The Annals of Statistics
2003, Vol. 31, No. 3, 807–832

ASYMPTOTICS FOR LIKELIHOOD RATIO TESTS UNDER LOSS OF IDENTIFIABILITY

BY XIN LIU[1] AND YONGZHAO SHAO[2]

Rockefeller University and Columbia University

This paper describes the large sample properties of the likelihood ratio test statistic (LRTS) when the parameters characterizing the true null distribution are not unique. It is well known that the classical asymptotic theory for the likelihood ratio test does not apply to such problems and the LRTS may not have the typical chi-squared type limiting distribution. This paper establishes a general quadratic approximation of the log-likelihood ratio function in a Hellinger neighborhood of the true density which is valid with or without loss of identifiability of the true distribution. Under suitable conditions, the asymptotic null distribution of the LRTS under loss of identifiability can be obtained by maximizing the quadratic form. These results extend the work of Chernoff and Le Cam. In particular, applications to testing the number of mixture components in finite mixture models are discussed.

1. Introduction. Many hypothesis testing problems involve a family of probability distributions $\{P_\theta, \theta \in \Theta\}$ which is assumed known except for some parameter θ in the parameter space Θ. Typically, Θ is a subset of some finite-dimensional metric space. We say that there is loss of identifiability in parameters if $P_\theta = P_{\theta'}$ for some $\theta \neq \theta'$ in Θ. The problem of loss of identifiability occurs in diverse areas such as econometrics, reliability theory and survival analysis [Prakasa Rao (1992)]. It is well known that the classical asymptotic theory for the likelihood ratio test (LRT) does not apply when there is loss of identifiability of the true distribution [Lindsay (1995)]. This paper provides a general approach for deriving the asymptotic null distribution of the likelihood ratio test statistic (LRTS) in this type of hypothesis testing problem.

For simplicity, we assume the null distribution, denoted by P, is unique. Denote by Θ_0 the set of true parameters corresponding to P; that is, $\Theta_0 = \{\theta \in \Theta : P_\theta = P\}$. A common hypothesis testing problem, with (or without) loss of identifiability, is to test

$$H_0 : P_\theta, \theta \in \Theta_0 \quad \text{against} \quad H_1 : P_\theta, \theta \in \Theta \setminus \Theta_0. \tag{1.1}$$

Received November 1999; revised April 2002.

[1]Supported in part by National Human Genome Research Institute Grant HG00008.

[2]Supported in part by NSF Grant DMS-96-26658.

AMS 2000 subject classifications. Primary 62F05; secondary 62H30, 62A10.

Key words and phrases. Donsker class, finite mixture model, Hellinger distance, likelihood ratio test, loss of identifiability.

The LRTS $2\lambda_n$ based on random observations $X_1, \ldots, X_n$ can be expressed as

$$2\lambda_n = 2 \sup_{\theta \in \Theta} L_n(\theta) \equiv 2 \sup_{\theta \in \Theta} \sum_{i=1}^{n} \log l_\theta(X_i),$$

where $l_\theta = dP_\theta / dP$ is the likelihood ratio function.

In the classical likelihood theory, the parameter characterizing the true null distribution is typically assumed to be a unique point θ_0 in some open subset Θ of $\Re^d$. The classical regularity conditions ensure the consistency of the maximum likelihood estimator (MLE) $\hat{\theta}_n$ and the existence of a quadratic approximation to $L_n(\theta)$ in a Euclidean $n^{-1/2}$-neighborhood of θ_0 [see, e.g., Chernoff (1954)],

$$2L_n(\theta) = 2\sqrt{n}(\theta - \theta_0)^T \nu_n \mathbf{S} - n(\theta - \theta_0)^T \mathbf{I}(\theta - \theta_0) + o_P(1), \tag{1.2}$$

where $\mathbf{S}(x) = l'_{\theta_0}(x)$ is the score function; $\mathbf{I} = E_P(\mathbf{S}\mathbf{S}^T)$ is the Fisher information matrix which is assumed nondegenerate; $\nu_n f = n^{-1/2} \sum_{i=1}^{n} (f(X_i) - \int f \, dP)$ for $f \in \mathcal{L}^1(P)$. Then the asymptotic normality of the MLE and the asymptotic null distribution of the LRTS can be obtained by maximizing the above quadratic form. However, when the parameters representing the true null distribution are not unique, the classical likelihood theory is no longer applicable and various difficulties arise in analyzing the asymptotic properties of the LRT. For example, the MLE $\hat{\theta}_n$ may not converge to any fixed point in Θ_0 and some directional scores may be zero, thereby leading to degeneracy of the Fisher information matrix $\mathbf{I}$ and failure of the quadratic expansion (1.2). In general, the limit distribution of the LRTS may not be the chi-squared type as predicted by the classical theory and it can be very hard to characterize. Typical hard problems of this kind include testing the number of components in finite mixture models and testing the order of a stationary ARMA process [Lindsay (1995) and Dacunha-Castelle and Gassiat (1999)].

This paper describes the asymptotic properties of the LRTS for (1.1) under loss of identifiability of the true distribution. Under some general conditions, the maximum likelihood density estimator is consistent even when there is loss of identifiability in the true parameters [Redner (1981)]. When the likelihood ratio is square integrable, the convergence of the ML density estimator can be measured by its $\mathcal{L}^2$ distance to the true density. Denote by $D(\cdot, \cdot)$ the $\mathcal{L}^2$ distance; that is, $D^2(\theta_1, \theta_2) = E_P(l_{\theta_1} - l_{\theta_2})^2$ for $\theta_1, \theta_2 \in \Theta$. Then it is natural to expand $L_n(\theta)$ in some $\mathcal{L}^2$-neighborhood of the true distribution. Moreover, when the true parameter is unique and the classical regularity conditions hold, the $\mathcal{L}^2$ distance $D(\theta) \equiv D(\theta, \theta_0)$ and the Euclidean distance $|\theta - \theta_0|$ are locally equivalent in some shrinking neighborhoods of θ_0. In particular, when the quadratic expansion in (1.2) holds for $|\theta - \theta_0| = O(n^{-1/2})$, it is easy to see that

$$D^2(\theta) = (\theta - \theta_0)^T \mathbf{I}(\theta - \theta_0) + o(|\theta - \theta_0|^2).$$

Thus $|\theta - \theta_0| = O(n^{-1/2})$ implies $D(\theta) = O(n^{-1/2})$ and vice versa. Moreover,

$$\sqrt{n}(\theta - \theta_0)^T \nu_n \mathbf{S} = \sqrt{n}\nu_n(l_\theta - 1) + o_P(1).$$

Define the *generalized score function* as $S_\theta = (l_\theta - 1)/D(\theta)$ for $\theta \in \Theta \setminus \Theta_0$. Then, when (1.2) holds,

$$2L_n(\theta) = 2\sqrt{n}D(\theta)\nu_n S_\theta - nD^2(\theta) + o_P(1). \tag{1.3}$$

Unlike the expansion in (1.2), (1.3) does not require uniqueness of the true parameter and holds under loss of identifiability of the true distribution. In many applications, the empirical process $\nu_n S_\theta$ converges in distribution to some Gaussian process. For instance, when $\{S_\theta : \theta \in \Theta \setminus \Theta_0\}$ is a P-Donsker class, $\nu_n S_\theta$ converges uniformly to some zero-mean Gaussian process $\{W_{S_\theta} : \theta \in \Theta \setminus \Theta_0\}$ with continuous sample paths. Then, using the almost sure representation for weak convergence, the quadratic form in (1.3) can be formulated as

$$2L_n(\theta) = 2\sqrt{n}D(\theta)W_{S_\theta} - nD^2(\theta) + o_P(1). \tag{1.4}$$

The supremum of $2L_n(\theta)$ can be obtained by maximizing the quadratic form of $\sqrt{n}D(\theta)$ in (1.4). Under some general regularity conditions, Theorem 3.1 of this paper asserts that on a suitable probability space

$$\lim_{n\to\infty} 2\lambda_n = \sup_{S\in\mathcal{F}} [\max(W_S, 0)]^2, \tag{1.5}$$

where $\mathcal{F}$ is the set of the $\mathcal{L}^2$ limits (i.e., cluster points) of the generalized score functions S_θ as $D(\theta) \to 0$.

As applications of this general approach, we characterize the asymptotics of the LRTS for testing the number of components in finite mixture models. Because of the importance of finite mixture models in applications, there are extensive investigations on the asymptotic behavior of the LRTS in mixture models [see, e.g., Titterington, Smith and Makov (1985), McLachlan and Basford (1988) and Lindsay (1995)]. Recently the asymptotic distribution of this LRTS has been derived by Dacunha-Castelle and Gassiat (1997, 1999) using "locally conic parameterization." Their approach is useful in deriving the index set $\mathcal{F}$ in (1.5); however, it does not lead to optimal assumptions. We characterize the asymptotic null distributions of the LRTS under some general conditions given in Section 4.

The paper is organized as follows: Section 2 establishes quadratic approximations to likelihood ratios in a Hellinger neighborhood of the true distribution without requiring square integrability of the likelihood ratios and also provides the asymptotic null distribution of the LRTS. Section 3 extends the results in Section 2 to square integrable likelihood ratios under P-Donsker conditions for the class of generalized score functions. The LRT of discrete models and hypothesis testing problems with a composite null hypothesis are also considered. Section 4 characterizes the asymptotic null distribution of the LRTS for testing the number of components in finite mixture models.

2. General quadratic approximations using Hellinger distance. In this section we establish quadratic approximations to likelihood ratios in a Hellinger neighborhood of the true distribution. Section 2.1 provides the notation used throughout this paper. Section 2.2 introduces the generalized differentiable in quadratic mean (GDQM) expansion which is used to obtain a specific quadratic approximation of the log-likelihood ratio in Section 2.3. The asymptotic null distribution of the LRTS is derived in Section 2.4.

2.1. *Notation.* The notation to be used is listed first for easy reference. Throughout this paper, vectors and matrices are denoted by boldface letters. We will use the abbreviation $Pf = \int f\,dP$ for an integrable function f and a signed measure P. For $f \in \mathcal{L}^2(P)$, define the $\mathcal{L}^2(P)$ norm as $\|f\| \equiv \sqrt{P(f^2)}$. The map $\Omega: \mathcal{L}^2(P) \to \mathcal{L}^2(P)$ is defined as $\Omega(f) \equiv f/\|f\|$ if $f \neq 0$. For a k-dimensional vector $\mathbf{x} = (x_1, \ldots, x_k)$, define $|\mathbf{x}| \equiv \sqrt{x_1^2 + \cdots + x_k^2}$ and the map $\omega: \Re^k \to \Re^k$ as $\omega(\mathbf{x}) \equiv \mathbf{x}/|\mathbf{x}|$ $(\mathbf{x} \neq 0)$. The *empirical measure* P_n of random observations $X_1, \ldots, X_n$ is defined as $P_n \equiv n^{-1}\sum_{i=1}^n \delta_{X_i}$, where $\delta_{X_i}(A) = I_A(X_i)$ for any measurable set A. Given a collection $\mathcal{F}$ of $\mathcal{L}^1(P)$ functions, the $\mathcal{F}$-indexed *empirical process* ν_n is given by $\{\nu_n f \equiv \sqrt{n}(P_n - P)f,\ f \in \mathcal{F}\}$. The envelope function of a class of functions $\mathcal{F}$ is defined as $F(x) \equiv \sup_{f\in\mathcal{F}} |f(x)|$.

DEFINITION 2.1. A family of random sequences $\{Y_n(g): g \in \mathcal{G},\ n = 1, 2, \ldots\}$ is said to be uniformly $O_P(1)$ if for every $\delta > 0$, there exist constants $M > 0$ and $N(\delta, M)$ such that $P(\sup_{g\in\mathcal{G}} |Y_n(g)| \leq M) \geq 1 - \delta$ for all $n \geq N(\delta, M)$. A family of random sequences $\{Y_n(g): g \in \mathcal{G},\ n = 1, 2, \ldots\}$ is said to be uniformly $o_P(1)$ if for every $\delta > 0$ and $\varepsilon > 0$, there exists a constant $N(\delta, \varepsilon)$ such that $P(\sup_{g\in\mathcal{G}} |Y_n(g)| < \varepsilon) \geq 1 - \delta$ for all $n \geq N(\delta, \varepsilon)$.

