Hypothesis Testing in Mixture Regression Models (Mathematical Details)

Hongtu Zhu and Heping Zhang

Yale University School of Medicine

Summary

As a technical supplement to Zhu and Zhang (2004), we give detailed information on how to establish asymptotic theory for both maximum likelihood estimate and maximum modified likelihood estimate in mixture regression models. Under specific and reasonable conditions, we show that the optimal convergence rate of $n^{-\frac{1}{4}}$ for estimating the mixing distribution is achievable for both the maximum likelihood and maximum modified likelihood estimates. We also derive the asymptotic distributions of the two log-likelihood ratio testing statistics for testing homogeneity.

1 Notation and Assumptions

We consider a random sample of n independent observations $\{\mathbf{y}_i, X_i\}_1^n$ with the following density function

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

where $g_i(\mathbf{x}_i)$ is the distribution function of X_i, $\omega = (\alpha, \beta, \mu_1, \mu_2)$ is the unknown parameter vector, in which β ($q_1 \times 1$) measures the strength of association 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. The log-likelihood function $L_n(\omega)$ is given by

$$L_n(\omega) = \sum_{i=1}^{n} \log[(1-\alpha) f_i(\beta, \mu_1)/f_{i*} + \alpha f_i(\beta, \mu_2)/f_{i*}], \tag{2}$$

where $f_{i*} = f_i(y_i, \mathbf{x}_i; \beta_*, \mu_*)$ and $f_i(y_i, \mathbf{x}_i; \beta, \mu) = f_i(\beta, \mu_1)$.

In light of the symmetry for α, without loss of generality, we only consider $\alpha \in [0, 0.5]$ only. Define the parametric space Ω as

$$\begin{aligned} \Omega &= \{\omega : \alpha \in [0, 0.5], \beta \in \mathcal{B}, ||\mu_1|| \leq M, ||\mu_2|| \leq M\} \\ &= [0, 0.5] \times \mathcal{B} \times B(\mathbf{0}, M) \times B(\mathbf{0}, M), \end{aligned} \tag{3}$$

where M is a large positive scalar such that $||\mu_*|| < M$, $B(\mathbf{0}, M)$ is a ball in R^{q_2} centered at $\mathbf{0}$ with radius M, and $\mathcal{B}$ is a subset of R^{q_1}.

One of the key hypotheses involving mixture models is whether the mixture regression is warranted. In family studies, it means whether or not the trait of interest is familial. This hypothesis can be stated as follows:

$$H_0: \quad \alpha_*(1-\alpha_*)||\mu_{1*} - \mu_{2*}|| = 0, \quad \text{v.s.} \quad H_1: \alpha_*(1-\alpha_*)||\mu_{1*} - \mu_{2*}|| \neq 0, \tag{4}$$

where $||\cdot||$ is the Euclidean norm of a vector.

Let Δ denote the incremental operator such as $\Delta\beta = \beta - \beta_*$. Define

$$(\Delta\beta)^T F_{i,1}(\beta, \mu_1) = \frac{f_i(\beta, \mu_1) - f_i(\beta_*, \mu_1)}{f_i(\beta_*, \mu_*)} \quad \text{and} \quad (\Delta\mu_1)^T F_{i,2}(\mu_1) = \frac{f_i(\beta_*, \mu_1) - f_i(\beta_*, \mu_*)}{f_i(\beta_*, \mu_*)}.$$

When $f_i(\beta, \mu_1)$ has a continuous derivative with respect to β, $F_{i,1}(\beta, \mu_1)$ is simply the first derivative.

Our results also require higher order increments. To avoid unnecessarily complicated matrix and tensor-product notations, we only define the higher order increments for a scalar μ_1. The notations can be defined by analogue for a vector μ_1. Define

$$\begin{aligned}
F_{i,3}(\beta, \mu_1)\Delta\beta &= F_{i,1}(\beta, \mu_1) - F_{i,1}(\beta_*, \mu_1), \\
F_{i,4}(\mu_1)\Delta\mu_1 &= F_{i,1}(\beta_*, \mu_1) - F_{i,1}(\beta_*, \mu_*), \\
F_{i,5}(\mu_1)\Delta\mu_1 &= F_{i,4}(\mu_1) - F_{i,4}(\mu_*), \\
F_{i,6}(\mu_1)(\Delta\mu_1) &= F_{i,2}(\mu_1) - F_{i,2}(\mu_*), \\
F_{i,7}(\mu_1)\Delta\mu_1 &= F_{i,6}(\mu_1) - F_{i,6}(\mu_*).
\end{aligned}$$

If all $f_i(\beta, \mu)$'s have partial derivatives up to third order with respect to (β, μ), these functions $\{F_{i,k} : k = 1, \cdots, 7\}$ are simple the corresponding partial derivatives.

Next, for any μ, we define the general form of $w_i(\mu)$; however, it should be noted that $F_{i,6}(\mu)$ is a symmetric matrix in this case. Let

$$\begin{aligned}
w_i(\mu) &= (F_{i,1}(\beta_*, \mu_*)^T, F_{i,2}(\mu_*)^T, \text{dvecs}[F_{i,6}(\mu)]^T)^T, \\
W_n(\mu) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i(\mu), \quad J_n(\mu) = \frac{1}{n} \sum_{i=1}^{n} w_i(\mu) w_i(\mu)^T,
\end{aligned}$$

where $W_n(\mu)$ is a $r \times 1$ vector and $J_n(\mu)$ is a $r \times r$ matrix, and for any symmetric $q_2 \times q_2$ matrix B, we define

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

Finally, let

$$\begin{aligned} k_1(\omega) &= (1-\alpha)(\Delta\mu_1) + \alpha(\Delta\mu_2), \\ k_2(\omega) &= \text{Vecs}[(1-\alpha)(\Delta\mu_1)^{\otimes 2} + \alpha(\Delta\mu_2)^{\otimes 2}], \\ K(\omega) &= (\Delta\beta, k_1(\omega), k_2(\omega)), \end{aligned}$$

in which $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$, $\Delta\mu_1 = \mu_1 - \mu_*$, $\Delta\mu_2 = \mu_2 - \mu_*$, and

$$\text{Vecs}(B) = (b_{11}, b_{21}, b_{22}, \cdots, b_{q_2 1}, \cdots, b_{q_2 q_2})^T$$

for a generic symmetric $q_2 \times q_2$ matrix B.

The following assumptions are sufficient conditions to derive our asymptotic results.

Assumptions:

(1.1) The sets $n^{1/2}(\mathcal{B} - \beta_*)/b_n$ can be locally approximated by a convex cone Λ_{β_*}, where $b_n \to \infty$ and $b_n n^{-1/2} < \infty$, and $\mathcal{B} - \beta_* = \{\beta - \beta_* : \beta \in \mathcal{B}\}$.

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

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

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

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

for every $\delta > 0$.

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

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

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

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

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

(1.4) $(W_n(\cdot), J_n(\cdot)) \Rightarrow (W(\cdot), J(\cdot))$, where these processes are indexed by $||\mu|| \le M$, and the stochastic process $\{(W(\mu), J(\mu)) : ||\mu|| \le M\}$ has bounded continuous sample paths with probability one. Each $J(\mu)$ is a symmetric matrix and $\infty > \sup_{||\mu||\le M}\lambda_{\max}[J(\mu)] \ge \inf_{||\mu||\le M}\lambda_{\min}[J(\mu)] > 0$ holds almost surely. The process $W(\mu)$ is a mean-vector R^r-valued Gaussian stochastic process $\{W(\mu) : ||\mu|| \le M\}$ such that $E[W(\mu)W(\mu)^T] = J(\mu)$ and $E[W(\mu)W(\mu')^T] = J(\mu,\mu')$ for any μ and μ' in $B(\mathbf{0}, M)$.

