to^{\mathrm{p}} 0$ is the uniform laws of large numbers. Some sufficient conditions for uniform laws of large numbers have been presented in the literature; see Pollard (1990). Moreover, assumption 2 is quite reasonable in most situations. In assumption 3, to prove that $W_n(\mu)$ weakly converges to a Gaussian process $W(\mu)$, we need to use the functional central limit theorem; see Pollard (1990). Moreover, assumption 3 assumes the positive definitiveness of $J(\mu)$, which is a generalization of the *strong identifiability conditions* that were used in Chen (1995) and Chen and Chen (2001). Under the identical and independent distribution framework, assumption (P0) of Dacunha-Castelle and Gassiat (1999) is a sufficient condition for assumptions 1 and 2. Moreover, Dacunha-Castelle and Gassiat's (1999) assumption (P1) is just the positive definitiveness of $J(\mu)$ in assumption 3.

References

Andrews, D. W. K. (1999) Estimation when a parameter is on a boundary: theory and applications. *Econometrica*, **67**, 1341–1383.

Bickel, P. and Chernoff, H. (1993) Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem. In *Statistics and Probability: a Raghu Raj Bahadur Festschrift* (eds J. K. Ghosh, S. K. Mitra, K. R. Parthasarathy and B. L. S. Prakasa Rao), pp. 83–96. New Delhi: Wiley.

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

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

Chen, J. (1995) Optimal rate of convergence for finite mixture models. *Ann. Statist.*, **23**, 221–233.

Cheng, R. C. H. and Liu, W. B. (2001) The consistency of estimators in finite mixture models. *Scand. J. Statist.*, **28**, 603–616.

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics*. London: Chapman and Hall.

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

Diggle, P. J., Liang, K. Y. and Zeger, S. L. (2002) *Analysis of Longitudinal Data*, 2nd edn. New York: Oxford University Press.

Hansen, B. E. (1996) Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, **64**, 413–430.

Haseman, J. K. and Elston, R. C. (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.*, **2**, 3–19.

Kosorok, M. R. (2003) Bootstraps of sums of independent but not identically distributed stochastic processes. *J. Multiv. Anal.*, **84**, 299–318.

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

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

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

McLachlan, G. and Peel, D. (2000) *Finite Mixture Models*. New York: Wiley.

Pauler, D. K. and Laird, N. M. (2000) A mixture model for longitudinal data with application to assessment of noncompliance. *Biometrics*, **56**, 464–472.

Pollard, D. (1990) *Empirical Processes: Theory and Applications*. Hayward: Institute of Mathematical Statistics.

Risch, N. and Zhang, H. P. (1995) Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science*, **268**, 1584–1589.

Teicher, H. (1963) Identifiability of finite mixtures. *Ann. Math. Statist.*, **34**, 1265–1269.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985) *The Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

van der Vaart, A. (1996) Efficient estimation in semiparametric models. *Ann. Statist.*, **24**, 862–878.

Wang, P. M. and Puterman, M. L. (1998) Mixed logistic regression models. *J. Agric. Biol. Environ. Statist.*, **3**, 175–200.

Wang, P. M., Puterman, M. L., Cockburn, I. and Le, N. (1996) Mixed Poisson regression models with covariate dependent rates. *Biometrics*, **52**, 381–400.

Zhang, H. P. and Bracken, M. (1995) Tree-based risk factor analysis of preterm delivery and small-for-gestational-age birth. *Am. J. Epidem.*, **141**, 70–78.

Zhang, H. P., Fui, R. and Zhu, H. T. (2003) A latent variable model of segregation analysis for ordinal outcome. *J. Am. Statist. Ass.*, to be published.

Zhang, H. P. and Merikangas, K. (2000) A frailty model of segregation analysis: understanding the familial transmission of alcoholism. *Biometrics*, **56**, 815–823.

Zhu, H. T. and Zhang, H. P. (2003) Hypothesis testing for finite mixture regression models (mathematical details). *Technical Report*. Yale University School of Medicine, New Haven.