We assume that the null distribution, denoted by P, is unique. All results in this paper are considered under the null hypothesis and all expectations are taken with respect to the null probability measure P. Suppose that $\{P_\theta;\ \theta \in \Theta\}$ is a family of probability distributions which is assumed known except for some parameters θ in the parameter space Θ. Always Θ will be a subset of some metric space. Denote by Θ_0 the set of parameters corresponding to P; that is, $\Theta_0 = \{\theta \in \Theta: P_\theta = P\}$. Suppose $X_1, \ldots, X_n$ are i.i.d. random observations from P. We investigate the asymptotic properties of the LRTS for testing the hypotheses

$$H_0: X_i \sim P_\theta,\ \theta \in \Theta_0 \quad \text{against} \quad H_1: X_i \sim P_\theta,\ \theta \in \Theta \setminus \Theta_0. \tag{2.1}$$

Denote by $l_\theta = dP_\theta/dP$ the Radon–Nikodym derivative and $L_n(\theta)$ the log-likelihood ratio function:

$$L_n(\theta) \equiv \sum_{i=1}^n \log(l_\theta(X_i)).$$

Let $x \vee y = \max(x, y)$. The LRT $2\lambda_n$ for (2.1) can be written as

$$2\lambda_n = 2 \sup_{\theta \in \Theta} L_n(\theta) = 2 \sup_{\theta \in \Theta \setminus \Theta_0} \big(L_n(\theta) \vee 0\big).$$

Let $H(\theta)$ be the P-Hellinger distance between P_θ and P; that is, $H^2(\theta) = P[(\sqrt{l_\theta} - 1)^2]/2$. When the likelihood ratios are square integrable, denote the Pearson type $\mathcal{L}^2$ distance by $D(\theta)$, where $D^2(\theta) = P(l_\theta - 1)^2$. In many applications, the likelihood ratio is a continuous function of θ. Without much loss of generality, we assume that $H(\theta)$ and $D(\theta)$ are bounded continuous functions for $\theta \in \Theta$. Under fairly general conditions, the maximum likelihood density estimator is consistent in Hellinger distance. Then for $\varepsilon > 0$, as $n \to \infty$, with probability going to 1, we have $\hat{\theta}_n \in \Theta_\varepsilon \cup \Theta_0$ and $2\lambda_n - 2 \sup_{\theta \in \Theta_\varepsilon} (L_n(\theta) \vee 0) \to 0$, where

$$\Theta_\varepsilon = \{\theta \in \Theta : 0 < H(\theta) \le \varepsilon\}.$$

The regularity conditions of this paper imply Hellinger consistency of the ML density estimator, so the asymptotic null distribution of the LRTS is determined by the local properties of the likelihood functions in a small Hellinger neighborhood Θ_ε of P for some $\varepsilon > 0$. Thus, we can focus on the limiting distribution of the restricted LRTS $\lim_{n\to\infty} 2 \sup_{\theta \in \Theta_\varepsilon} (L_n(\theta) \vee 0)$, for some $\varepsilon > 0$.

2.2. *Generalized DQM expansion.* Let $h_\theta = \sqrt{l_\theta} - 1$. Le Cam's DQM condition [Le Cam (1970)] can be formulated as

$$h_\theta = (\theta - \theta_0)^T \mathbf{S} + r_\theta, \tag{2.2}$$

where θ_0 is the true parameter, $\|r_\theta\| = o(|\theta - \theta_0|)$ as $\theta \to \theta_0$, $P\mathbf{S} = 0$ and $\mathbf{I} = 4P(\mathbf{S}\mathbf{S}^T)$ is positive definite. The DQM condition (2.2) holds very generally and is one of the best known regularity conditions for local asymptotic normality of the model. However, if Θ_0 contains more than one point, the DQM condition is no longer feasible. In this case, we use the generalized differentiable in quadratic mean (GDQM) expansion:

DEFINITION 2.2 (GDQM expansion). A trio $(S_\theta, \sigma(\theta), R_\theta)$ is said to satisfy the GDQM expansion if for some $\varepsilon > 0$ and all $\theta \in \Theta_\varepsilon$, we have $PS_\theta = PR_\theta = 0$, $\sigma(\theta) > 0$ and

$$h_\theta = \sigma(\theta) S_\theta - H^2(\theta) + H^2(\theta) R_\theta. \tag{2.3}$$

When the true parameter is unique (denoted by θ_0), the GDQM expansion yields Le Cam's DQM expansion by letting $S_\theta = (\theta - \theta_0)^T \mathbf{S}/|\theta - \theta_0|$, $\sigma(\theta) = |\theta - \theta_0|$ and $R_\theta = r_\theta / H^2(\theta) + 1$. The GDQM expansion always exists and is not unique.

For instance, let

$$S_\theta = \sqrt{2}\frac{h_\theta + H^2(\theta)}{H(\theta)}, \qquad \sigma(\theta) = \frac{H(\theta)}{\sqrt{2}}, \qquad R_\theta = 0;$$

or

$$S_\theta = \frac{l_\theta - 1}{D(\theta)}, \qquad \sigma(\theta) = \frac{D(\theta)}{2}, \qquad R_\theta = 1 - \frac{h_\theta^2}{2H^2(\theta)}.$$

Many different choices for the GDQM expansion are equivalent in the sense that, under suitable conditions, they yield the same limiting distribution of the LRTS.

When the leading term in (2.3), $\sigma(\theta)$, has the same order as the Hellinger distance $H(\theta)$, the GDQM expansion yields a useful quadratic expansion of $H(\theta)$. It is for this reason we assume that $\sigma(\theta)/H(\theta)$ and $H(\theta)/\sigma(\theta)$ are uniformly bounded on Θ_ε in the following GDQM condition.

DEFINITION 2.3 (GDQM condition). The GDQM expansion $(S_\theta, \sigma(\theta), R_\theta)$ is said to satisfy the GDQM condition if $\sup_{\theta\in\Theta_\varepsilon} \sigma(\theta)/H(\theta) < \infty$, $\sup_{\theta\in\Theta_\varepsilon} H(\theta)/\sigma(\theta) < \infty$ and $\sup_{\theta\in\Theta_\varepsilon}(S_\theta^2 + |R_\theta|) \in \mathcal{L}^1(P)$ for some $\varepsilon > 0$.

Since $H(\theta)$ is assumed continuous and bounded, the GDQM condition ensures that $\sup_{\theta\in\Theta_\varepsilon} \sigma(\theta) < \infty$. When a trio $(S_\theta, \sigma(\theta), R_\theta)$ satisfies the GDQM condition, let $S'_\theta = [\sigma(\theta)/H(\theta)]S_\theta$, $\sigma'(\theta) = H(\theta)$ and $R'_\theta = R_\theta$. Then $(S'_\theta, \sigma'(\theta), R'_\theta)$ also satisfies the GDQM condition. The results of this paper which are valid for $(S_\theta, \sigma(\theta), R_\theta)$ also hold for $(S'_\theta, \sigma'(\theta), R'_\theta)$, and vice versa. For convenience, we assume $\sigma(\theta) = H(\theta)$ in the following sections.

2.3. *Quadratic approximations of log-likelihood ratio functions.* In this subsection, we obtain quadratic approximations of $L_n(\theta) \vee 0$ under the GDQM condition. We first study a quadratic approximation of $L_n(\theta)$ in the neighborhood $\Theta_{c/\sqrt{n}}$ for $c > 0$.

THEOREM 2.1. *Assume that $(S_\theta, H(\theta), R_\theta)$ satisfies the GDQM condition and for all $c > 0$,*

$$\sup_{\theta\in\Theta_{c/\sqrt{n}}} |v_n(S_\theta)| = O_P(1), \qquad \sup_{\theta\in\Theta_{c/\sqrt{n}}} |P_n(R_\theta)| = o_P(1). \tag{2.4}$$

Then, as $n \to \infty$, in probability,

$$\sup_{\theta\in\Theta_{c/\sqrt{n}}} \left|L_n(\theta) - 2\sqrt{n}H(\theta)v_n(S_\theta) + nH^2(\theta)[2 + P_n(S_\theta^2)]\right| \to 0.$$

To prove Theorem 2.1, we need the following fact whose proof is omitted.

LEMMA 2.1. *Suppose $X_1, \ldots, X_n$ are i.i.d. random variables and $G(X_i) \in \mathcal{L}^q$ for some $q > 0$. Then*

$$\max_{1 \le i \le n} |G(X_i)| = o_P(n^{1/q}).$$

PROOF OF THEOREM 2.1. We first consider the approximation to $P_n(h_\theta^2)$. The GDQM expansion yields

$$\begin{aligned} nP_n(h_\theta^2) &= nH^2(\theta)P_n(S_\theta^2) + 2nH^3(\theta)P_n[S_\theta(R_\theta - 1)] \\ &\quad + nH^4(\theta)P_n[(1 - R_\theta)^2]. \end{aligned} \tag{2.5}$$

Let $G = \sup_{\theta \in \Theta_\varepsilon}(S_\theta^2 + |R_\theta|)$ for some small enough $\varepsilon > 0$. Then $G \in \mathcal{L}^1(P)$, so by Lemma 2.1, $\max_{1 \le i \le n} G(X_i) = o_P(n)$. Consequently,

$$\max_{1 \le i \le n} \sup_{\theta \in \Theta_\varepsilon} |S_\theta(X_i)| = o_P(n^{1/2}), \qquad \max_{1 \le i \le n} \sup_{\theta \in \Theta_\varepsilon} |R_\theta(X_i)| = o_P(n). \tag{2.6}$$

The second and third terms in (2.5) can be bounded as follows:

$$\sup_{\theta \in \Theta_\varepsilon} |P_n[S_\theta(R_\theta - 1)]| = P_n(G + 1) \max_{1 \le i \le n} \sup_{\theta \in \Theta_\varepsilon} |S_\theta(X_i)| = o_P(n^{1/2}),$$

$$\sup_{\theta \in \Theta_\varepsilon} P_n[(1 - R_\theta)^2] = P_n(G + 1) \max_{1 \le i \le n} \sup_{\theta \in \Theta_\varepsilon} \big(|R_\theta(X_i)| + 1\big) = o_P(n).$$

Hence, $nP_n(h_\theta^2) = nH^2(\theta)P_n(S_\theta^2) + o_P(1)$, where $o_P(1)$ holds uniformly for $\theta \in \Theta_{c/\sqrt{n}}$. Similarly, (2.4) and (2.6) yield

$$\max_{1 \le i \le n} \sup_{\theta \in \Theta_{c/\sqrt{n}}} |h_\theta(X_i)| = o_P(1).$$

Note that

$$\begin{aligned} nP_n(h_\theta) &= \sqrt{n}H(\theta)\nu_n(S_\theta) - nH^2(\theta) + nH^2(\theta)P_n(R_\theta) \\ &= \sqrt{n}H(\theta)\nu_n(S_\theta) - nH^2(\theta) + o_P(1). \end{aligned} \tag{2.7}$$

Using the Taylor expansion of $2\log(1 + x) = 2x - x^2(1 + o(1))$ for small x, we have uniformly for $\theta \in \Theta_{c/\sqrt{n}}$,

$$\begin{aligned} L_n(\theta) &= 2\sum_{i=1}^{n} \log\big(1 + h_\theta(X_i)\big) = nP_n\big(2h_\theta - [1 + o_P(1)]h_\theta^2\big) \\ &= 2\sqrt{n}H(\theta)\nu_n(S_\theta) - nH^2(\theta)[2 + P_n(S_\theta^2)] + o_P(1). \end{aligned}$$

This completes the proof of Theorem 2.1. □

It is important to observe that $L_n(\theta)$ may diverge to $-\infty$ for some $\theta \in \Theta_\varepsilon$. Consequently, it is very difficult to find a general approximation of $L_n(\theta)$ with

a uniform residual term $o_P(1)$ on Θ_ε. Since $2\lambda_n \geq 0$, it suffices to maximize $L_n(\theta) \vee 0$, which has a general quadratic approximation as shown in the following Theorem 2.2. By expanding $L_n(\theta) \vee 0$ locally in a Hellinger neighborhood of the true distribution, we circumvent the difficulties encountered in the classical approaches of Chernoff (1954) and Le Cam (1970) and are able to characterize the asymptotic null distribution of the LRT under loss of identifiability.