Comments on (1.1): Assumption (1.1) is related to the following definition of Andrews (1999), who generalized a definition of Chernoff (1954). Let's introduce two geometrical concepts: a convex cone and the distance between a point and a set. A set $\Lambda \subset R^{q_1}$ is called a *convex cone* if "$\lambda \in \Lambda$" implies "$a\lambda \in \Lambda$ for all $a > 0$ and "$\lambda_1, \lambda_2 \in \Lambda$" implies "$a_1\lambda_1 + (1-a_1)\lambda_2$" for all $a_1 \in [0,1]$. The distance between a point $z \in R^{q_1}$ and a set $\Lambda \subset R^{q_1}$ is defined by $\text{dist}(z,\Lambda) = \inf_{\lambda\in\Lambda}||z-\lambda||$. A sequence of sets $\{n^{1/2}(\mathcal{B}-\beta_*)/b_n \subset R^{q_1} : n \ge 1\}$ are called locally approximated by a cone Λ_{β_*} provided that

$$\text{dist}(k_n(\beta), \Lambda_{\beta_*}) = o(||k_n(\beta)||) \;\text{ for any }\; k_n(\beta) \in n^{1/2}(\mathcal{B}-\beta_*)/b_n \text{ such that } \; ||k_n(\beta)|| \to 0,$$

and

$$\text{dist}(\lambda_n, n^{1/2}(\mathcal{B}-\beta_*)/b_n) = o(||\lambda_n||) \;\text{ for any }\; \lambda_n \in \Lambda_{\beta_*} \;\text{ such that }\; ||\lambda_n|| \to 0.$$

Detailed discussions about this assumption can be found in Subsection 4.3 of Andrews (1999).

Comments on (1.2): Assumption (1.2) is a generalized definition of the identifiable uniqueness; see Definition 3.1 of Pötscher and Prucha (1991). Assumption $\sup_{\omega\in\Omega} n^{-1}|L_n(\omega) - \overline{L}_n(\omega)| \to^p 0$ is the uniform laws of large numbers (LLN). Some sufficient conditions for the uniform LLN have been presented in the literature; see Andrews (1992) and Pollard (1990).

Comments on (1.3): Assumption (1.3) generalizes Assumption (P0) of Dacunha-Castelle and Gassiat (1999). The major difference lies in that Assumption (1.3) only requires that all $f_i(\beta,\mu)$'s possess partial derivatives up to order 3. Especially, if $\{\mathbf{y}_i, X_i\}_1^n$ are independent and identical observations, a simple sufficient condition for Assumption (1.3) is that

$$\sup_{\omega\in\Omega} ||\partial_\omega^2 f_i(y_i, X_i;\omega)/f_{i*}(y_i, X_i)|| \le m(y_i, X_i) \quad , \int [m^3(y_i, X_i) f_{i*}(y_i, X_i) g_i(X_i)] dy_i dX_i < +\infty.$$

We prefer Assumption (1.3) since it is the minimum requirement to prove Theorem 1. The existence and measurability of $\{F_{i,k} : k = 1, \cdots, 7\}$ and corresponding supreme functions might be handled by imposing some measurable condition on all $\{F_{i,k}, f_i(\beta,\mu) : i = 1, \cdots, n; k = 1, \cdots, 7\}$ and using results in Appendix C of Pollard (1984). Other conditions on $\{F_{i,k} : k = 1, \cdots, 7\}$ are quite reasonable. For instance, Theorem 1 of Andrews (1992) can be used to validate that

$$\sup_{(\beta,\mu)\in\mathbf{B}_{\delta_0}} ||\frac{1}{n}\sum_{i=1}^n F_{i,3}(\beta,\mu)|| \to 0 \text{ in probability.}$$

Comments on (1.4): Assumption (1.4) is a direct generalization of Assumption (P1) of Dacunha-Castelle and Gassiat (1999) and the *strong identifiability conditions* used in Chen (1995) and Chen and Chen (2001). Some discussions of this assumption in some explicit examples can be found in those papers. A sufficient condition for $\sup_{|\mu|\le M} ||J_n(\mu) - J(\mu)|| \to^p 0$ is given by following:

$$\sup_{|\mu|\le M} ||J_n(\mu) - EJ_n(\mu)|| \to^p 0 \text{ and } \lim_{n\to\infty} \sup_{|\mu|\le M} ||J(\mu) - EJ_n(\mu)|| = 0.$$

For example, we can use the uniform LLN results to prove $\sup_{|\mu|\le M} ||J_n(\mu) - EJ_n(\mu)|| \to^p 0$. 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).

2 Maximum Likelihood Estimate and Log-likelihood Ratio Statistic

In this section, we mainly present the asymptotic theory for the maximum likelihood estimate under the model (1) and the log-likelihood ratio statistic for testing the hypothesis (4) if H_0 is true.

Consider the maximum likelihood estimate $\hat{\omega}_M = (\hat{\alpha}_M, \hat{\beta}_M, \hat{\mu}_{1M}, \hat{\mu}_{2M}) = \text{argmax}_{\omega \in \Omega} L_n(\omega)$. The key idea of our approach is that we first establish a quadratic approximation for $L_n(\omega)$, which is the cornerstone of the asymptotics; see Andrews (1999) and Zhu and Zhang (2002). We will show that

$$2L_n(\omega) = 2K(\omega)W_n(\mu_2) - K(\omega)^T J_n(\mu_2)K(\omega) + o_p(1) \tag{5}$$

holds uniformly in $\{\omega \in \Omega : ||K(\omega)|| \leq C_0/\sqrt{n}\}$ for some constant C_0. By using this approximation of $L_n(\omega)$, we have the following asymptotic result.

THEOREM 1. *Under Assumptions (1.1)-(1.4), the following results hold:*

(i) $K(\hat{\omega}_M) = O_p(n^{-1/2})$.

(ii)

$$\begin{aligned} 2L_n(\hat{\omega}_M) &= \sup_{||\mu_2|| \leq M} \left\{ W_n(\mu_2)^T J_n(\mu_2)^{-1} W_n(\mu_2) - \inf_{\omega \in \Omega_{\mu_2}} Q_n(\sqrt{n}K(\omega), \mu_2) \right\} + o_p(1) \\ &\to^d \sup_{||\mu_2|| \leq M} V^{0T}(\mu_2) J(\mu_2) V^0(\mu_2), \end{aligned} \tag{6}$$

where $\Omega_{\mu_2} = \{\omega \in \Omega : \mu_2 \quad \textit{is fixed}\}$,

$$Q_n(\sqrt{n}K(\omega), \mu_2) = [\sqrt{n}K(\omega) - J_n(\mu_2)^{-1}W_n(\mu_2)]^T J_n(\mu_2)[\sqrt{n}K(\omega) - J_n(\mu_2)^{-1}W_n(\mu_2)],$$

and $V^0(\mu_2)$ *is defined in the proof of Theorem 1.*

To our knowledge, Theorem 1 (i) provides the convergence rate of $K(\hat{\omega}_M)$ for the first time under model (1). In the literature, the common effort is to derive the convergence rate for the estimate, $\hat{\omega}_M$, itself. Our result has several implications. First, we confirm that the convergence rate of $\hat{\beta}$ (maximum likelihood estimate) is $n^{-1/2}$ under the defined conditions. Van der Vaart (1996) proved a similar result under semiparametric mixture models. Second, under H_0, we prove that only $k_1(\hat{\omega}_M)$ and $k_2(\hat{\omega}_M)$ can reach the rate of $n^{-1/2}$, which is useful for determining the asymptotic distributions of the estimates.