THEOREM 2.2. *Assume that* $(S_\theta, H(\theta), R_\theta)$ *satisfies the GDQM condition, and for some* $\varepsilon > 0$ *and all* $c > 0$,

$$\sup_{\theta\in\Theta_\varepsilon} |\nu_n S_\theta| = O_P(1), \qquad \sup_{\theta\in\Theta_{c/\sqrt{n}}} |P_n(R_\theta)| = o_P(1), \tag{2.8}$$
$$\inf_{\theta\in\Theta_\varepsilon} (1 - P_n R_\theta) > 0.$$

Then uniformly for $\theta \in \Theta_\varepsilon$,

$$L_n(\theta) \vee 0 = \big(2\sqrt{n}H(\theta)\nu_n S_\theta - nH^2(\theta)[2 + P_n(S_\theta^2)]\big) \vee 0 + o_P(1). \tag{2.9}$$

PROOF. It suffices to show that (2.9) holds uniformly for $\theta \in A_{1,n} \cup A_{2,n}$, where

$$A_{1,n} = \{\theta \in \Theta_\varepsilon : L_n(\theta) > 0\},$$
$$A_{2,n} = \big\{\theta \in \Theta_\varepsilon : 2\sqrt{n}H(\theta)\nu_n S_\theta - nH^2(\theta)[2 + P_n(S_\theta^2)] > 0\big\},$$

because, otherwise, both sides of (2.9) are 0. Applying the inequality $\log(1 + x) \leq x$ to the log-likelihood ratio function yields

$$\begin{aligned} L_n(\theta) &= 2\sum_{i=1}^{n} \log\big(1 + h_\theta(X_i)\big) \\ &\leq 2nP_n(h_\theta) \\ &= 2\sqrt{n}H(\theta)\nu_n S_\theta - 2nH^2(\theta)(1 - P_n R_\theta). \end{aligned}$$

The inequality above and (2.8) yield $\sup_{\theta\in A_{1,n}} H(\theta) = O_P(n^{-1/2})$. It can be shown in a similar fashion that $\sup_{\theta\in A_{2,n}} H(\theta) = O_P(n^{-1/2})$. Thus, for $\delta > 0$, we can find constants $c > 0$ and $N(\delta, M)$ such that

$$P\left\{\sup_{\theta\in A_{1,n}\cup A_{2,n}} H(\theta) > cn^{-1/2}\right\} < \delta/2, \tag{2.10}$$

for all $n \geq N(\delta, M)$. By Theorem 2.1, given δ and c, for any $\varepsilon > 0$, we can find $N(\delta, \varepsilon, c)$ such that

$$\begin{aligned} P\left\{\sup_{\theta\in\Theta_{c/\sqrt{n}}} \big|L_n(\theta) - \big\{2\sqrt{n}H(\theta)\nu_n S_\theta - nH^2(\theta)[2 + P_n(S_\theta^2)]\big\}\big| \geq \varepsilon\right\} \\ \leq \delta/2, \end{aligned} \tag{2.11}$$

for all $n \geq N(\delta, \varepsilon, c)$. Then (2.10) and (2.11) yield that

$$P\left\{ \sup_{\theta \in A_n} \left| L_n(\theta) - \left\{ 2\sqrt{n} H(\theta) \nu_n S_\theta - n H^2(\theta)[2 + P_n(S_\theta^2)] \right\} \right| \geq \varepsilon \right\} \leq \delta,$$

for all $n \geq \max(N(\delta, \varepsilon), N(\delta, \varepsilon, c))$. This completes the proof of Theorem 2.2. □

2.4. *Asymptotic null distribution of the LRT.* In this section, we derive the asymptotic null distribution of the LRT based on the quadratic approximation of the log-likelihood ratio function obtained in the previous section. Direct maximization of the quadratic form in (2.9) by $\sqrt{n} H(\theta)$ yields

$$(\nu_n S_\theta \vee 0)^2 / (2 + P_n S_\theta^2) \approx (\nu_n S_\theta^* \vee 0)^2 / 2,$$

where S_θ^*, which standardizes S_θ, is defined as

$$S_\theta^* = S_\theta \Big/ \sqrt{1 + P S_\theta^2 / 2}. \tag{2.12}$$

Define $\mathcal{F}$ as the set of all $\mathcal{L}^2$ limits (cluster points) of $\nu_n S_\theta^*$ as $H(\theta) \to 0$, that is,

$$\mathcal{F} = \Bigg\{ S \in \mathcal{L}^2(P) : \exists \{\theta^m\} \in \Theta_\varepsilon \text{ s.t. } \lim_{m \to \infty} H(\theta^m) = \lim_{m \to \infty} \left\| S_{\theta^m} \Big/ \sqrt{1 + P S_{\theta^m}^2 / 2} - S \right\| = 0 \Bigg\}. \tag{2.13}$$

Throughout this paper, we assume that, in (2.13) the $\mathcal{L}^2$ convergence implies pointwise convergence; that is, there exist $\mathcal{L}^2$ representations of S and $\{S_{\theta^m}^*\}$ in (2.13), such that $S_{\theta^m}^*$ converges to S pointwise. Furthermore, we assume that $\mathcal{F}$ is *complete* and admits *continuous paths* as follows.

DEFINITION 2.4. $\mathcal{F}$ is complete if for any sequence $\{\theta^m\} \in \Theta_\varepsilon$ ($\varepsilon > 0$) with $H(\theta^m) \to 0$, there exists a subsequence $\{\theta^{m_k}\}$ of $\{\theta^m\}$ such that $S_{\theta^{m_k}}^*$ converges to some $S \in \mathcal{F}$ in $\mathcal{L}^2$. $\mathcal{F}$ admits continuous paths if for all $S \in \mathcal{F}$, there exists a path $\{\theta(t, S) : 0 < t \leq \varepsilon\} \subset \Theta_\varepsilon$ such that $\theta(t, S)$ is continuous in t, $H(\theta(t, S)) = t$ and $\lim_{t \to 0} S_{\theta(t,S)}^* = S$ in $\mathcal{L}^2$.

Define $D_{\mathcal{F}}(S_\theta^*)$ as the $\mathcal{L}^2$ distance between S_θ^* and $\mathcal{F}$; that is, $D_{\mathcal{F}}(S_\theta^*) = \inf_{S \in \mathcal{F}} \|S - S_\theta^*\|$. The completeness of $\mathcal{F}$ immediately yields the following lemma, whose proof will be omitted.

LEMMA 2.2. *If $\mathcal{F}$ is complete, then* $\limsup_{H(\theta) \to 0} D_{\mathcal{F}}(S_\theta^*) = 0$.

According to the GDQM condition, S_θ has a square integrable envelope function on Θ_ε. Then the dominated convergence theorem implies that, when $\mathcal{F}$ is complete, $\mathcal{F}$ is also compact and $PS = 0$ for all $S \in \mathcal{F}$. Moreover, there exists a function in $\mathcal{F}$, denoted by $\widetilde{S}_\theta^*$, achieving the minimum distance $D_{\mathcal{F}}(S_\theta^*)$; that is,

$$\|S_\theta^* - \widetilde{S}_\theta^*\| = D_{\mathcal{F}}(S_\theta^*). \tag{2.14}$$

By Lemma 2.2, $\|S_\theta^* - \widetilde{S}_\theta^*\|$ converges to 0 uniformly as $H(\theta) \to 0$.

THEOREM 2.3. *Assume that*

(a) $(S_\theta, H(\theta), R_\theta)$ *satisfies the GDQM condition*.
(b) *Equations* (2.8) *in Theorem* 2.2 *hold*.
(c) $\mathcal{F}$ *in* (2.13) *is complete and admits continuous paths*.
(d) *On the same probability space as the empirical process* ν_n, *there exists a centered Gaussian process* $\{W_S : S \in \mathcal{F}\}$ *with uniformly continuous sample paths*; *that is*, *in probability*

$$\limsup_{\|S_1 - S_2\| \to 0} |W_{S_1} - W_{S_2}| = 0;$$

and covariance kernel

$$P(W_{S_1} W_{S_2}) = P(S_1 S_2) \qquad \forall\, S_1, S_2 \in \mathcal{F},$$

such that, *for all* $c > 0$, *the following two conditions hold*:

$$\sup_{\theta \in \Theta_{c/\sqrt{n}}} |\nu_n S_\theta^* - W_{\tilde{S}_\theta^*}| = o_P(1), \qquad \sup_{\theta \in \Theta_{c/\sqrt{n}}} |P_n S_\theta^2 - P S_\theta^2| = o_P(1). \tag{2.15}$$

Under the assumptions (a)–(d), *the LRTS for* (2.1) *satisfies*

$$\lim_{n \to \infty} 2\lambda_n = \sup_{S \in \mathcal{F}} (W_S \vee 0)^2.$$

PROOF. Since $(S_\theta, H(\theta), R_\theta)$ satisfies the GDQM condition and (2.8) holds, Theorem 2.2 yields (2.9). From (2.10) and (2.15),

$$\sup_{\{\theta : L_n(\theta) > 0\}} \nu_n S_\theta^* \vee 0 \le \sup_{S \in \mathcal{F}} W_S \vee 0 + o_P(1).$$

Then by Theorem 2.2,

$$\begin{aligned}
2\lambda_n &= \sup_{\{\theta : L_n(\theta) > 0\}} \left(4\sqrt{n} H(\theta) \nu_n S_\theta - 2nH^2(\theta)[2 + P_n(S_\theta^2)]\right) \vee 0 + o_P(1) \\
&= \sup_{\{\theta : L_n(\theta) > 0\}} \Big[2\sqrt{2n} H(\theta) \sqrt{2 + P S_\theta^2}\, \nu_n S_\theta^* \\
&\qquad\qquad - 2nH^2(\theta)(2 + P S_\theta^2)\Big] \vee 0 + o_P(1) \qquad (2.16) \\
&\le \sup_{\{\theta : L_n(\theta) > 0\}} (\nu_n S_\theta^* \vee 0)^2 + o_P(1) \\
&\le \sup_{S \in \mathcal{F}} (W_S \vee 0)^2 + o_P(1).
\end{aligned}$$

Thus to prove the theorem, it suffices to show that $2\lambda_n \ge \sup_{S \in \mathcal{F}} (W_S \vee 0)^2 + o_P(1)$, which is equivalent to showing that for $\delta > 0$, there exists a constant $N(\delta)$ such that

$$P\left(2\lambda_n \ge \sup_{S \in \mathcal{F}} (W_S \vee 0)^2 - \delta\right) \ge 1 - \delta, \tag{2.17}$$

for all $n \geq N(\delta)$. From previous discussions, $\mathcal{F}$ is compact and W_S has uniformly continuous sample paths on $(\mathcal{F}, \|\cdot\|)$. Hence we can assume that for some small $\eta > 0$,

$$P\Big(\sup_{\|S_1 - S_2\| \leq \eta} \big|W_{S_1} - W_{S_2}\big| \geq \delta/2\Big) \leq \delta/2.$$

Note that $\sup_{S\in\mathcal{F}} |W_S| = O_P(1)$ and $W_S \to (W_S \vee 0)^2$ is a continuous map. Without loss of generality, we can assume

$$P\Big(\sup_{\|S_1 - S_2\| \leq \eta} \big|(W_{S_1} \vee 0)^2 - (W_{S_2} \vee 0)^2\big| \geq \delta/2\Big) \leq \delta/2. \tag{2.18}$$

Since $\mathcal{F}$ is compact, there exists an η-net $\{S_1, \ldots, S_k\}$ on $\mathcal{F}$. Denote by $S_{i(S)}$ the closest point in the η-net to $S \in \mathcal{F}$. Note that $\|S - S_{i(S)}\| \leq \eta$. Therefore, to prove (2.17), it suffices to show that, for $\delta > 0$, there exists a constant $N(\delta)$ such that for all $n \geq N(\delta)$,

$$P\Big(2\lambda_n \geq \max_{1\leq i\leq k} \big(W_{S_i} \vee 0\big)^2 - \delta/2\Big) \geq 1 - \delta/2. \tag{2.19}$$

Clearly, (2.19) holds if we can prove that for all $S \in \mathcal{F}$,

$$2\lambda_n \geq (W_S \vee 0)^2 + o_P(1).$$

Without loss of generality, we assume $W_S > 0$. According to the assumptions, S has a continuous path in θ_ε, denoted by $\{\theta(t, S)\}$. Then $\lim_{t\to 0} \|S^*_{\theta(t,S)} - S\| = 0$ and $H(\theta(t, S)) = t$. With probability going to 1, the equation

$$t\sqrt{2n\big(2 + PS^2_{\theta(t,S)}\big)} = W_S \tag{2.20}$$

has a solution with $\theta(t, S) \in \Theta_\varepsilon$. To prove this, define $g(t) = t\sqrt{2n(2 + PS^2_{\theta(t,S)})}$. Note that PS^2_θ is bounded. Then $g(0) = 0 < W_S$ and $g(\varepsilon) > 2\varepsilon\sqrt{2n} > W_S$ with probability going to 1. Since $g(t)$ is continuous, (2.20) has a solution with probability going to 1. Denote the solution by t_n and $\theta(t_n, S)$ by θ^n. Note that $\lim_{n\to\infty} H(\theta^n) = 0$, so $\lim_{n\to\infty} \|S^*_{\theta^n} - S\| = 0$. The triangle inequality then yields $\lim_{n\to\infty} \|\widetilde{S}^*_{\theta^n} - S\| = 0$. Since the Gaussian process $\{W_S : S \in \mathcal{F}\}$ has uniformly continuous sample paths,

$$W_{\widetilde{S}^*_{\theta^n}} - W_S = o_P(1).$$

Note that $H(\theta^n) = O_P(n^{-1/2})$. Then (2.15) yields

$$\nu_n S^*_{\theta^n} = W_{\widetilde{S}^*_{\theta^n}} + o_P(1) = W_S + o_P(1).$$

Plugging the equation $H(\theta^n)\sqrt{2n(2+PS_{\theta^n}^2)} = W_S$ into (2.16),

$$\begin{aligned}2\lambda_n &\geq \Big(2\Big[H(\theta^n)\sqrt{2n(2+PS_{\theta^n}^2)}\Big]\nu_n S_{\theta^n}^* - 2nH^2(\theta^n)(2+PS_{\theta^n}^2)\Big) \vee 0 + o_P(1)\\ &= 2W_S\nu_n S_{\theta^n}^* - W_S^2 + o_P(1)\\ &= W_S^2 + o_P(1).\end{aligned}$$

Note that $W_S > 0$. This completes the proof of Theorem 2.3. □

The GDQM expansion used by the above theorems assumes that $\sigma(\theta) = H(\theta)$. For the general GDQM expansion $(S_\theta, \sigma(\theta), R_\theta)$, one can simply replace S_θ by $[\sigma(\theta)/H(\theta)]S_\theta$ and obtain similar results. The general form of standardized score function in Theorem 2.3, S_θ^*, is then

$$S_\theta^* = \sigma(\theta)S_\theta \big/ \sqrt{H^2(\theta) + \sigma^2(\theta)P(S_\theta^2)/2}. \tag{2.21}$$

$\mathcal{F}$ and $\widetilde{S}_\theta^*$ can be redefined in an analogous fashion to (2.13) and (2.14), respectively. We state the following theorem without proof.

THEOREM 2.4. *Assume that* $(S_\theta, \sigma(\theta), R_\theta)$ *satisfies the GDQM condition and that assumption* (2.8) *in Theorem* 2.2 *holds. Then uniformly for* $\theta \in \Theta_\varepsilon$,

$$L_n(\theta) \vee 0 = \big(2\sqrt{n}\sigma(\theta)\nu_n S_\theta - n[2H^2(\theta) + \sigma^2(\theta)P_n(S_\theta^2)]\big) \vee 0 + o_P(1).$$

Moreover, under the assumptions (a)–(d) *in Theorem* 2.3, *the LRTS for* (2.1) *satisfies*

$$\lim_{n\to\infty} 2\lambda_n = \sup_{S\in\mathcal{F}} (W_S \vee 0)^2.$$

3. LRT with square integrable likelihood ratios. For square integrable likelihood ratios, it is more convenient to use the Pearson type $\mathcal{L}^2$ distance $D(\theta)$ instead of the Hellinger distance $H(\theta)$. We call $S_\theta = (l_\theta - 1)/D(\theta)$ the *generalized score function*. Then $PS_\theta = 0$ and $PS_\theta^2 = 1$. One choice of the GDQM expansion for square integrable likelihood ratios is $S_\theta = (l_\theta - 1)/D(\theta)$, $\sigma = D(\theta)/2$ and $R(\theta) = 1 - h_\theta^2/[2H^2(\theta)]$. The standardized score function in (2.21) becomes

$$S_\theta^* = \Big[D(\theta)\big/\sqrt{4H^2(\theta) + D^2(\theta)/2}\Big]S_\theta.$$

Under suitable conditions, we can prove that $D(\theta)/\sqrt{4H^2(\theta) + D^2(\theta)/2} \to 1$ as $D(\theta) \to 0$. Thus, S_θ^* and S_θ are equivalent in the sense that they yield the same $\mathcal{L}^2$ limit. Then $\mathcal{F}$ can be formulated as the set of limits of generalized score functions:

$$\begin{aligned}\mathcal{F} = \Big\{ & S \in \mathcal{L}^2 : \exists\, \{\theta^m\} \in \Theta_\varepsilon, \text{ s.t.}\\ & \lim_{m\to\infty} D(\theta^m) = 0, \lim_{m\to\infty} \|(l_{\theta^m} - 1)/D(\theta^m) - S\| = 0\Big\}.\end{aligned} \tag{3.1}$$

3.1. *Quadratic approximation and asymptotic null distribution.* In this subsection, we extend the asymptotic results obtained in Section 2 to square integrable likelihood ratios. First we present a lemma on the local equivalence between $H(\theta)$ and $D(\theta)$ as $D(\theta) \to 0$.

LEMMA 3.1. *If the generalized score function,* $S_\theta = (l_\theta - 1)/D(\theta)$, *has a square integrable envelope function on* Θ_ε *for some* $\varepsilon > 0$, *then*

$$\lim_{D(\theta)\to 0} 8H^2(\theta)/D^2(\theta) = \lim_{H(\theta)\to 0} 8H^2(\theta)/D^2(\theta) = 1.$$

PROOF. Let $G = \sup_{\theta\in\Theta_\varepsilon} |S_\theta|$. Then $G \in \mathcal{L}^2(P)$. It is straightforward to verify that

$$\begin{aligned} D^{-2}(\theta)|8H^2(\theta) - D^2(\theta)| &= D^{-2}(\theta)|P\big((\sqrt{l_\theta}+3)(\sqrt{l_\theta}-1)^3\big)| \\ &\le 3P\big((G+1)^2|\sqrt{l_\theta}-1|(\sqrt{l_\theta}+1)^{-1}\big). \end{aligned}$$

When $D(\theta) \to 0$ or $H(\theta) \to 0$, $\sqrt{l_\theta} \to 1$ in probability. Since $(G+1)^2|\sqrt{l_\theta}-1| \times (\sqrt{l_\theta}+1)^{-1}$ is dominated by $(G+1)^2$, the dominated convergence theorem yields Lemma 3.1. □

To apply Theorem 2.3, one needs to verify the conditions for the convergence of $\nu_n S_\theta^*$ and $P_n S_\theta^2$ in a $n^{-1/2}$-Hellinger neighborhood of the true distribution. For square integrable likelihood ratios, these conditions are directly implied by the P-Donsker class and Glivenko–Cantelli class conditions. Next we give their definitions. For further details, see van der Vaart and Wellner (1996) and Dudley (1999).

DEFINITION 3.1. A family of measurable functions $\mathcal{G} \in \mathcal{L}^1(P)$ forms a Glivenko–Cantelli class, written as $\mathcal{G} \in GC(P)$, if in probability $\sup_{g\in\mathcal{G}} |P_n g - Pg| \to 0$ as $n \to \infty$.

DEFINITION 3.2. A family of measurable functions $\mathcal{G} \in \mathcal{L}^2(P)$ forms a P-Donsker class, if, on some suitable probability space, there exists a version $\tilde{\nu}_n$ of the empirical process ν_n and a centered Gaussian process $\{W_g, g \in \mathcal{G}\}$ with covariance kernel

$$\operatorname{cov}(W_{g_1}, W_{g_2}) = Pg_1g_2 - Pg_1Pg_2 \qquad \text{for } g_1, g_2 \in \mathcal{G}$$

such that in probability,

$$\sup_{g\in\mathcal{G}} |\tilde{\nu}_n(g) - W_g| \to 0,$$

as n tends to infinity, where the Gaussian process $\{W_g, g \in \mathcal{G}\}$ has continuous sample paths with respect to the $\mathcal{L}^2$ distance $e(\cdot,\cdot)$ on $\mathcal{G}: e(g_1, g_2) = P(g_1 - g_2)^2 - (Pg_1 - Pg_2)^2$.

Since there exists a probability space holding both the Gaussian process $\{W_S : S \in \mathcal{F}\}$ and the empirical process ν_n, we simply denote by P their common probability measure. Next we present the main theorem of this section.

THEOREM 3.1. *Assume, for some* $\varepsilon > 0$, $\mathcal{F}_\varepsilon = \{S_\theta = (l_\theta - 1)/D(\theta) : \theta \in \Theta_\varepsilon\}$ *forms a* P*-Donsker class with a square integrable envelope function. Then uniformly for* $\theta \in \Theta_\varepsilon$,

$$2L_n(\theta) \vee 0 = \big(2\sqrt{n}D(\theta)\nu_n S_\theta - nD^2(\theta)\big) \vee 0 + o_P(1). \tag{3.2}$$

Moreover, assume the set $\mathcal{F}$ *in* (3.1) *is complete and admits continuous paths. Then on some probability space, there exists a centered Gaussian process* $\{W_S : S \in \mathcal{F}\}$, *equipped with the same probability measure* P *as* ν_n, *with continuous sample paths and covariance kernel,* $P(W_{S_1}W_{S_2}) = PS_1S_2$, *for* $S_1, S_2 \in \mathcal{F}$. *The LRTS for* (2.1) *satisfies*

$$\lim_{n\to\infty} 2\lambda_n = \sup_{S\in\mathcal{F}} (W_S \vee 0)^2.$$

PROOF. First we prove that the GDQM condition is satisfied for the expansion $S_\theta = (l_\theta - 1)/D(\theta)$, $\sigma(\theta) = D(\theta)/2$, and $R_\theta = 1 - h_\theta^2/[2H^2(\theta)]$. Then we obtain (3.2) by verifying condition (2.8) of Theorem 2.2.

By Lemma 3.1, it is clear that there exists a constant $c \in (0,1)$ such that $c \le D(\theta)/H(\theta) \le 1/c$ for $\theta \in \Theta_\varepsilon$. Denote the envelope function of S_θ on Θ_ε by G. Then

$$\begin{aligned} S_\theta^2 + |R_\theta| &\le S_\theta^2 + \frac{(l_\theta - 1)^2}{2H^2(\theta)} + 1 \\ &\le S_\theta^2 + \frac{D^2(\theta)}{2H^2(\theta)} S_\theta^2 + 1 \le \frac{1 + 2c^2}{2c^2} G^2 + 1 \in \mathcal{L}^1(P). \end{aligned}$$

Thus $(S_\theta, \sigma(\theta), R_\theta)$ satisfies the GDQM condition. The P-Donsker class condition implies $\sup_{\theta\in\Theta_\varepsilon} |\nu_n S_\theta| = O_P(1)$. To prove (3.2), by Theorem 2.2, it suffices to show that

$$\sup_{\theta\in\Theta_{c/\sqrt{n}}} |P_n(R_\theta)| = o_P(1) \quad \text{and} \quad \inf_{\theta\in\Theta_\varepsilon} (1 - P_n R_\theta) > 0.$$

The proof for $\sup_{\theta\in\Theta_{c/\sqrt{n}}} |P_n(R_\theta)| = o_P(1)$ is presented first. By Lemma 3.1 and the LLN,

$$\begin{aligned} 1 - P_n R_\theta &= \frac{D^2(\theta)}{2H^2(\theta)} P_n\bigg(\frac{(l_\theta - 1)^2}{D^2(\theta)} \frac{1}{(\sqrt{l_\theta} + 1)^2}\bigg) \\ &= P_n\bigg(\frac{4S_\theta^2}{(\sqrt{l_\theta} + 1)^2}\bigg) + o_P(1) \qquad (3.3) \\ &= P_n(S_\theta^2) + P_n\bigg(S_\theta^2(1 - l_\theta)\frac{(\sqrt{l_\theta} + 3)}{(\sqrt{l_\theta} + 1)^3}\bigg) + o_P(1). \qquad (3.4) \end{aligned}$$