To illustrate the usefulness of Theorem 1, let us explain how Theorem 1(i) strengthens and generalizes a theorem in Chen (1995). Consider the true mixing distribution $G(\mu) = I\{\mu \geq \mu_*\}$ in the case of $q_2 = 1$. Using $\hat{\omega}_M$, we construct $\hat{G}_M(\mu) = G(\hat{\mu}_{1M}, \hat{\mu}_{2M}, \hat{\alpha}_M) = (1 - \hat{\alpha}_M)I\{\mu \geq \hat{\mu}_{1M}\} + \hat{\alpha}_M I\{\mu \geq \hat{\mu}_{2M}\}$. Theorem 1 of Chen (1995) states that the optimal rate convergence for estimating $G(\mu)$ by using $\hat{G}_M(\mu)$ in L_1 norm is at most $n^{-1/4}$. However, it does not imply the exact rate for the L_1 distance between $\hat{G}_M(\mu)$ and $G(\mu)$, $\int_\mu |\hat{G}_M(\mu) - G(\mu)| d\mu$. In contrast, Theorem 1(i) can deduce that $\hat{G}_M(\mu)$ converges to $G(\mu)$ at the optimal rate under a broader model than the one considered in Chen (1995). Specifically, we have

Corollary 1. *Under assumptions of Theorem 1, we have*

$$\int_\mu |\hat{G}_M(\mu) - G(\mu)| d\mu = O_p(n^{-1/4}). \tag{7}$$

Now, we are ready to consider the log-likelihood ratio statistic as follows:

$$LR_n = \sup_{\omega \in \Omega} 2L_n(\omega) - \sup_{\omega \in \Omega_0} 2L_n(\omega),$$

where $\Omega_0 = \{\omega \in \Omega : \alpha = 0.5, \mu_1 = \mu_2\}$. It is noteworthy that even if we confine $\omega \in \Omega_0$, the equation (5) is still true in that $k_2(\omega) = 0$. Therefore, (5) is a unified equation, regardless $\omega \in \Omega_0$. Thus,

$$\sup_{\omega \in \Omega_0} 2L_n(\omega) = W_n(\mu_*)^T J_n^{-1}(\mu_*) W_n(\mu_*) - \inf_{\omega \in \Omega_0} Q_n(\sqrt{n} K(\omega), \mu_*) + o_p(1).$$

To better understand the forgoing approximation, let us examine the case, in which β_* is an interior point of $\mathcal{B}$. The standard asymptotic theory assures that $\sup_{\omega \in \Omega_0} 2L_n(\omega)$ converges to $\chi^2_{q_1+q_2}$ in distribution and

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

where $H = [I_{q_1+q_2}, \mathbf{0}^T]^T$ is a $(q_1+q_2+q_2(q_2-1)/2)\times(q_1+q_2)$ matrix. Thus, the log-likelihood ratio statistic

$$\begin{aligned} LR_n &= 2L_n(\hat{\omega}_M) - \sup_{\omega\in\Omega_0} 2L_n(\omega) \\ &= \sup_{||\mu_2||\le M} \hat{V}(\mu_2)^T J_n(\mu_2)\hat{V}(\mu_2) - W_n(\mu_*)^T H(H^T J_n(\mu_*)H)^{-1}H^T W_n(\mu_*) + o_p(1), \end{aligned}$$

where $Q_n(\hat{V}(\mu_2),\mu_*) = \inf_{\lambda\in\Lambda_{\beta_*}\times\Lambda_{\mu_2}} Q_n(\lambda,\mu_*)$ and Λ_{μ_2} will be defined in the proof of Theorem 1.

To prove Theorem 1, we need the following two lemmas. Lemma 1 has been presented in Zhu and Zhang (2002). We include the proof of Lemma 1 to be self-contained. Lemma 2 follows easily from some traditional inequalities such as Cauchy-Schwarz inequality, and its detailed proof is omitted.

LEMMA 1. *We assume that*

(a.1) there is a continuous function of ω, $K:\Omega\to\mathcal{K}$ where $\mathcal{K}=\{K(\omega):\omega\in\Omega\}$. Let $\Omega_*=\{\omega: K(\omega)=\mathbf{0}\}$ and $\mathbf{0}$ belongs to $cl(\mathcal{K})$, the closed set of $\mathcal{K}$. There is an estimate $\hat{\omega}\in cl(\mathcal{K})$ such that $L_n(\hat{\omega}) = O_p(1)+\sup_{\omega\in\Omega} L_n(\omega)$, $\sup_{\omega\in\Omega} L_n(\omega)\ge 0$ and $K(\hat{\omega})$ converges to zero in probability.

(a.2) An inequality

$$L_n(\omega) \le \sqrt{n}K(\omega)^T\tilde{W}_n(\omega) - \frac{n}{2}K(\omega)^T\tilde{J}_n(\omega)K(\omega) + o_p(n||K(\omega)||^2+1) \qquad (8)$$

holds uniformly for the neighborhood $N_\omega^K[o_p(1)] = \{\omega: K(\omega)=o_p(1)\}$.

(a.3) $\sup_{\omega\in N_\omega^K[o_p(1)]} ||\tilde{W}_n(\omega)|| = O_p(1)$.

(a.4) For a fixed $C_l>0$, $\inf_{\omega\in N_\omega^K[o_p(1)]}\lambda_{\min}(\tilde{J}_n(\omega))\ge C_l^2$ holds almost surely.

Then, we have $K(\hat{\omega}) = O_p(n^{-1/2})$.

Proof. Let $F_n(\omega) = K(\omega)^T \tilde{W}_n(\omega)/||K(\omega)||$. According to (a.1) and inequality in (a.2), the following inequality

$$-L_n(\omega)/n + o_p(n^{-1}) + ||K(\omega)||F_n(\omega)/\sqrt{n} \geq \frac{1}{2}K(\omega)^T \tilde{J}_n(\omega)K(\omega) + o_p(||K(\omega)||^2)$$

holds uniformly for all ω in the neighborhood $N_\omega^K[o_p(1)]$. Moreover, (a.4) implies that

$$K(\omega)^T \tilde{J}_n(\omega)K(\omega) \geq C_l^2 ||K(\omega)||^2$$

almost surely. Also, the $o_p(||K(\omega)||^2)$ term can be bounded by $C_l^2||K(\omega)||^2/4$ as ω is in $N_\omega^K[o_p(1)]$. Therefore, we obtain

$$-L_n(\omega)/n + o_p(n^{-1}) \geq C_l^2||K(\omega)||^2/4 - F_n(\omega)||K(\omega)||/\sqrt{n}.$$

Substituting $\hat{\omega}$ into the above inequality and adding $n^{-1}F_n(\hat{\omega})^2/C_l^2 = O_p(n^{-1})$ to both sides, we have

$$[C_l||K(\hat{\omega})|| - \frac{2}{\sqrt{n}}F_n(\hat{\omega})/C_l]^2 \leq -4n^{-1}L_n(\hat{\omega}) + 4n^{-1}F_n(\hat{\omega})^2/C_l^2 + o_p(n^{-1})$$

$$\leq O_p(n^{-1}) - 4\sup_{\omega\in\Omega} L_n(\omega) + n^{-1}4F_n(\hat{\omega})^2/C_l^2 + o_p(n^{-1}) \leq O_p(n^{-1}) + n^{-1}4F_n(\hat{\omega})^2/C_l^2 + o_p(n^{-1}).$$

Condition (a.3) implies that the right-hand side of the above inequality is in the order of $O_p(n^{-1}) + o_p(n^{-1})$. Thus, $K(\hat{\omega}) = O_p(n^{-1/2})$, since $F_n(\hat{\omega}) = O_p(1)$. This completes the proof for Lemma 1.

Lemma 2. *If* $S_i = F_i + e_i$ *and* $e_i = \sum_{j=1}^m e_{ij}$, *then*

$$\sum_{i=1}^n S_i^3 \leq C_1[\sum_{i=1}^n |F_i|^3 + \sum_{i=1}^n \sum_{j=1}^m |e_{ij}|^3],$$

and

$$\sum_{i=1}^n S_i^2 \geq C_2 \sum_{i=1}^n F_i^2 - C_3\sqrt{\sum_{i=1}^n F_i^2 \sum_{i=1}^n \sum_{j=1}^m e_{ij}^2},$$

where C_1, C_2 and C_3 are some universal constants.

Proof of Theorem 1.

First, we deal with the consistency of the maximum likelihood estimate. To prove that $\hat{\beta}_M - \beta_*$, $\hat{\mu}_{1M} - \mu_*$ and $\hat{\alpha}_M(\hat{\mu}_{2M} - \mu_*)$ converge to zero in probability, we need to use Assumption (1.2) to show that for any $\delta > 0$, $d(\hat{\omega}_M, \Omega_{*,\delta}) \to 0$ in probability, where $d(\mathbf{x}, A)$ is defined to be $\inf_{\mathbf{y}\in cl(A)} ||\mathbf{x} - \mathbf{y}||$. Thus, $K(\hat{\omega}_M) = o_p(1)$. To see this, Assumption (1.2) implies that given $\delta > 0$, there exist $\epsilon_0 > 0$ and large N such that $\omega \notin \Omega_{*,\delta}$ implies that $n^{-1}[\overline{L}_n(\overline{\omega}_n) - \overline{L}_n(\omega)] \geq \epsilon_0$ for all $n \geq N$. Thus,

$$P(\hat{\omega}_M \notin \Omega_{*,\delta}) \leq P(\overline{L}_n(\overline{\omega}_n) - L_n(\overline{\omega}_n) + L_n(\overline{\omega}_n) - L_n(\hat{\omega}_M) + L_n(\hat{\omega}_M) - \overline{L}_n(\hat{\omega}_M) \geq n\epsilon_0)$$

$$\leq P(\overline{L}_n(\overline{\omega}_n) - L_n(\overline{\omega}_n) + L_n(\hat{\omega}_M) - \overline{L}_n(\hat{\omega}_M) \geq n\epsilon_0) \leq P(2\sup_{\omega\in\Omega} n^{-1}|\overline{L}_n(\omega) - L_n(\omega)| \geq \epsilon_0) \to^p 0.$$

From now on, we only consider the consistent estimates in $N_\omega^K[o_p(1)]$.

Second, we start to consider the case such that $|\Delta\mu_2| \leq x_n$, where x_n converges to zero in probability as n tends to infinity. To ease the notation, we set $q_1 = q_2 = 1$ and will discuss the general case later. Let

$$k_1(\omega) = (1-\alpha)\Delta\mu_1 + \alpha\Delta\mu_2 \quad \text{and} \quad k_2(\omega) = (1-\alpha)(\Delta\mu_1)^2 + \alpha(\Delta\mu_2)^2.$$

With above preparations, $S_i(\omega) = p_i(y_i, \mathbf{x}_i; \omega)/f_i(y_i, \mathbf{x}_i; \beta_*, \mu_*) - 1$ can be expressed as

$$S_i(\omega) = K(\omega)^T w_i(\omega_*) + \mathbf{e}_{i,1}(\omega),$$

where $K(\omega) = (\Delta\beta^T, k_1(\omega), k_2(\omega))^T$ and

$$\mathbf{e}_{i,1}(\omega) = (\Delta\beta)^T F_{i,4}(\mu_*)k_1(\omega) + (1-\alpha)(\Delta\beta)F_{i,5}(\mu_1)(\Delta\mu_1)^2 + (1-\alpha)(\Delta\beta)^2 F_{i,3}(\beta, \mu_1) +$$

$$\alpha(\Delta\beta)[F_{i,5}(\mu_2)](\Delta\mu_2)^2 + \alpha(\Delta\beta)^2 F_{i,3}(\beta, \mu_2) + (1-\alpha)F_{i,7}(\mu_1)(\Delta\mu_1)^3 + \alpha F_{i,7}(\mu_2)(\Delta\mu_2)^3.$$

By using Assumption (1.3), it is easy to handle the negligible terms $(\Delta\beta)^T F_{i,4}(\mu_*)k_1(\omega)$, $(1-\alpha)(\Delta\beta)^T F_{i,3}(\beta, \mu_1)(\Delta\beta)$ and $\alpha(\Delta\beta)^T F_{i,3}(\beta, \mu_2)(\Delta\beta)$.

Let's consider some negligible terms as follows:

$$\left|\sum_{i=1}^{n}\left[(1-\alpha)(\Delta\mu_1)^3F_{i,7}(\mu_1)+\alpha(\Delta\mu_2)^3F_{i,7}(\mu_2)\right]\right| \leq n^{1/2}[(1-\alpha)|\Delta\mu_1|^3+\alpha|\Delta\mu_2|^3]O_p(1)$$

$$\leq n^{1/2}k_2(\omega)[o_p(1)+x_n]O_p(1),$$

$$\left|\sum_{i=1}^{n}\left\{(1-\alpha)^2(\Delta\mu_1)^6[F_{i,7}(\mu_1)]^2+\alpha^2(\Delta\mu_2)^6[F_{i,7}(\mu_2)]^2\right\}\right| \leq n[(1-\alpha)^2|\Delta\mu_1|^6+\alpha^2|\Delta\mu_2|^6]O_p(1)$$

$$\leq nk_2^2(\omega)[o_p(1)+x_n^2]O_p(1),$$

and $|(\Delta\beta)\sum_{i=1}^{n}[(1-\alpha)F_{i,5}(\mu_1)(\Delta\mu_1)^2+\alpha F_{i,5}(\mu_2)(\Delta\mu_2)^2]| \leq n|\Delta\beta||k_2(\omega)|o_p(1)$.

It follows from Assumptions (1.3) and (1.4), the inequality $\log(1+x) \leq x - x^2/2 + x^3/3$, Lemma 2 and above discussions that

$$L_n(\omega) = \sum_{i=1}^{n}\log[1+S_i(\omega)] \leq \sqrt{n}K(\omega)W_n(\mu_*) - \frac{n}{2}K(\omega)^TJ_n(\mu_*)K(\omega)$$

$$+o_p(n||K(\omega)||^2[\text{Constant}+g_1(x_n)]+1)+n^{1/2}k_2(\omega)x_nO_p(1)$$

holds for all $|\Delta\mu_2| \leq x_n$ and $\omega \in N_\omega^K[o_p(1)]$, where $g_1(x_n)$ is a function of x_n such that $\lim_{x_n\to 0} g_1(x_n) = 0$; see for example, $g_1(x_n)$ has the term $x_n^2O_p(1)$.

As above, let consider $|\Delta\mu_2| > x_n$, then

$$S_i(\omega) = K(\omega)^Tw_i(\mu_2) + \mathbf{e}_{i,2}(\omega),$$

where

$$\mathbf{e}_{i,2}(\omega) = (1-\alpha)(\Delta\beta)^2F_{i,3}(\beta,\mu_1)+\alpha(\Delta\beta)^2F_{i,3}(\beta,\mu_2)+k_1(\omega)F_{i,4}(\mu_1)\Delta\beta$$

$$+\alpha\Delta\mu_2(\Delta\beta)[F_{i,4}(\mu_1)-F_{i,4}(\mu_2)]+(1-\alpha)[F_{i,6}(\mu_1)-F_{i,6}(\mu_2)](\Delta\mu_1)^2.$$

Similar to previous arguments, we only need to focus on two terms $\alpha\Delta\mu_2(\Delta\beta)[F_{i,4}(\mu_1)-F_{i,4}(\mu_2)]$ and $(1-\alpha)[F_{i,6}(\mu_1)-F_{i,6}(\mu_2)](\Delta\mu_1)^2$. It is easy to find the following relationships:

$$(\Delta\mu_1)^2 \leq [|k_1(\omega)|+|\alpha\Delta\mu_2|]^2 \leq O_p(1)[k_1^2(\omega)+k_2^2(\omega)x_n^{-2}],$$

$$(\Delta\mu_1)^4 \leq O_p(1)[k_1^4(\omega) + k_2^4(\omega)x_n^{-4}], \quad (\Delta\mu_1)^6 \leq O_p(1)[k_1^6(\omega) + k_2^6(\omega)x_n^{-6}],$$

and

$$\alpha|\Delta\mu_2\Delta\beta| \leq k_2(\omega)|\Delta\beta|x_n^{-1} \leq [k_2^2(\omega)x_n^{-2} + |\Delta\beta|^2].$$

Similar to previous arguments, we can get

$$L_n(\omega) \leq \sqrt{n}K(\omega)W_n(\mu_2) - \frac{n}{2}K(\omega)^T J_n(\mu_2)K(\omega)$$

$$+o_p(n||K(\omega)||^2[\text{Const} + g_2(x_n)] + 1) + n^{1/2}k_2(\omega)^2 x_n^{-2}O_p(1),$$

for $|\Delta\mu_2| > x_n$ and $\omega \in N_\omega^K[o_p(1)]$, where $g_2(x_n)$ is a function of x_n such that $\lim_{x_n\to 0} g_2(x_n) = 0$.