By Lemma 2.10.14 in van der Vaart and Wellner (1996), $\{S_\theta^2 : \theta \in \Theta_\varepsilon\}$ forms a Glivenko–Gantelli class. Hence $\sup_{\theta\in\Theta_\varepsilon} |P_n(S_\theta^2) - 1| = o_P(1)$. Note that $\sup_{\theta\in\Theta_{c/\sqrt{n}}} D(\theta) = O(n^{-1/2})$ and $\max_{1\le i\le n} G(X_i) = o_P(n^{1/2})$. Thus

$$\max_{1\le i\le n} \sup_{\theta\in\Theta_{c/\sqrt{n}}} |l_\theta(X_i) - 1| \le \max_{1\le i\le n} G(X_i) \sup_{\theta\in\Theta_{c/\sqrt{n}}} D(\theta) = o_P(1).$$

Then (3.4) yields

$$\sup_{\theta\in\Theta_{c/\sqrt{n}}} |P_n R_\theta| \le o_P(1) + o_P(1) \sup_{\theta\in\Theta_{c/\sqrt{n}}} P_n\left(G^2 \frac{(\sqrt{l_\theta}+3)}{(\sqrt{l_\theta}+1)^3}\right) = o_P(1).$$

Next we verify that $\inf_{\theta\in\Theta_\varepsilon}(1 - P_n R_\theta) > 0$. By (3.3), for $k > 0$,

$$\begin{aligned} 1 - P_n(R_\theta) &\ge \frac{c^2}{2} P_n\left(\frac{S_\theta^2}{l_\theta + 1}\right) \ge \frac{1}{4c^2(k+1)} P_n\big(\mathbb{1}_{\{l_\theta \le k\}} S_\theta^2\big) \\ &\ge \frac{1}{4c^2(k+1)}\big[P_n(S_\theta^2) - P_n\big(\mathbb{1}_{\{l_\theta > k\}} S_\theta^2\big)\big]. \end{aligned}$$

It is clear that almost surely,

$$\lim_{k\to\infty} \sup_{\theta\in\Theta_\varepsilon} P_n\big(\mathbb{1}_{\{|l_\theta|>k\}} S_\theta^2\big) \le \lim_{k\to\infty} P_n\big(\mathbb{1}_{\{(G+1)>k/c\}} G^2\big) = 0.$$

We can choose k so large that in probability $\sup_{\theta\in\Theta_\varepsilon} P_n(\mathbb{1}_{\{|l_\theta|>k\}} S_\theta^2) < 1/4$. By the definition of Glivenko–Cantelli class, $\inf_{\theta\in\Theta_\varepsilon} P_n(S_\theta^2) > 1/2$ with probability going to 1. Thus $\inf_{\theta\in\Theta_\varepsilon}(1 - P_n R_\theta) > 1/[16c^2(k+1)] > 0$ with probability going to 1. Applying Theorem 2.4 yields that uniformly for $\theta \in \Theta_\varepsilon$,

$$\begin{aligned} 2L_n(\theta) \vee 0 &= \big(2\sqrt{n} D(\theta) \nu_n S_\theta - 2n[H^2(\theta) + D^2(\theta) P_n({S_\theta}^2)/2]\big) \vee 0 + o_P(1) \\ &= \big(2\sqrt{n} D(\theta) \nu_n S_\theta - n D^2(\theta)\big) \vee 0 + o_P(1). \end{aligned}$$

In the last equation we used the fact that $8H^2(\theta)/D^2(\theta) = 1 + o_P(1)$, because Theorem 2.2 ensures that $\sup_{\{\theta\,:\,L_n(\theta)\ge 0\}} H(\theta) = O_P(n^{-1/2})$.

Lemma 3.1 ensures that S_θ^* and S_θ have the same $\mathcal{L}^2$ limits as $D(\theta)$ tends to 0. Therefore, the definitions of $\mathcal{F}$ in (2.13) and (3.1) are equivalent. Because $\mathcal{F}$ is complete, it is clear that the closure of $\mathcal{F}_\varepsilon$ satisfies $\bar{\mathcal{F}}_\varepsilon = \mathcal{F}_\varepsilon \cup \mathcal{F}$ and $\mathcal{F}$ is a closed set. Theorem 2.10.2 in van der Vaart and Wellner (1996) yields that $\bar{\mathcal{F}}_\varepsilon$ is a P-Donsker class. Hence, on the same probability space as the empirical process ν_n, there exists a centered Gaussian process $\{W_S : S \in \bar{\mathcal{F}}_\varepsilon\}$ as defined in Definition 3.2 with continuous sample paths and covariance kernel,

$$P(W_{S_1} W_{S_2}) = P S_1 S_2 \qquad \forall\, S_1, S_2 \in \bar{\mathcal{F}}_\varepsilon.$$

Finally, we will verify (2.15) and thereby obtain the asymptotic null distribution of the LRTS by Theorem 2.3. Note that $\{S_\theta^2;\ \theta \in \Theta_\varepsilon\} \in GC(P)$. It directly yields

$\sup_{\theta\in\Theta_\varepsilon}|P_nS_\theta^2 - PS_\theta^2| = o_P(1)$ in (2.15). Next we consider $\nu_nS_\theta^*$. Note that S_θ^*/S_θ converges to 1 as $H(\theta)$ tends to 0. We have $\sup_{\theta\in\Theta_{c/\sqrt{n}}}\|S_\theta - S_\theta^*\| = o(1)$ and $\sup_{\theta\in\Theta_{c/\sqrt{n}}}|\nu_nS_\theta - \nu_nS_\theta^*| = o_P(1)$. By Lemma 2.2 and (2.14), there is a function $\widetilde{S}_\theta^*$ such that $\|S_\theta^* - \widetilde{S}_\theta^*\| = o(1)$ as $H(\theta)$ tends to 0. Then the triangle inequality and the uniform continuity of the Gaussian process $\{W_S : S \in \bar{\mathcal{F}}_\varepsilon\}$ yield that

$$\sup_{\theta\in\Theta_{c/\sqrt{n}}}\|S_\theta - \widetilde{S}_\theta^*\| = o_P(1) \quad \text{and} \quad \sup_{\theta\in\Theta_{c/\sqrt{n}}}|W_{S_\theta} - W_{\widetilde{S}_\theta^*}| = o_P(1).$$

The Donsker class condition ensures $\sup_{\theta\in\Theta_{c/\sqrt{n}}}|\nu_nS_\theta - W_{S_\theta}| = o_P(1)$. The triangle inequality then gives

$$\sup_{\theta\in\Theta_{c/\sqrt{n}}}|\nu_nS_\theta^* - W_{\widetilde{S}_\theta^*}| = o_P(1),$$

thus (2.15). By Theorem 2.3, we complete the proof of Theorem 3.1. □

The P-Donsker class condition for $\{S_\theta; \theta \in \Theta_\varepsilon\}$ in Theorem 3.1 can be verified using the empirical process techniques discussed in van der Vaart and Wellner (1996). For instance, in the next section, we show that the generalized score functions in discrete models form a P-Donsker class. The following lemma is helpful when S_θ is Lipschitz in θ.

LEMMA 3.2. *Assume that $\mathcal{G} = \{g_\theta; \theta \in \Theta\} \subset L^1(P)$. If Θ is a compact set in $\Re^d$ and for any $\theta_1, \theta_2 \in \Theta$, there exists a function $G \in L^2(P)$ such that*

$$|g_{\theta_1}(x) - g_{\theta_2}(x)| \le |\theta_1 - \theta_2|G(x),$$

then $\mathcal{G}$ is a P-Donsker class and $\{g_\theta^2; \theta \in \Theta\} \in GC(P)$.

PROOF. Theorem 2.7.11 in van der Vaart and Wellner (1996) implies $\mathcal{G}$ is a P-Donsker class. By Lemma 6.3.5 and Theorem 6.1.7 in Dudley (1999), $\{g_\theta^2; \theta \in \Theta\} \in GC(P)$. □

To obtain the limiting distribution of the LRTS, one needs to derive the index set $\mathcal{F}$, then find the distribution of the supremum of the Gaussian process W_S on $\mathcal{F}$ [see, e.g., Azaïs, Cierco-Ayrolles and Croquette (1999) and Azaïs and Wschebor (2000)]. This can also be done using simulations. Since the covariance kernel of $\{W_S; S \in \mathcal{F}\}$ is known, the Gaussian process can be simulated via standard Monte Carlo methods. Then we obtain the asymptotic null distribution of the LRTS by maximizing the simulated $(W_S \vee 0)^2$ over $\mathcal{F}$. In the following, we demonstrate how to derive $\mathcal{F}$ in the admixture models. The admixture models have also been studied by many researchers [see, e.g., Dacunha-Castelle and Gassiat (1997) and Lemdani and Pons (1999)].

EXAMPLE 3.1 (Admixture models). The hypothesis testing problem in admixture models is to test

$$H_0: f_0 \quad \text{against} \quad H_1: pf_\phi + (1-p)f_0,\ 0 < p \le 1,\ \phi \in \Phi, \tag{3.5}$$

where f_0 and f_ϕ are density functions, f_0 is known. Denote $\Phi_0 = \{\phi \in \Phi : \|f_\phi - f_0\| = 0\}$ and $\theta = (\phi, p)$. Then, the $\mathcal{L}^2$ distance is $D(\theta) = p\|f_\phi/f_0 - 1\|$ and the generalized score function is

$$S_\theta = (f_\phi/f_0 - 1)/\|f_\phi/f_0 - 1\|,$$

for $\phi \in \Phi \setminus \Phi_0$. Since the generalized score function does not depend on p, we denote it by S_ϕ. Recall that $\Omega(f)$ is defined as $\Omega(f) = f/\|f\|$ for $f \in \mathcal{L}^2(P)$ $(f \neq 0)$. There are two types of limits of the generalized score functions, depending on whether $\|f_{\phi^m} - f_0\|$ converges to zero:

$$\mathcal{F}_1 = \{\Omega(f_\phi/f_0 - 1) : \phi \in \Phi \setminus \Phi_0\},$$

$$\mathcal{F}_2 = \Big\{S : \exists\{\phi^m\} \in \Phi \setminus \Phi_0, \text{ s.t. } \lim_{m\to\infty} \|f_{\phi^m} - f_0\| = 0, \lim_{m\to\infty} \|\Omega(f_{\phi^m} - f_0) - S\| = 0\Big\}.$$

Thus $\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2$. It is clear that $\mathcal{F}_2$ is a subset of the closure of $\mathcal{F}_1$. Therefore, $\mathcal{F} = \bar{\mathcal{F}}_1$. Under the conditions of Theorem 3.1, the Gaussian process $\{W_S;\ S \in \mathcal{F}\}$ has uniformly continuous sample paths. We have

$$\lim_{n\to\infty} 2\lambda_n = \sup_{S\in\bar{\mathcal{F}}_1} (W_S \vee 0)^2 = \sup_{S\in\mathcal{F}_1} (W_S \vee 0)^2.$$

3.2. *Discrete models and composite null hypothesis.* In this subsection, we first present a theorem for testing hypotheses of discrete models. We prove that the Donsker class condition is automatically satisfied by discrete models and obtain an explicit representation of the Gaussian process in Theorem 3.1. Then we consider hypothesis testing problems with composite null hypotheses.

When the distribution function is discrete, without loss of generality, we may assume that the sample space is $\mathcal{X} = \{1, \ldots, k\}$. An observation on $\mathcal{X}$ can be written as a vector $\mathbf{X} = (X(1), \ldots, X(k))$ with a multinomial distribution with probabilities: $\mathbf{f}_\theta = (f_\theta(i))_{i=1}^k$, where $f_\theta(i) = P_\theta(X(i) = 1)$. Denote the true null distribution by $\mathbf{f} = \mathbf{f}_{\theta_0}$ and assume that $f(i) > 0$ for $1 \le i \le k$. Note that a vector $\mathbf{x}$ on $\mathcal{X}$ can be also regarded as a function. Define the $\mathcal{L}^2$ norm $\|\cdot\|$ on $\mathcal{X}$ as $\|\mathbf{x}\|^2 = \sum_{i=1}^k f(i)x(i)^2$. For any $\theta \in \Theta \setminus \Theta_0$, define the *generalized score vector* as $\mathbf{S}_\theta = (\mathbf{f}_\theta/\mathbf{f} - 1)/\|\mathbf{f}_\theta/\mathbf{f} - 1\|$ and the set of limits of $\mathbf{S}_\theta$ as $\mathbf{F}$, that is,

$$\mathbf{F} = \Big\{\mathbf{S} : \exists\{\theta^m\} \in \Theta \setminus \Theta_0 \text{ s.t. } \lim_{m\to\infty} \|\mathbf{f}_{\theta^m} - \mathbf{f}\| = 0, \lim_{m\to\infty} \|\mathbf{S}_{\theta^m} - \mathbf{S}\| = 0\Big\}. \tag{3.6}$$

The following theorem gives the asymptotic null distribution of the LRT for discrete models.

THEOREM 3.2. *Assume that the distributions in* (2.1) *are discrete, and* $\mathbf{F}$ *in* (3.6) *is complete and admits continuous paths. The LRTS for* (2.1) *satisfies*

$$\lim_{n\to\infty} 2\lambda_n = \sup_{\mathbf{S}\in\mathbf{F}} (\mathbf{S}^T\mathbf{W} \vee 0)^2,$$

where $\mathbf{W} \sim N(0, \operatorname{diag}(\mathbf{f}) - \mathbf{f}\mathbf{f}^T)$ *and* $\mathbf{f}$ *is the true null probability vector.*

PROOF. First we verify the P-Donsker condition for discrete models. It suffices to show that the empirical process of the generalized score function, $\nu_n S_\theta$, converges uniformly to some Gaussian process. Based on our notation, the generalized score function S_θ can be expressed as

$$S_\theta(\mathbf{X}) = \left(\prod_{i=1}^{k} [f_\theta(i)/f(i)]^{X(i)} - 1 \right) \Big/ \|\mathbf{f}_\theta/\mathbf{f} - 1\| = \mathbf{S}_\theta^T \mathbf{X}.$$

Consider i.i.d. random observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from the true distribution $\mathbf{f}$. Denote $\mathbf{n} = (n(1), \ldots, n(k))$, where $n(i) = \#\{j : X_j(i) = 1,\ 1 \le j \le n\}$, for $1 \le i \le k$, and let $\mathbf{W}_n = \mathbf{n} - n\mathbf{f}$. Note that $\mathbf{S}_\theta^T \mathbf{f} = 0$, we have $\nu_n S_\theta = \mathbf{S}_\theta^T \mathbf{W}_n$. By the central limit theorem, $\mathbf{W}_n$ converges to a Gaussian random vector $\mathbf{W} \sim N(0, \operatorname{diag}(\mathbf{f}) - \mathbf{f}\mathbf{f}^T)$. Without loss of generality, we assume $\mathbf{W}_n$ converges to $\mathbf{W}$ almost surely [see, e.g., Dudley (1999)]. Note that $\|\mathbf{W}_n - \mathbf{W}\| = o_P(1)$ and $\|\mathbf{S}_\theta\| = 1$. Thus $\nu_n S_\theta$ converges to $\mathbf{S}_\theta^T \mathbf{W}$ uniformly for $\theta \in \Theta \setminus \Theta_0$. This verifies the P-Donsker class condition in Theorem 3.1. For discrete models, it is obvious that the $\mathcal{L}^2$ convergence of $\mathbf{S}_\theta$ is equivalent to the pointwise convergence. By Theorem 3.1, the proof is completed. □

In many hypothesis testing problems, the null hypothesis is composite:

$$H_0 : f_\theta,\ \theta \in \Theta_1, \quad \text{against} \quad H_1 : f_\theta,\ \theta \in \Theta_2 \setminus \Theta_1, \tag{3.7}$$

where f_θ is a density function, Θ_1 and Θ_2 are subsets of $\Re^d$. Without loss of generality, we assume $\Theta_1 \subset \Theta_2$. Denote the true null distribution by f_0. For $i = 1, 2$, let

$$\Theta_{i0} = \{\theta \in \Theta_i : f_\theta = f_0\},$$
$$\Theta_{i\varepsilon} = \{\theta \in \Theta_i : 0 < D(\theta) \le \varepsilon\} \quad \text{and} \quad \lambda_{in} = \sup_{\theta\in\Theta_{i\varepsilon}} L_n(\theta) \vee 0.$$

Then the LRTS for (3.7) can be written as $2\lambda_n^* = 2\lambda_{2n} - 2\lambda_{1n}$. Denote by $\mathcal{F}_i$ the limits of generalized score functions on Θ_i as $D(\theta) \to 0$; that is,

$$\mathcal{F}_i = \Big\{ S : \exists\, \{\theta^m\} \in \Theta_{i\varepsilon} \setminus \Theta_{i0},\ \text{s.t.} \lim_{m\to\infty} \|f_{\theta^m} - f_0\| = 0,\ \lim_{m\to\infty} \|\Omega(f_{\theta^m}/f_0 - 1) - S\| = 0 \Big\}. \tag{3.8}$$

We assume that, for any $S \in \mathcal{F}_2$, it can be decomposed as the linear combination of its projection $S^{\|}$ in $\mathcal{F}_1$ and a function $S^{\perp}$ in $\mathcal{F}_2$ orthogonal to $\mathcal{F}_1$; that is,

$$S = aS^{\|} + bS^{\perp}, \tag{3.9}$$

where $P(S^{\perp}T) = 0$ for all $T \in \mathcal{F}_1$, $a, b \geq 0$. Denote by $\mathcal{F}^{\perp}$ the set of the decomposed functions of $\mathcal{F}_2$ orthogonal to $\mathcal{F}_1$. Under suitable conditions, the LRTS for (3.7) converges in distribution to $\sup_{S\in\mathcal{F}^{\perp}}(W_S \vee 0)^2$. That is the following theorem.

THEOREM 3.3. *For the test in* (3.7), *we assume that for* $i = 1, 2$, $\mathcal{F}_{i\varepsilon} = \{S_\theta : \theta \in \Theta_{i\varepsilon}\}$ *is a* P*-Donsker class with square integrable envelope function for some* $\varepsilon > 0$, *and* $\mathcal{F}_i$ *defined in* (3.8) *is complete and admits continuous paths, and* $\mathcal{F}_2$ *is convex. We also assume that each* $S \in \mathcal{F}_2$ *has the decomposition* (3.9). *Then on a suitable probability space, there exists a centered Gaussian process* $\{W_S : S \in \mathcal{F}^{\perp}\}$, *equipped with the same probability measure as* ν_n, *with continuous sample paths and covariance kernel,*

$$P(W_{S_1}W_{S_2}) = PS_1S_2 \qquad \text{for } S_1, S_2 \in \mathcal{F}^{\perp},$$

such that the LRTS for (3.7) *satisfies*

$$\lim_{n\to\infty} 2\lambda_n^* = \sup_{S\in\mathcal{F}^{\perp}}(W_S \vee 0)^2.$$

PROOF. By Theorem 3.1, there exists a centered, uniformly continuous Gaussian process $\{W_S : S \in \bar{\mathcal{F}}_{2\varepsilon}\}$ such that

$$\lim_{n\to\infty} 2\lambda_n^* = \sup_{S\in\mathcal{F}_2}(W_S \vee 0)^2 - \sup_{S\in\mathcal{F}_1}(W_S \vee 0)^2.$$

For any $S \in \mathcal{F}_2$, in (3.9) we have $a^2 + b^2 = 1$. Then,

$$W_S = aW_{S^{\|}} + bW_{S^{\perp}} \leq \sqrt{(W_{S^{\|}} \vee 0)^2 + (W_{S^{\perp}} \vee 0)^2}. \tag{3.10}$$

Since $\mathcal{F}_2$ is convex, for any $S_1 \in \mathcal{F}_1$, $S_2 \in \mathcal{F}^{\perp}$ and $0 \leq p \leq 1$, we have $S_p = pS_1 + \sqrt{1-p^2}S_2 \in \mathcal{F}_2$. Let $\hat{p} = (W_{S_1} \vee 0)/\sqrt{(W_{S_1} \vee 0)^2 + (W_{S_2} \vee 0)^2}$. Then,

$$\left(W_{S_{\hat{p}}} \vee 0\right)^2 = \left(W_{S_1} \vee 0\right)^2 + \left(W_{S_2} \vee 0\right)^2. \tag{3.11}$$

Combining (3.10) and (3.11) proves the theorem. □

4. Applications to finite mixture models. In this section we study the LRT for testing the number of components in finite mixture models. According to Theorem 3.1, essentially one needs to verify the P-Donsker class condition for $\{S_\theta;\ \theta \in \Theta_\varepsilon\}$ and to derive $\mathcal{F}$. Note that the most widely used mixture models are those from the exponential families which have square integrable likelihood ratios and other nice analytic properties. For these models the Donsker class condition can be directly verified using Lemma 3.2 and other techniques in van der Vaart and Wellner (1996). In this section, we assume that the P-Donsker class condition is satisfied and focus on deriving the index set $\mathcal{F}$.

4.1. *General set-up in deriving* $\mathcal{F}$. We first present a set-up for deriving $\mathcal{F}$ in general hypothesis testing problems. When there are difficulties obtaining all limits of the generalized score functions directly, the following lemma shows that $\mathcal{F}$ can be determined by a class of functions equivalent to generalized score functions. We say that a family of functions $\{g_\theta : \theta \in \Theta\}$ satisfies $g_\theta = o(1)$ uniformly as $D(\theta) \to 0$ if there exists a square integrable function G such that

$$\limsup_{D(\theta)\to 0} \|g_\theta / G\| = 0.$$

LEMMA 4.1. *If* $S_\theta - T_\theta = o(1)$ *uniformly as* $D(\theta) \to 0$, *then* $\mathcal{F}$ *in* (3.1) *can be formulated as*

$$\mathcal{F} = \left\{ S \in \mathcal{L}^2 : \exists \{\theta^m\} \in \Theta_\varepsilon, \ s.t. \lim_{m\to\infty} D(\theta^m) = 0, \ \lim_{m\to\infty} \|T_{\theta^m} - S\| = 0 \right\}.$$

The proof of Lemma 4.1 is straightforward and will be omitted.

By Lemma 4.1, $\mathcal{F}$ can be derived using functions equivalent to the generalized score functions. One way to find the equivalent functions is to use Taylor expansions of the likelihood ratios. In many applications, including mixture models, there exists a reparameterization $\theta = (\phi, \psi) \in \Phi \otimes \Psi$ such that the equation $D(\theta) = 0$ is equivalent to the condition that $\phi = \phi_0$ for all $\psi \in \Psi$, where Φ and Ψ are subsets of $\Re^d$ and $\Re^k$, respectively. Then the first-order Taylor expansion at $\phi = \phi_0$ is

$$l_{(\phi,\psi)} = 1 + (\phi - \phi_0)^T \frac{\partial l_{(\phi_0,\psi)}}{\partial \phi} + o(|\phi - \phi_0|). \tag{4.1}$$

If $(\phi - \phi_0)^T \frac{\partial l_{(\phi_0,\psi)}}{\partial \phi}$ is not degenerate, by Lemma 4.1, $\mathcal{F}$ can be formulated as the $\mathcal{L}^2$ limits of $\Omega((\phi - \phi_0)^T \frac{\partial l_{(\phi_0,\psi)}}{\partial \phi})$ as $|\phi - \phi_0| \to 0$. Without loss of generality, we assume ϕ_0 is an interior point of Φ. Then the limits of $(\phi - \phi_0)/|\phi - \phi_0|$ form the unit ball in $\Re^d$. Note that $\mathcal{F}$ is closed, so $\mathcal{F} = \bar{\mathcal{F}}_0$, where

$$\mathcal{F}_0 = \left\{ \Omega\left(\beta^T \frac{\partial l_{(\phi_0,\psi)}}{\partial \phi} \right) : |\beta| = 1, \ \beta \in \Re^d, \ \psi \in \Psi \right\}. \tag{4.2}$$

From (4.2), the LRTS converges to the supremum of the square of a Gaussian process indexed by the closure of the convex cone of directional score functions, which gives the results of Lindsay (1995) and Dacunha-Castelle and Gassiat (1997, 1999) for mixture models. The local conic parameterization approach of Dacunha-Castelle and Gassiat is very useful in identifying $\mathcal{F}$. However, (4.2) fails when the directional score functions are linearly correlated; that is, there exist $\beta \in \Re^d$ $(\beta \neq 0)$ and $\phi \in \Phi$ such that $\beta^T \frac{\partial l_{(\phi_0,\psi)}}{\partial \phi} = 0$. In general, the closure of the convex cone of the directional score functions is only a subset of $\mathcal{F}$. To obtain $\mathcal{F}$, one may need to expand the likelihood ratios by Taylor expansion to a higher order or

use other approximations. For instance, when (4.1) fails, the Taylor expansion of the likelihood ratio function to the second order may still hold; that is,

$$l_{(\phi,\psi)} - 1 = (\phi - \phi_0)^T \frac{\partial l_{(\phi_0,\psi)}}{\partial \phi} + \frac{1}{2}(\phi - \phi_0)^T \frac{\partial^2 l_{(\phi_0,\psi)}}{\partial^2 \phi}(\phi - \phi_0) + o(D(\theta)). \tag{4.3}$$

In the next section, we show how to derive the explicit form of $\mathcal{F}$ for mixture models under (4.3).