Therefore, we have

$$\begin{aligned} L_n(\omega) &\leq \sqrt{n}K(\omega)[I\{|\Delta\mu_2| > x_n\}W_n(\mu_2) + I\{|\Delta\mu_2| \leq x_n\}W_n(\mu_*)] \\ &- \frac{n}{2}K(\omega)^T[I\{|\Delta\mu_2| > x_n\}J_n(\mu_2) + I\{|\Delta\mu_2| \leq x_n\}J_n(\mu_*)]K(\omega) \\ &+ o_p(n||K(\omega)||^2[\text{Const} + g_1(x_n) + g_2(x_n)] + 1) + n^{1/2}k_2(\omega)[k_2(\omega)/x_n^2 + x_n]O_p(1). \end{aligned}$$

If we minimize the function $k_2(\omega)/x_n^2 + x_n$, it is easy to get $x_n = [2k_2(\omega)]^{1/3}$; therefore, we set x_n to be $k_2(\omega)^{1/3}$. Therefore,

$$n^{1/2}k_2(\omega)[k_2(\omega)/x_n^2 + x_n]O_p(1) \leq n^{1/2}k_2(\omega)k_2(\omega)^{1/3}O_p(1) = [1 + nk_2^2(\omega)]o_p(1).$$

Finally,

$$\begin{aligned} L_n(\omega) &\leq \sqrt{n}K(\omega)[I\{|\Delta\mu_2| > k_2(\omega)^{1/3}\}W_n(\mu_2) + I\{|\Delta\mu_2| \leq k_2(\omega)^{1/3}\}W_n(\mu_*)] \\ &- \frac{n}{2}K(\omega)^T[I\{|\Delta\mu_2| > k_2(\omega)^{1/3}\}J_n(\mu_2) + I\{|\Delta\mu_2| \leq k_2(\omega)^{1/3}\}J_n(\mu_*)]K(\omega) \\ &+ o_p(n||K(\omega)||^2 + 1) \end{aligned}$$

holds uniformly for all $\omega \in N_\omega^K[o_p(1)]$. By using Assumption (1.4) and above preparations, we can check all conditions in Lemma 1; therefore, $K(\hat{\omega}_M) = O_p(n^{-1/2})$.

From now on, we only consider $\omega \in \Omega_{*,C_0/\sqrt{n}} = \{\omega \in \Omega : \ ||K(\omega)|| \leq C_0 n^{-1/2}\}$ for any $C_0 > 0$. Since

$$S_i(\omega) = (1-\alpha)\Delta\beta F_{i,1}(\beta,\mu_1) + \alpha\Delta\beta F_{i,1}(\beta,\mu_2) + (1-\alpha)\Delta\mu_1[F_{i,2}(\mu_1) - F_{i,2}(\mu_2)] + k_1(\omega)F_{i,2}(\mu_2),$$

we get

$$\max_{1\leq i\leq n} \sup_{\omega\in\Omega_{*,C_0/\sqrt{n}}} |S_i(\omega)| \leq \frac{O_p(1)}{\sqrt{n}} \sup_{\omega\in B_{\delta_0}} \{|F_{i,1}(\beta,\mu)|\} + n^{-1/4}O_p(1) \sup_{|\mu_2|\leq M} \{|F_{i,2}(\mu_2)|\} = o_p(1).$$

Therefore, $L_n(\omega) = \sum_{i=1}^n S_i(\omega) - 0.5\sum_{i=1}^n S_i^2(\omega)[1+o_p(1)]$ holds uniformly for all $\omega \in \Omega_{*,C_0/\sqrt{n}}$. Moreover, for $\omega \in \Omega_{*,C_0/\sqrt{n}}$, $|\sum_{i=1}^n \mathbf{e}_{i,2}(\omega)| = o_p(1)$. Finally, we can show that

$$L_n(\omega) = \sqrt{n}K(\omega)W_n(\mu_2) - \frac{n}{2}K(\omega)^T J_n(\mu_2)K(\omega) + o_p(1) \tag{9}$$

holds uniformly over $\omega \in \Omega_{*,C_0/\sqrt{n}}$.

Until now, we focus on the case where $q_1 = q_2 = 1$. It is easy to extend q_1 from 1 to any positive integers; in contrast, it is little tedious to increase the value of q_2, because we have to introduce many new notations. However, all previous arguments still work for general (q_1, q_2).

For fixed μ_2 with $||\mu_2 - \mu_*|| \neq 0$, we can show that

$$\mathcal{K}_{\mu_2} = \{K(\omega) : \omega \in \Omega \text{ and } \mu_2 \text{ is fixed}\}$$

can be locally approximated at $\omega(\mu_2)_* = (\lambda_*, \mu_2) = (0, \beta_*, \mu_*, \mu_2)$ by the convex cone $\Lambda_\beta \times \Lambda_{\mu_2}$ (see Subsection 4.1 in Andrews (1999)), where

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

The reason is that for fixed μ_2, we only need to consider $\lambda = (\alpha, \beta, \mu_1)$ in the compact set $[0, 0.5] \times \mathcal{B} \times B(\mathbf{0}, M)$. Moreover, for fixed μ_2, $K(\lambda, \mu_2) = \mathbf{0}$ implies that $\beta = \beta_*$, $\alpha = 0$ and

$\mu_1 = \mu_*$ and

$$K(\lambda, \mu_2) - K(\lambda_*, \mu_2) = \begin{bmatrix} I_{q_1} & \mathbf{0} & \mathbf{0} \\ 0 & \Delta\mu_2 & I_{q_2} \\ \mathbf{0} & \mathrm{Vecs}(\Delta\mu_2^{\otimes 2}) & \mathbf{0} \end{bmatrix} \begin{pmatrix} \beta - \beta_* \\ \alpha \\ \mu_1 - \mu_* \end{pmatrix} + o(||\lambda - \lambda_*||).$$

Now similar to arguments in Lemma 1 of Andrews (1999) or Theorem 4 of Zhu and Zhang (2002), we can show that

$$\inf_{K \in \sqrt{n}\mathcal{K}_{\mu_2}} Q_n(K, \mu_2) = \inf_{\lambda \in \Lambda_{\beta_*} \times \Lambda_{\mu_2}} Q_n(\lambda, \mu_2) + o_p(1)$$

and

$$2 \sup_{\omega \in \Omega_{\mu_2}} L_n(\omega) = \hat{V}(\mu_2)^T J_n(\mu_2) \hat{V}(\mu_2) + o_p(1)$$

where $\hat{V}(\mu_2)$ is defined to be the unique minimizer of $Q_n(\lambda, \mu_2)$ over $\Lambda_{\beta_*} \times \Lambda_{\mu_2}$ for fixed μ_2 with $||\mu_2 - \mu_*|| \neq 0$.

Now, we focus on the case such that $||\mu_2 - \mu_*|| = 0$. Since α can take any value in $[0, 0.5]$, we can temporarily fix α. By using similar arguments above, we can show that

$$\mathcal{K}_{\alpha,\mu_*} = \{K(\omega) : \omega \in \Omega \text{ and } \alpha \text{ is fixed and } \mu_2 = \mu_*\}$$

can be locally approximated at $\omega(\alpha)_* = (\alpha, \beta_*, \mu_*, \mu_*)$ by the convex cone $\Lambda_{\beta_*} \times R^{q_2} \times \mathbf{0}_{q_2}$, which is independent of α. Therefore,

$$\inf_{K \in \sqrt{n}\mathcal{K}_{\mu_*}} Q_n(K, \mu_*) = \inf_{\lambda \in \Lambda_{\beta_*} \times R^{q_2} \times \mathbf{0}_{q_2}} Q_n(\lambda, \mu_*) + o_p(1)$$

and

$$2 \sup_{\omega \in \Omega_{\mu_*}} L_n(\omega) = \hat{V}(\mu_*)^T J_n(\mu_*) \hat{V}(\mu_*) + o_p(1)$$

where $\hat{V}(\mu_*)$ is defined to be the unique minimizer of $Q_n(\lambda, \mu_*)$ over $\Lambda_{\beta_*} \times R^{q_2} \times \mathbf{0}_{q_2}$. Moreover, we can find that $\lim_{\mu_2 \to \mu_*} \hat{V}(\mu_2) = \hat{V}(\mu_*)$.

Define

$$Q(V^0(\mu_2), \mu_2) = \inf_{\lambda \in \Lambda_{\beta_*} \times \Lambda_{\mu_2}} Q(\lambda, \mu_2), \tag{10}$$

where $Q(\lambda, \mu_2) = [\lambda - J(\mu_2)^{-1}W(\mu_2)]^T J(\mu_2)[\lambda - J(\mu_2)^{-1}W(\mu_2)]$. By using continuous mapping theorem, we can show that

$$2 \sup_{\omega \in \Omega} L_n(\omega) \to^d \sup_{||\mu_2|| \le M} [V^0(\mu_2)]^T J(\mu_2) V^0(\mu_2).$$

Proof of Corollary 1. Due to the symmetry for μ_1 and μ_2, without loss of generality, we assume that $\hat{\mu}_{1M} \le \hat{\mu}_{2M}$. Moreover, we have to distinguish two different cases: $A = \{\hat{\mu}_{1M} \le \mu_* \le \hat{\mu}_{2M}\}$ and the complement of A, denoted by $\overline{A}$. It is easy to see that

$$\int_\mu |\hat{G}_M(\mu) - G(\mu)| d\mu = I\{A\} \| \hat{\alpha}_M \Delta\hat{\mu}_{2M} - (1 - \hat{\alpha}_M)\Delta\hat{\mu}_{1M}| + I\{\overline{A}\}|k_1(\hat{\omega}_M)|.$$

According to the definitions of $k_1(\omega)$ and $k_2(\omega)$, we can show that

$$\alpha\Delta\mu_2 - (1-\alpha)\Delta\mu_1 = (1-2\alpha)k_1(\omega) + 2\sqrt{\alpha(1-\alpha)[k_2(\omega) - k_1(\omega)^2]}.$$

Since $k_1(\hat{\omega}_M) = O_p(n^{-1/2})$ and $k_2(\hat{\omega}_M) = O_p(n^{-1/2})$, we know that $\hat{\alpha}_M \Delta\hat{\mu}_{2M} - (1 - \hat{\alpha}_M)\Delta\hat{\mu}_{1M}$ at least has the same convergence rate as $O_p(\sqrt{\hat{\alpha}_M k_2(\hat{\omega}_M)}) = \sqrt{\hat{\alpha}_M} O_p(n^{-1/4})$. Therefore, the convergence rate of $\int_\mu |\hat{G}_M(\mu) - G(\mu)| d\mu$ is at least as small as $\sqrt{\hat{\alpha}_M} O_p(n^{-1/4})$. By using Chen's (1995) Theorem 1, we know that $\int_\mu |\hat{G}_M(\mu) - G(\mu)| d\mu$ is exactly $O_p(n^{-1/4})$.

3 Maximum Modified Likelihood Estimate and Modified Log-likelihood Ratio Statistic

As in Chen *et al* (2001), we consider an alternative approach by using a modified likelihood function

$$ML_n(\omega) = L_n(\omega) + \log(M) \log\{4\alpha(1-\alpha)\}, \tag{11}$$

where M is as defined before. Let $\hat{\omega}_P$ be the resulting estimate as $\hat{\omega}_P = \text{argmax}_{\omega\in\Omega} ML_n(\omega)$. Knowing $L_n(\hat{\omega}_M) = O_p(1)$ from Theorem 1, we have

$$0 \le ML_n(\hat{\omega}_P) \le L_n(\hat{\omega}_M) = O_p(1)$$

and

$$0 \le -\log(M)\log[4\hat{\alpha}_P(1-\hat{\alpha}_P)] \le L_n(\hat{\omega}_P) \le L_n(\hat{\omega}_M) = O_p(1).$$

Thus, $\hat{\alpha}_P(1-\hat{\alpha}_P) = O_p(1)$. In addition, using a similar proof of Theorem 1, we can prove $K(\hat{\omega}_P) = O_p(n^{-1/2})$.

THEOREM 2. *Under the assumptions of Theorem 1, we have the following results:*

(i) $\hat{\alpha}_P = O_p(1), \Delta\hat{\beta}_P = O_p(n^{-1/2}), \quad \Delta\hat{\mu}_{1P} = O_p(n^{-1/4}), \quad \Delta\hat{\mu}_{2P} = O_p(n^{-1/4})$.

(ii) The modified log-likelihood $2ML_n(\hat{\omega}_P)$ *is asymptotically equivalent to*

$$W_n(\mu_*)^T J_n(\mu_*)^{-1} W_n(\mu_*) - \inf_{\omega\in\Omega}\left\{Q_n(\sqrt{n}K(\omega), \mu_*) - 2\log(M)\log[4\alpha(1-\alpha)]\right\},$$

which converges to $\inf_{\lambda\in\Lambda_{\beta_*}\times\Lambda_0} Q(\lambda,\mu_*)$ *in distribution, where* Λ_0 *and* $Q(\lambda,\mu_*)$ *will be defined in the proof of Theorem 2.*

(iii) Furthermore, if $q_2 = 1$, $\int_\mu |\hat{G}_P(\mu) - G(\mu)|d\mu = O_p(n^{-1/4})$, *where* $\hat{G}_P(\mu) = (1-\hat{\alpha}_P)I\{\mu \ge \hat{\mu}_{1P}\} + \hat{\alpha}_P I\{\mu \ge \hat{\mu}_{2P}\}$.

Theorem 2 (i) gives the exact convergence rate of $\hat{\omega}_P$. Theorem 2 (ii) determines the asymptotic distribution of $2ML_n(\hat{\omega}_P)$. While the explicit form of this distribution is still complicated in general, this corollary leads to a neat asymptotic distribution, $0.5\chi^2_{q_1+2} + 0.5\chi^2_{q_1+1}$, for $2ML_n(\hat{\omega}_P)$ when $q_2 = 1$ and β_* is an interior point of $\mathcal{B}$. This coincides with Theorem 1 of Chen *et al* (2001) when there are no covariates, that is, $q_1 = 0$. Theorem 2 (iii) shows that the $n^{-1/4}$-consistent rate for estimating the mixing distribution $G(\mu)$ is reachable by using $\hat{\omega}_P$. Furthermore, to test the hypothesis (4), the modified log-likelihood ratio statistic is defined as

$$MLR_n = 2ML_n(\hat{\omega}_P) - \sup_{\omega\in\Omega_0} 2ML_n(\omega). \tag{12}$$

Similar to the log-likelihood ratio statistic, we find that when β_* is an interior point of $\mathcal{B}$,

$$MLR_n = \inf_{\lambda \in R^{q_1} \times \Lambda_0} Q_n(\lambda, \mu_*) - W_n(\mu_*)^T H (H^T J_n(\mu_*) H)^{-1} H^T W_n(\mu_*) + o_p(1). \quad (13)$$

Proof of Theorem 2. Similar to Theorem 1,

$$2ML_n(\omega) \leq \sqrt{n} K(\omega) W_n(\mu_2) - \frac{n}{2} K(\omega)^T J_n(\mu_2) K(\omega) + o_p(n||K(\omega)||^2 + 1)$$

holds uniformly for all $\omega \in N_\omega^K[o_p(1)]$. Thus, $K(\hat{\omega}_P) = O_p(n^{-1/2})$. Furthermore, we can show that

$$ML_n(\omega) = \sqrt{n} K(\omega) W_n(\mu_*) - \frac{n}{2} K(\omega)^T J_n(\mu_*) K(\omega) + \log(M) \log[4\alpha(1-\alpha)] + o_p(1)$$

holds uniformly over $\omega \in \Omega_{*, C_0/\sqrt{n}}$ and

$$\begin{aligned} 2ML_n(\hat{\omega}_P) &= W_n(\mu_*)^T J_n(\mu_*)^{-1} W_n(\mu_*) - \inf_{\omega \in \Omega} \left\{ Q_n(\sqrt{n} K(\omega), \mu_*) - 2\log(M) \log[4\alpha(1-\alpha)] \right\} \\ &= W_n(\mu_*)^T J_n(\mu_*)^{-1} W_n(\mu_*) - \inf_{\alpha \in [0, 0.5]} \left\{ \inf_{\omega \in \Omega_\alpha} Q_n(\sqrt{n} K(\omega), \mu_*) - 2\log(M) \log[4\alpha(1-\alpha)] \right\}. \end{aligned}$$