4.2. *Finite mixture models.* Suppose $\{f_\alpha : \alpha \in A\}$ is a family of density functions with $A \subset \Re^d$ a compact, convex set. For two known integers $l < m$, testing a mixture model with l components against a mixture with m components can be expressed as testing

$$H_0 : \sum_{i=1}^{l} p_{0i} f_{\alpha_{0i}} \quad \text{against} \quad H_1 : \sum_{j=1}^{m} p_j f_{\alpha_j}, \tag{4.4}$$

where $0 \le p_{0i}, p_j \le 1$, $\alpha_{0i}, \alpha_j \in A$; $\sum_{i=1}^{l} p_{0i} = \sum_{j=1}^{m} p_j = 1$; p_{0i}, α_{0i} $(i = 1, \ldots, l)$ are assumed known and p_j, α_j $(j = 1, \ldots, m)$ are unknown parameters. Without loss of generality, we assume that $p_{0i} > 0$, α_{0i}'s are interior points of A and their values are different from each other. The likelihood ratio function for (4.4) is

$$l_{\boldsymbol{\alpha},\mathbf{p}} = \left(\sum_{j=1}^{m} p_j f_{\alpha_j}\right) \Big/ \left(\sum_{i=1}^{l} p_{0i} f_{\alpha_{0i}}\right).$$

We assume that the likelihood function is identifiable in the following sense:

$$l_{\boldsymbol{\alpha},\mathbf{p}} = 1 \iff \sum_{i=1}^{l} p_{0i}\delta_{\alpha_{0i}} = \sum_{j=1}^{m} p_j \delta_{\alpha_j}. \tag{A1}$$

In the following, we find a reparameterization of θ based on assumption (A1) and then obtain the Taylor expansion in (4.3). When $l_{\boldsymbol{\alpha},\mathbf{p}} = 1$, there exists a vector $\mathbf{t} = (t_i)_{i=0}^{l}$ such that $0 = t_0 < t_1 < \cdots < t_l \le m$ and, up to permutations, $(\boldsymbol{\alpha}, \mathbf{p})$ can be presented as $\alpha_{t_{i-1}+1} = \cdots = \alpha_{t_i} = \alpha_{0i}$, $\sum_{j=t_{i-1}+1}^{t_i} p_j = p_{0i}$, $i = 1, \ldots, l$; and $\alpha_j \notin \{\alpha_{01}, \ldots, \alpha_{0l}\}$, $p_j = 0$ for $j = t_l + 1, \ldots, m$. Define $\mathbf{s} = (s_i)_{i=1}^{l}$, $\mathbf{u} = (u_i)_{i=1}^{l}$ and $\mathbf{q} = (q_j)_{j=1}^{t_l}$, where

$$s_i = \sum_{j=t_{i-1}+1}^{t_i} p_j - p_{0i}, \qquad u_i = s_i - s_l p_{0i}/p_{0l}, \qquad q_j = p_j \Big/ \sum_{j=t_{i-1}+1}^{t_i} p_j,$$

for $i = 1, \ldots, l,\ j = t_{i-1} + 1, \ldots, t_i$. Note that $u_l = 0$. Define the reparameterization $\theta = (\boldsymbol{\phi}_\mathbf{t}, \boldsymbol{\psi}_\mathbf{t})$ by

$$\boldsymbol{\phi}_\mathbf{t} = \big((\alpha_j)_{j=1}^{t_l}, (s_i)_{i=1}^{l-1}, (p_j)_{j=t_l+1}^{m}\big), \qquad \boldsymbol{\psi}_\mathbf{t} = \big((q_j)_{j=1}^{t_l}, (\alpha_j)_{j=t_l+1}^{m}\big).$$

Then $\boldsymbol{\phi}_{0\mathbf{t}} = (\underbrace{\alpha_{01}, \ldots, \alpha_{01}}_{t_1}, \ldots, \underbrace{\alpha_{0l}, \ldots, \alpha_{0l}}_{t_l - t_{l-1}}, \underbrace{0, \ldots, 0}_{l-1}, \underbrace{0, \ldots, 0}_{m-t_l})$. Let $l_\alpha = f_\alpha / (\sum_{i=1}^{l} p_{0i} f_{\alpha_{0i}}) - 1$. Then

$$l_\theta - 1 = s \sum_{i=1}^{l} (s_i + p_{0i}) \sum_{j=t_{i-1}+1}^{t_i} q_j l_{\alpha_j} + \sum_{j=t_l+1}^{m} p_j l_{\alpha_j}.$$

Define the $\boldsymbol{\phi}_\mathbf{t}$-derivatives of the likelihood ratio function at $\boldsymbol{\phi}_{0\mathbf{t}}$ by

$$l_i' = \partial l_{\alpha_{0i}} / \partial \alpha_i \quad \text{and} \quad l_i'' = \partial^2 l_{\alpha_{0i}} / \partial^2 \alpha_i.$$

We then make the second general assumption that the likelihood ratio has the following Taylor expansion:

$$\begin{aligned} l_{(\phi_\mathbf{t}, \psi_\mathbf{t})} &= 1 + (\boldsymbol{\phi}_\mathbf{t} - \boldsymbol{\phi}_{0\mathbf{t}})^T l'_{\phi_{0t}, \psi_t} \\ &\quad + 0.5(\boldsymbol{\phi}_\mathbf{t} - \boldsymbol{\phi}_{0\mathbf{t}})^T l''_{\boldsymbol{\phi}_{0t}, \boldsymbol{\psi}_t} (\boldsymbol{\phi}_\mathbf{t} - \boldsymbol{\phi}_{0\mathbf{t}}) + o\big(D(\phi_\mathbf{t}, \psi_\mathbf{t})\big), \end{aligned} \tag{A2}$$

where

$$\begin{aligned} (\boldsymbol{\phi}_\mathbf{t} - \boldsymbol{\phi}_{0\mathbf{t}})^T l'_{\phi_{0t}, \psi_t} &= \sum_{i=1}^{l} p_{0i} \Bigg(\sum_{j=t_{i-1}+1}^{t_i} q_j \alpha_j - \alpha_{0i} \Bigg)^T l_i' \\ &\quad + \sum_{i=1}^{l-1} u_i l_{\alpha_{0i}} + \sum_{i=t_l+1}^{m} p_i l_{\alpha_i}, \end{aligned}$$

$$\begin{aligned} (\boldsymbol{\phi}_\mathbf{t} - \boldsymbol{\phi}_{0\mathbf{t}})^T l''_{\phi_{0t}, \psi_t} (\boldsymbol{\phi}_\mathbf{t} - \boldsymbol{\phi}_{0\mathbf{t}}) &= \sum_{i=1}^{l} \Bigg[2 s_i \Bigg(\sum_{j=t_{i-1}+1}^{t_i} q_j \alpha_j - \alpha_{0i} \Bigg)^T l_i' \\ &\quad + p_{0i} \sum_{j=t_{i-1}+1}^{t_i} q_j (\alpha_j - \alpha_{0i})^T l_i'' (\alpha_j - \alpha_{0i}) \Bigg]. \end{aligned}$$

Our third general assumption is that

$$\begin{aligned} &(\boldsymbol{\phi}_\mathbf{t} - \boldsymbol{\phi}_{0\mathbf{t}})^T l'_{\phi_{0t}, \psi_t} + 0.5(\boldsymbol{\phi}_\mathbf{t} - \boldsymbol{\phi}_{0\mathbf{t}})^T l''_{\phi_{0t}, \psi_t} (\boldsymbol{\phi}_\mathbf{t} - \boldsymbol{\phi}_{0\mathbf{t}}) = 0 \\ &\quad \Longleftrightarrow \quad \boldsymbol{\phi}_\mathbf{t} = \boldsymbol{\phi}_{0\mathbf{t}}. \end{aligned} \tag{A3}$$

Assumptions (A1)–(A3) allow us to obtain the explicit form of $\mathcal{F}$ and show that it is complete and admits continuous paths in the following theorem.

THEOREM 4.1. *In the hypothesis testing problem* (4.4), *assume that* (A1)–(A3) *hold. Then* $\mathcal{F} = \bigcup_{\mathbf{t}} \mathcal{F}_{\mathbf{t}}$ *with*

$$\mathcal{F}_{\mathbf{t}} = \Bigg\{ \Omega\Bigg(\sum_{i=1}^{l} \lambda_{i\mathbf{t}}^T l_i' + \sum_{i=1}^{l-1} \lambda_{(i+l)\mathbf{t}} l_{\alpha_{0i}} + \sum_{i=1}^{m-t_l} \lambda_{(i+2l-1)\mathbf{t}} l_{\alpha_{(i+t_l)\mathbf{t}}} + \delta \sum_{i=1}^{l} \sum_{j=t_{i-1}+1}^{t_i} \gamma_{j\mathbf{t}}^T l_i'' \gamma_{j\mathbf{t}} \Bigg) : \tag{4.5}$$

$$\lambda_{1\mathbf{t}}, \ldots \lambda_{l\mathbf{t}} \in \Re^{d*}, \lambda_{(l+1)\mathbf{t}}, \ldots, \lambda_{(2l-1)\mathbf{t}} \in \Re,$$

$$\lambda_{(2l)\mathbf{t}}, \ldots, \lambda_{(m+2l-t_l)\mathbf{t}} \in \Re^+; \gamma_1, \ldots, \gamma_{t_l} \in \Re^d;$$

$$\alpha_{t_l+1}, \ldots, \alpha_m \in A \setminus \{\alpha_{01}, \ldots, \alpha_{0l}\}; |\boldsymbol{\lambda}_{\mathbf{t}}| + \delta |\boldsymbol{\gamma}_{\mathbf{t}}| = 1 \Bigg\},$$

where in (4.5) $\delta = 1$, *if there exists a vector* $\mathbf{q}$ *such that* $q_j \geq 0$, $\sum_{j=t_{i-1}}^{t_i} q_j = 1$ *and* $\sum_{j=t_{i-1}}^{t_i} \sqrt{q_j}\gamma_j = 0$, *for* $i = 1, \ldots, l$; *and* 0 *otherwise*; *the union runs over all possible* $\mathbf{t}$ *with* $0 = t_0 < t_1 < \cdots < t_l = m$. *Moreover*, $\mathcal{F}$ *is complete and admits continuous paths*.