Let $\Omega_\alpha = \{\omega \in \Omega : \alpha \text{ is fixed}\}$. It follows from the definition of $K(\omega)$ that $\mathcal{K}_\alpha = \{K(\omega) : \omega \in \Omega_\alpha\}$ can be locally approximated by $\Lambda_{\beta_*} \times \Lambda_\alpha$, where Λ_α is a closed cone defined by

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

Using the same arguments in Lemma 1 of Andrews (1999), we can show that

$$\inf_{\omega \in \Omega_\alpha} Q_n(\sqrt{n} \tilde{K}(\omega), \mu_*) = \inf_{\lambda \in \Lambda_{\beta_*} \times \Lambda_\alpha} Q_n(\lambda, \mu_*) + o_p(1),$$

holds uniformly for all $\alpha \in [\alpha_0, 0.5]$, where α_0 is any positive scalar. Moreover, because $\hat{\alpha}_P = O_p(1)$, Λ_α is independent of α and $4\alpha(1-\alpha)$ is maximized at $\alpha = 0.5$, we can deduce that

$$\begin{aligned} 2ML(\hat{\omega}_P) &= \inf_{\lambda \in \Lambda_{\beta_*} \times \Lambda_\alpha} Q_n(\lambda, \mu_*) - 2\log(M) \sup_{\alpha \in [0, 0.5]} \log[4\alpha(1-\alpha)] + o_p(1) \\ &= \inf_{\lambda \in \Lambda_{\beta_*} \times \Lambda_\alpha} Q_n(\lambda, \mu_*) + o_p(1). \end{aligned}$$

Therefore, $2ML(\hat{\omega}_P)$ converges to $\inf_{\lambda \in \Lambda_{\beta_*} \times \Lambda_\alpha} Q(\lambda, \mu_*)$ in distribution.

4 Asymptotic Local Power

Theoretically, many authors are interested in the asymptotic local power of the testing statistics, because power considerations may guide the choice of sample size and compare alternative tests of significance (Cox and Hinkley, 1975, p.103). It becomes apparent from the quadratic approximations that 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. Therefore, it is worthwhile to explore its property under a sequence of local alternatives. Consider a sequence of local alternatives ω^n consisting of

$$\begin{aligned}
\alpha^n &= \alpha_0, \quad \beta^n = \beta_* + n^{-1/2}\mathbf{h}_1, \\
\mu_1^n &= \mu_* - n^{-1/4}\{\alpha_0/(1-\alpha_0)\}^{0.5}\mathbf{h}_2, \\
\mu_2^n &= \mu_* + n^{-1/4}\{(1-\alpha_0)/\alpha_0\}^{0.5}\mathbf{h}_2,
\end{aligned}$$

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 $\mathbf{h}^T = (\mathbf{h}_1^T, \mathbf{0}^T, \text{Vecs}(\mathbf{h}_2^{\otimes 2})^T)$.

Theorem 3. *Under assumptions (1.1)-(1.4) and the alternatives ω^n, $J_n(\mu_2)^{-1}W_n(\mu_2) \rightarrow^d N(J(\mu_2)^{-1}J(\mu_2,\mu_*)\mathbf{h}, J(\mu_2)^{-1})$.*

Proof of Theorem 3. Under the local alternative ω^n, we find that $\Delta\beta^n = n^{-1/2}\mathbf{h}_1, k_1(\omega^n) = \mathbf{0}$ and $k_2(\omega^n) = n^{-1/2}\text{Vecs}(\mathbf{h}_2^{\otimes 2})$. Similar to Theorem 1, we have

$$L_n(\omega^n) = \mathbf{h}^T W_n(\mu_*) - \frac{1}{2}\mathbf{h}^T J_n(\mu_*)\mathbf{h} + o_p(1).$$

Therefore, under H_0, we can show that $(J_n^{-1}(\mu_2)W_n(\mu_2), L_n(\omega^n))$ converges to normal distribution with mean $(\mathbf{0}^T, -0.5\sigma_{22})^T$ and covariance matrix

$$\begin{pmatrix} J^{-1}(\mu_2) & J(\mu_2)^{-1}J(\mu_2,\mu_*)\mathbf{h} \\ \mathbf{h}^T J(\mu_*,\mu_2)J(\mu_2)^{-1} & \sigma_{22} \end{pmatrix}$$

where $\sigma_{22} = \mathbf{h}^T J(\mu_*)\mathbf{h}$. The completed proof follows from LeCam's third lemma (van der Vaart, 1998, p.90).

5 Two Theoretical Examples

To demonstrate the usefulness of our general theory, we examine two examples and apply our theory.

Example 1. For the purpose of illustration, let us consider a simplified logistic mixture regression model as follow

$$\text{logit}(P\{Y_{ij} = 1|U_i\}) = \mathbf{x}_{ij}^T\beta + [U_i\mu_1 + (1 - U_i)\mu_2], \tag{14}$$

where μ_1 and μ_2 are scalars.

Furthermore, Theorem 1 is also applicable to another interesting case in which $\mathbf{x}_{ij} = \mathbf{z}_{ij}$. Namely,

$$\text{logit}(P\{Y_{ij} = 1|U_i\}) = \mathbf{z}_{ij}^T[\beta + U_i\mu_1 + (1 - U_i)\mu_2].$$

Specifically, $\omega_i(\mu)$ is given by

$$\omega_i(\mu) = \begin{pmatrix} \sum_{j=1}^{n_i} e_{ij}(\mu_*)\mathbf{z}_{ij} \\ \text{Vecs}\{(\sum_{j=1}^{n_i} e_{ij}(\mu)\mathbf{z}_{ij})^{\otimes 2} - \sum_{j=1}^{n_i}(\mathbf{z}_{ij})^{\otimes 2} p_{ij}(\mu)[1 - p_{ij}(\mu)]\} \end{pmatrix},$$

in which $e_{ij}(\mu) = y_{ij} - p_{ij}(\mu)$ and $p_{ij}(\mu) = \exp(\mathbf{z}_{ij}^T\mu)/[1 + \exp(\mathbf{z}_{ij}^T\mu)]$.

Let's show that Theorem 1 holds for this important case by using the following assumptions.

(2.1) The same as (1.1).

(2.2) For all $i = 1, \cdots, n$ and $j = 1, \cdots, n_i$, $|n_i| \le N_0 < \infty$ and $||\mathbf{x}_j^i|| \le N_0$, where N_0 is a large scalar.

(2.3) For every $\delta > 0$, $\sup_{\omega \in \Omega/\Omega_{*,\delta}} n^{-1}E[L_n(\omega)] < 0$, where

$$E[L_n(\omega)] = \sum_{i=1}^{n} \int \log[p_i(y_i, \mathbf{x}_i, \omega)/f_i^*] f_i^*(y_i, \mathbf{x}_i) g_i(\mathbf{x}_i) dy_i d\mathbf{x}_i,$$

which means that $E[L_n(\omega)]$ is the average of the Kullback-Leibler distance to the true model from all families. (2.3) implies that we only need a proportion of families based on which the model is identifiable around the true model. Sometimes, the techniques used by Cheng and Liu (2001) can be used here to justify this assumption.

(2.4) There exists a $J(\mu)$ such that $\sup_{|\mu|\le M} ||EJ_n(\mu) - J(\mu)|| \to 0$ in probability and $\infty > \sup_{|\mu|\le M} \lambda_{\max}[J(\mu)] \ge \inf_{|\mu|\le M} \lambda_{\min}[J(\mu)] > 0$ holds almost surely. To prove the positivity of $J(\mu)$, the characteristic function technique used in Section 4 of Chen (1995) can be used to validate this assumption under the independent and identical framework.

Check Assumptions (1.1)-(1.4): We need to check all assumptions (1.1)-(1.4). Starting from (1.2), we need to show that $\sup_{\omega\in\Omega} n^{-1}||L_n(\omega) - EL_n(\omega)||$ converges to zero in probability. First, we can show that for each $\omega \in \Omega$, $\mathrm{Var}\{n^{-1}[L_n(\omega) - EL_n(\omega)]\}$ converges to zero as $n \to \infty$; that is, $n^{-1}|L_n(\omega) - EL_n(\omega)| = o_p(1)$ for each $\omega \in \Omega$. Secondly, $n^{-1}\{L_n(\omega) - EL_n(\omega)\}$ is stochastically equicontinuous on Ω, since $n^{-1}|L_n(\omega) - EL_n(\omega) - L_n(\omega') + EL_n(\omega')| \le \mathrm{Const} \times ||\omega - \omega'||$. Combining above arguments, we can deduce that $\sup_{\omega\in\Omega} n^{-1}||L_n(\omega) - EL_n(\omega)||$ converges to zero in probability; see Theorem 1 of Andrews (1992). By using Jensen inequality, we know that for all ω corresponding to P_*, $EL_n(\omega)$ reaches the maxima *zero*. Thus (1.2) follows from (2.2) and (2.3).

All functions in Assumptions (1.3) are bounded; therefore,

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

By using Theorem 1 of Andrews (1992), we know

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

Let $Q_n(\mu) = \sum_{i=1}^{n} F_{i,7}(\mu)/\sqrt{n}$. Since $\mathrm{Var}[Q_n(\mu)] = \sum_{i=1}^{n}\mathrm{Var}[F_{i,7}(\mu)]/n < \infty$ and $E[Q_n(\mu) - Q_n(\mu')]^2 \le \mathrm{Const} \times (\mu - \mu')^2$, $\sup_{(\beta,\mu)\in\mathbf{B}_{\delta_0}} ||\sum_{i=1}^{n} F_{i,7}(\mu)/\sqrt{n}|| = O_p(1)$. Similarly, we can check all conditions in (1.3).

To check (1.4), we first prove that $\sup_{|\mu|\leq M}||J_n(\mu)-EJ_n(\mu)||$ converges to zero in probability by using Theorem 1 of Andrews (1992). Second, using (2.4), we can show that for each $\mu \in [-M, M]$, $W_n(\mu) = \sum_{i=1}^n w_i(\mu)/\sqrt{n}$ converges to $N[\mathbf{0}, J(\mu)]$ in distribution. Moreover, since $E[W_n(\mu) - W_n(\mu')]^2 \leq \text{const} \times (\mu - \mu')^2$, we invoke the weak convergence theorem on the space $C[-M, M]$ with uniform metric to prove that $W_n(\mu)$ converges to a Gaussian process indexed by $\mu \in [-M, M]$; see Pollard (1990).

In this example, we put several simple conditions to validate the high-level assumptions (1.1)-(1.4). The soundness of these new conditions is another important issue. Let's focus on the most difficult assumption (2.4) under the model (14). We are going to demonstrate that assumption (2.4) is quite reasonable in this case. First, we can show that

$$\omega_i(\mu) = \left(\sum_{j=1}^{n_i} e_{ij}\mathbf{x}_{ij}, \sum_{j=1}^{n_i} e_{ij}, [\sum_{j=1}^{n_i} e_{ij}(\mu)]^{\otimes 2} - \sum_{j=1}^{n_i} p_{ij}(\mu)[1 - p_{ij}(\mu)] \right).$$

Second, since $J_n(\mu) = \sum_{i=1}^n E[\omega_i(\mu)^{\otimes 2}]$, we need to focus on all $E[\omega_i(\mu)^{\otimes 2}]$s. If we can show that $E[\omega_i(\mu)^{\otimes 2}]$ is positive for each i, the positivity of $J_n(\mu)$ will be reasonable. Especially, under the i.i.d framework, $J_n(\mu) = E[\omega_1(\mu)^{\otimes 2}]$; therefore, assumption (2.4) is obvious. In general situation, if $J_n(\mu)$ is positive, the assumption that $J_n(\mu)$ converges to $J(\mu)$ is quite natural in the literature.

A sufficient condition for the positivity of $E[\omega_i(\mu)^{\otimes 2}]$ is that $\text{Var}[\omega_i(\mu)^{\otimes 2}]$ is positive definite. Further, for simplicity, we set $n_i = 1$ and get rid of subscript j. By using direct arithmatic, we know that

$$\text{Var}[\omega_i(\mu)^{\otimes 2}] = E_{\mathbf{x}_i} \left\{ A_i(\mu) E[b_i(\mathbf{x}_i)^{\otimes 2} | \mathbf{x}_i] A_i(\mu)^T \right\} \tag{15}$$

where $A_i(\mu)$ and $b_i(\mathbf{x}_i)$ are given by

$$A_i(\mu) = \begin{pmatrix} I_{q_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 1 & 0 \\ 0 & 2[p_i(\mu_*) - p_i(\mu)] & 1 \end{pmatrix}, \quad b_i(\mathbf{x}_i) = \begin{pmatrix} e_i\mathbf{x}_i \\ e_i \\ e_i^2 - p_i(\mu_*)[1 - p_i(\mu_*)] \end{pmatrix}.$$

If $\mathrm{Var}[(\mathbf{x}_i^T, 1)]$ is positive definite, then both $E[b_i(\mathbf{x}_i)^{\otimes 2}]$ and $\mathrm{Var}[\omega_i(\mu)^{\otimes 2}]$ are positive definite by noting the fact (15).

Example 2. Let's consider a simple case as follows:

$$y_{ij} = \mathbf{z}_{ij}^T[\beta + U_i\mu_1 + (1 - U_i)\mu_2] + \epsilon_{ij}.$$

In this case,

$$\omega_i(\mu) = 0.5\sigma^{-4} \begin{pmatrix} 2\sigma^2 \sum_{j=1}^{n_i} e_{ij}(\mu_*)\mathbf{z}_{ij} \\ \sum_{j=1}^{n_i} [e_{ij}(\mu_*)]^2 - n_i\sigma_*^2 \\ 2\mathrm{Vecs}\{(\sum_{j=1}^{n_i} e_{ij}(\mu)\mathbf{z}_{ij})^{\otimes 2} - \sigma_*^2 \sum_{j=1}^{n_i}(\mathbf{z}_{ij})^{\otimes 2}]\} \end{pmatrix},$$

where $e_{ij}(\mu) = y_{ij} - \mu$. When $n_i = n_0$, $\mu_* = 0$, $\sigma_* = 1$, and $\mathbf{z}_{ij} = 1$, we find that $J_n(\xi)$ converges to

$$0.25\sigma^{-8} \begin{pmatrix} 4\sigma^4 n_0 & 4\sigma^2 n_0 \Delta\mu & 8\sigma^2 n_0^2 \Delta\mu \\ & 2n_0 + 4n_0(\Delta\mu)^2 & 4n_0 + 8n_0^2(\Delta\mu)^2 \\ & & 8n_0^2 + 16n_0^3(\Delta\mu)^2 \end{pmatrix}$$

almost surely. It can be seen that above matrix is positive definite if and only if $n_0 > 1$, which actually checks Assumption (1.4). Similar to assumptions (2.1)-(2.4), we can develop some assumptions of this linear mixture regression model for $n_0 > 1$ such that Theorems 1-3 hold. However, Theorem 1 is not applicable when $n_0 = 1$, although this special case can be accommodated by modifying the recent results in Chen and Chen (2003). Surprisingly, an interesting observation is that the asymptotic distribution depends on the number of observations in each cluster.

References

Andrews, D. W. K. (1992). Generic Uniform convergence. *Econometric Theory*, **8**, 241-257.

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

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

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. and Chen, J. (2003). Tests for homogeneity in normal mixtures with presence of a structural parameter: technical details. Preprint, *Statistica Sinica.*

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

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

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics*, **54**, 573-578.

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

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.

Pötscher, B. and Prucha, I. (1991). Basic structure of the asymptotic theory in dynamic nonlinear econometric models. I. Consistency and approximation concepts. *Econometric Reviews*, **10**, 125-216.

Pollard, D. (1984). *Convergence of Stochastic Processes.* New York: Springer-Verlag.

Pollard, D. (1990). *Empirical Processes: Theory and Applications.* CBMS Regional Conference Series in Probability and Statistics, Volumn 2, Institute of Mathematical Statistics: Hayward, CB.

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

van der Vaart (1998). *Asymptotic Statistics.* Cambridge University Press, Cambridge, UK.

Zhu, H.T. and Zhang, H.P. (2002). Asymptotics for estimation and testing procedures under loss of identifiability. Technical Report. Yale University School of Medicine.

Zhu, H.T. and Zhang, H.P. (2004). Hypothesis testing in mixture regression models, JRSS-B, 66:3-16.