PROOF. To further simplify notation, let

$$\boldsymbol{\gamma} = \Bigg(\Bigg[(p_{0i} + s_i) \Bigg(\sum_{j=t_{i-1}+1}^{t_i} q_j \alpha_j - \alpha_{0i} \Bigg) \Bigg]_{i=1}^{l}, (u_i)_{i=1}^{l-1}, (p_i)_{i=t_l+1}^{m} \Bigg);$$

$$\boldsymbol{\lambda} = \big([\sqrt{q}_j (\alpha_j - \alpha_{0i})]_{j=t_{i-1}+1}^{t_i} \big)_{i=1}^{l}; \qquad \eta = |\boldsymbol{\lambda}|^2 / (2|\boldsymbol{\gamma}| + |\boldsymbol{\lambda}|^2);$$

$$\mathbf{L}' = \operatorname{diag}(l_1', \ldots, l_l', l_{\alpha_{01}}, \ldots, l_{\alpha_{0(l-1)}}, \ldots, l_{\alpha_{t_l+1}}, l_{\alpha_m});$$

$$\mathbf{L}'' = \operatorname{diag}(\underbrace{l_1', \ldots, l_1'}_{t_1}, \underbrace{l_2', \ldots, l_2'}_{t_2 - t_1}, \ldots, \underbrace{l_l', \ldots, l_l'}_{t_l - t_{l-1}}).$$

Recall that $\omega(\mathbf{x}) = \mathbf{x}/|\mathbf{x}|$, for $\mathbf{x} \neq 0$. The Taylor expansion in (A2) can be expressed as

$$\begin{aligned} l_\theta - 1 &= \boldsymbol{\gamma}^T \mathbf{L}' + 0.5 \boldsymbol{\lambda}^T \mathbf{L}'' \boldsymbol{\lambda} + o(D(\theta)) \\ &= |\boldsymbol{\gamma}| + 0.5 |\boldsymbol{\lambda}|^2 \big[(1-\eta) \omega(\boldsymbol{\gamma})^T \mathbf{L}' + \eta \omega(\boldsymbol{\lambda})^T \mathbf{L}'' \omega(\boldsymbol{\lambda}) + o(1) \big]. \end{aligned} \tag{4.6}$$

By definition, for $S \in \mathcal{F}$, there exists a sequence $\{\theta^r\} \in \Theta \setminus \Theta_0$ such that $D(\theta^r)$ tends to zero and $\Omega(l_{\theta^r} - 1)$ converges to S in $\mathcal{L}^2$ as $r \to 0$. For any function $g(\theta)$ bounded in $\Theta \setminus \Theta_0$, we can choose a suitable subsequence $\{\theta^{r_k}\}$ of $\{\theta^r\}$ such that $g(\theta^{r_k})$ converges. Without loss of generality, we may assume that $(\eta^r, \omega(\boldsymbol{\gamma}^r), \omega(\boldsymbol{\lambda}^r), \mathbf{q}^r)$ converges to $(\eta, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathbf{q})$ as $r \to 0$. Assumption (A3) ensures that $\|(1-\eta)\boldsymbol{\gamma}^T \mathbf{L}' + \eta \boldsymbol{\lambda}^T \mathbf{L}'' \boldsymbol{\lambda})\| > 0$. Thus $\Omega((1-\eta^r)\omega(\boldsymbol{\gamma}^r)^T \mathbf{L}' + \eta^r \omega(\boldsymbol{\lambda}^r)^T \mathbf{L}'' \omega(\boldsymbol{\lambda}^r))$ converges to $S = \Omega((1-\eta)\boldsymbol{\gamma}^T \mathbf{L}' + \eta \boldsymbol{\lambda}^T \mathbf{L}'' \boldsymbol{\lambda})$.

Now we can show that S can be written in the form of (4.5). When $\eta = 1$, note that $\boldsymbol{\gamma}^r \to 0$. We have $\boldsymbol{\gamma}^r/|\boldsymbol{\lambda}^r| \to 0$ and

$$(p_{0i} + s_i^r) \sum_{j=t_{i-1}+1}^{t_i} \sqrt{q_j^r}\omega(\boldsymbol{\gamma}^r)_j = \gamma_i^r/|\boldsymbol{\lambda}^r|,$$

for $i = 1, \ldots, l$. Since $p_{0i} > 0$ and s_i^r converges to 0, we have $\sum_{j=t_{i-1}+1}^{t_i} \sqrt{q_j} \times \boldsymbol{\gamma}_j = 0$. Simple linear transformations of the parameters allow $\Omega((1-\eta)\boldsymbol{\gamma}^T\mathbf{L}' + \eta\boldsymbol{\lambda}^T\mathbf{L}''\boldsymbol{\lambda})$ to be expressed as (4.5). When $\eta = 0$, (4.5) is obvious. Therefore, $\mathcal{F}$ is a subset of $\bigcup_{\mathbf{t}} \mathcal{F}_{\mathbf{t}}$.

Next we prove $\mathcal{F} = \bigcup_{\mathbf{t}} \mathcal{F}_{\mathbf{t}}$ by showing that for any $S \in \mathcal{F}_{\mathbf{t}}$, there exists a continuous path θ^r such that S is the pointwise and $\mathcal{L}^2$ limit of S_{θ^r} as r tends to 0. Here we consider the case $\eta = 1$ only. The proof of the case $\eta = 0$ is similar and not presented here. By definition of $\mathcal{F}_{\mathbf{t}}$, S can be defined in (4.5) by parameters $((\lambda_i)_{i=1}^{2l+m-t_l-1}, (\gamma_i)_{i=1}^{t_l}, (q_i)_{i=1}^{t_l}, (\alpha_i)_{i=t_l}^{m})$, where $\sum_{j=t_{i-1}+1}^{t_i} \sqrt{q_j}\gamma_j = 0$, $\sum_{j=t_{i-1}+1}^{t_i} q_j = 1$. For simplicity, we express it as $S = \Omega((1-\eta)\boldsymbol{\gamma}^T\mathbf{L}' + \eta\boldsymbol{\lambda}^T\mathbf{L}''\boldsymbol{\lambda})$ where $\eta = |\boldsymbol{\gamma}|^2/(2|\boldsymbol{\lambda}| + |\boldsymbol{\gamma}|^2) > 0$. Define a sequence $\{\theta^r\}$ for small r as follows: $\alpha_i^r = \alpha_i$, $p_i^r = r^4\lambda_{i-t_l+2l-1}$ for $i = t_l + 1, \ldots, m$; $r_i = r^4\lambda_{l+i}$ for $i = 1, \ldots, l-1$; and

$$q_j^r = \begin{cases} r^2, & \text{if } q_j = 0, \\ q_j\left(1 - r^2 \sum_{j=t_{i-1}+1}^{t_l} \delta_0(q_j)\right), & \text{if } q_j \neq 0, \end{cases}$$

$$\alpha_j^r - \alpha_{0i} = \begin{cases} r\gamma_j, & \text{if } q_j = 0, \\ r^2\gamma_j/\sqrt{q_j} - r^3 \sum_{j=t_{i-1}+1}^{t_l} \delta_0(q_j)\gamma_j + r^4\lambda_i/p_{0i}, & \text{if } q_j \neq 0, \end{cases}$$

for $i = 1, \ldots, m$, $j = t_{i-1} + 1, \ldots, t_i$, where $\delta_0(x) = 0$ if $x = 0$; and 1, otherwise. $(\boldsymbol{\gamma}^r, \boldsymbol{\lambda}^r, \eta^r)$ is defined by θ^r as above.

$$\begin{aligned} &\sum_{j=t_{i-1}+1}^{t_i} q_j^r\,(\alpha_j^r - \alpha_{0i}) \\ &\quad = r^3 \sum_{q_j=0} \gamma_j \\ &\qquad + r^2 \sum_{j=t_{i-1}+1}^{t_i} q_j\Big(1 - r^2 \sum \delta_0(q_j)\Big)\Big(\gamma_j/\sqrt{q_j} - r \sum_{q_j=0} \gamma_j + r^2\lambda_i/p_{0i}\Big) \\ &\quad = r^2\gamma_i \sum_{j=t_{i-1}+1}^{t_i} \sqrt{q_j}\gamma_i + r^3\left(1 - \sum_{j=t_{i-1}+1}^{t_i} q_j\right) \sum_{q_j=0} \gamma_j t^4 \gamma_i \end{aligned}$$

$$+ r^4\left(\gamma_i \Big/ p_{0i} \sum_{j=t_{i-1}+1}^{t_i} q_i - \sum_{j=t_{i-1}+1}^{t_i} \delta_0(q_j) \sum_{j=t_{i-1}+1}^{t_i} \sqrt{q_j}\gamma_i\right) + o(r^4)$$

$$= r^4 \gamma_i / p_{0i} + o(r^4),$$

where we have used the equations $\sum_{j=t_{i-1}+1}^{t_i} \sqrt{q_j}\gamma_i = 0$ and $\sum_{j=t_{i-1}+1}^{t_i} q_i = 1$. Clearly $\boldsymbol{\lambda}^r = r^2\boldsymbol{\lambda} + o(r^2)$. Then $(r^{-4}\boldsymbol{\gamma}^r, r^{-2}\boldsymbol{\lambda}^r) \to (\boldsymbol{\gamma}, \boldsymbol{\lambda})$ as r tends to 0. Equation (4.6) yields

$$l_{\theta^r} - 1 = r^4\big[(1-\eta)\boldsymbol{\lambda}^T \mathbf{L}' + \eta \boldsymbol{\gamma}^T \mathbf{L}'' \boldsymbol{\gamma} + o(1)\big].$$

Note that l_θ is a continuous function in θ. Therefore $\{\theta^r\}$ is a continuous path for S such that $S_{\theta^r} \to S$ in $\mathcal{L}^2$ as $r \to 0$ and $D(\theta^r)$ is a continuous function of r. Similarly, we can show that $\mathcal{F}$ is complete and the $\mathcal{L}^2$ convergence of $\mathcal{F}$ implies pointwise convergence. We omit the details. □

Testing homogeneity in mixture models is a frequently met problem in applications. The homogeneity test corresponds to $l = 1$ in (4.4). For simplicity, let $\alpha_0 = \alpha_{01}$, $l' = l'_1$ and $l'' = l''_1$. The vector $\mathbf{t}$ in Theorem 4.1 is then reduced to a scalar parameter t where $1 \le t \le m$. We give the following corollary without proof.

COROLLARY 4.1. *In the hypothesis testing problem* (4.4), *assume that* $l = 1$ *and that assumptions* (A1)–(A3) *hold. Then* $\mathcal{F} = \bigcup_{1\le t\le m} \mathcal{F}_t$, *where*

$$\mathcal{F}_t = \left\{ \Omega\left(\lambda_1^T l' + \sum_{i=1}^{m-t} \lambda_{i+1} l_{\alpha_{i+t}} + \sum_{i=1}^{t} \gamma_i^T l'' \gamma_i\right): \right.$$
$$\lambda_1 \in \Re^d, \lambda_2, \ldots \lambda_{m-t+1} \in \Re^+; \gamma_1, \ldots, \gamma_t \in \Re^d \text{ and}$$
$$\left. \text{Rank}\,(\boldsymbol{\gamma}) < t; \alpha_{1+t}, \ldots, \alpha_m \in A \setminus \alpha_0;\ |\boldsymbol{\lambda}| + |\boldsymbol{\gamma}| = 1 \right\}.$$

Acknowledgments. We thank the referees who provided valuable suggestions which have improved every aspect of the paper and who brought to our attention some important references.

REFERENCES

AZAÏS, J.-M., CIERCO-AYROLLES, C. and CROQUETTE, A. (1999). Bounds and asymptotic expansions for the distribution of the maximum of a smooth stationary Gaussian process. *ESAIM Probab. Statist.* **3** 107–129.

AZAÏS, J.-M. and WSCHEBOR, M. (2000). Computing the distribution of the maximum of a Gaussian process. Univ. Sabatier, Toulouse. Preprint.

CHERNOFF, H. (1954). On the distribution of the likelihood ratio. *Ann. Math. Statist.* **25** 573–578.

CHERNOFF, H. and LANDER, E. (1995). Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. *J. Statist. Plann. Inference* **43** 19–40.

DACUNHA-CASTELLE, D. and GASSIAT, E. (1997). Testing in locally conic models and application to mixture models. *ESAIM Probab. Statist.* **1** 285–317.

DACUNHA-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.

DUDLEY, R. M. (1999). *Uniform Central Limit Theorems*. Cambridge Univ. Press.

LE CAM, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Statist.* **41** 802–828.

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.* IMS, Hayward, CA.

MCLACHLAN, G. J. and BASFORD, K. E. (1988). *Mixture Models*: *Inference and Applications to Clustering*. Dekker, New York.

PRAKASA RAO, B. L. S. (1992). *Identifiability in Stochastic Models*: *Characterization of Probability Distributions.* Academic Press, London.

REDNER, R. A. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Ann. Statist.* **9** 225–228.

TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer, New York.

LAB OF STATISTICAL GENETICS
ROCKEFELLER UNIVERSITY
1230 YORK AVENUE
NEW YORK, NEW YORK 10021
E-MAIL: liuxin@linkage.rockefeller.edu

DEPARTMENT OF STATISTICS
COLUMBIA UNIVERSITY
2990 BROADWAY
NEW YORK, NEW YORK 10027
E-MAIL: yshao@stat.columbia.edu