A note on profile likelihood for exponential tilt mixture models

BY Z. TAN

Department of Statistics, Rutgers University, Piscataway, New Jersey 08854, U.S.A.
ztan@stat.rutgers.edu

SUMMARY

Suppose that independent observations are drawn from multiple distributions, each of which is a mixture of two component distributions such that their log density ratio satisfies a linear model with a slope parameter and an intercept parameter. Inference for such models has been studied using empirical likelihood, and mixed results have been obtained. The profile empirical likelihood of the slope and intercept has an irregularity at the null hypothesis so that the two component distributions are equal. We derive a profile empirical likelihood and maximum likelihood estimator of the slope alone, and obtain the usual asymptotic properties for the estimator and the likelihood ratio statistic regardless of the null. Furthermore, we show the maximum likelihood estimator of the slope and intercept jointly is consistent and asymptotically normal regardless of the null. At the null, the joint maximum likelihood estimator falls along a straight line through the origin with perfect correlation asymptotically to the first order.

Some key words: Empirical likelihood; Exponential tilt; Likelihood ratio statistic; Maximum likelihood; Mixture model; Profile likelihood.

1. INTRODUCTION

Suppose that a sample of $n_j$ independent observations, $\{x_{j1}, \ldots, x_{jn_j}\}$, is drawn from each of $m (\geq 2)$ distributions $P_j$ ($j = 1, \ldots, m$), and each $P_j$ is a mixture of two component distributions, $G_0$ and $G_1$, related in an exponential tilt form:

$$dP_j = \lambda_j dG_0 + (1 - \lambda_j) dG_1,$$

where $0 \leq \lambda_j \leq 1$ are known mixture proportions, $\lambda_j \neq \lambda_k$ ($j \neq k$), and

$$dG_1 = \exp(\beta_0 + \beta_1 x) dG_0,$$

where $G_0$ is an unknown baseline distribution, $\beta_1$ is an unknown parameter, and $\beta_0 = -\log[\int \exp(\beta_1 x) dG_0]$. Note that $\beta_1 = 0$ implies that $\beta_0 = 0$ for any $G_0$. The $j$th sample, $\{x_{j1}, \ldots, x_{jn_j}\}$, is obtained from

$$dP_j = (\lambda_j + (1 - \lambda_j) \exp(\beta_0 + \beta_1 x)) dG_0. \quad (1)$$

The special case of model (1) where $m = 2$, $\lambda_1 = 1$, and $\lambda_2 = 0$ is known as the two-sample exponential tilt model (Qin, 1998), first studied by Anderson (1972) and Prentice & Pyke (1979) in the context of case-control studies.

Inference for $\beta = (\beta_1, \beta_0)'$ and $G_0$ in model (1) has been studied using empirical likelihood (Owen, 2001), and mixed results have been obtained. Qin (1999) derived a profile empirical likelihood and the resulting maximum likelihood estimator of $\beta$ in the case where $m = 3$, $\lambda_1 = 1$, $\lambda_2 = 0$, and $0 < \lambda_3 < 1$ may be unknown. However, Zou et al. (2002) pointed out that the profile empirical likelihood of $\beta$ has an irregularity at the null hypothesis that $G_1 = G_0$, i.e., $\beta = 0$, and the asymptotic arguments in Qin (1999) for the maximum likelihood estimator of $\beta$ are not applicable there. Zou & Fine (2002) claimed that
the maximum likelihood estimator of \( \beta \) is not consistent when \( \beta = 0 \). To address this, Zou et al. (2002) proposed a partial profile empirical likelihood, and showed that the maximum partial likelihood estimator of \( \beta \) is consistent and asymptotically normal whether or not \( \beta = 0 \), and the profile loglikelihood ratio for testing \( \beta = 0 \) has a chi-squared distribution with one degree of freedom asymptotically. A limitation of this method is that, when \( \beta \neq 0 \), the maximum partial likelihood estimator of \( \beta \) always has asymptotic variance no smaller than that of the maximum likelihood estimator in Qin (1999). See Zou & Fine (2002) for a further discussion on the construction of the maximum partial likelihood estimator and the efficiency comparisons of several related estimators.

In this note we first derive a profile empirical likelihood and the resulting maximum likelihood estimator of \( \beta \) alone, and show that the profile empirical likelihood of \( \beta \) does not have an irregularity at \( \beta = 0 \), in contrast to that of \( \beta \). In fact, under mild regularity conditions, the profile empirical likelihood and the maximum likelihood estimator of \( \beta \) behave like usual profile likelihoods and maximum likelihood estimators with the following properties, whether or not \( \beta = 0 \): the maximum likelihood estimator is consistent and asymptotically normal; the inverse curvature of the profile loglikelihood at its maximum is a consistent estimator of the asymptotic variance of the maximum likelihood estimator; and the loglikelihood ratio statistic has a chi-squared distribution with one degree of freedom asymptotically. Second, we derive in our approach the maximum likelihood estimator of \( \beta \), which agrees with that in Qin (1999), and show that it is consistent and asymptotically normal, and always has asymptotic variance no greater than that of the maximum partial likelihood estimator of \( \beta \) in Zou et al. (2002), whether or not \( \beta = 0 \). A subtle point is that when \( \beta = 0 \), the maximum likelihood estimator of \( \beta \) falls along a straight line through the origin with perfect correlation asymptotically to the first order.

The proofs of the main results are provided in the Appendix.

2. Profile Likelihood and Maximum Likelihood Estimator

The loglikelihood of \((\beta_1, G_0)\) is

\[
{\ell(\beta_1, G_0) = \sum_{j=1}^m \sum_{i=1}^{n_j} \left[ \log(\lambda_j + (1 - \lambda_j)\exp(\beta_0 + \beta_1 x_{ji})) + \log G_0(\{x_{ji}\}) \right]},
\]

where \(\beta_0 = -\log\left( \int \exp(\beta_1 x) \, dG_0 \right)\) is a function of \((\beta_1, G_0)\). The profile loglikelihood of \(\beta_1\) is defined for each fixed \(\beta_1\) by maximizing \(l(\beta_1, G_0)\) over all possible probability distributions \(G_0\), that is, \(pl(\beta_1) = \max_{G_0} l(\beta_1, G_0)\). The maximum likelihood estimator of \(\beta_1\) is defined by maximizing the profile loglikelihood, that is, \(\hat{\beta}_1 = \arg\max_{\beta_1} pl(\beta_1)\). We provide a formula for computing the profile loglikelihood \(pl(\beta_1)\) in Proposition 1 (i). The formula involves the function

\[
\kappa(\beta, \alpha_0) = \sum_{j=1}^m \sum_{i=1}^{n_j} \log \left( \frac{\lambda_j + (1 - \lambda_j)\exp(\beta_0 + \beta_1 x_{ji})}{\alpha_0 + (1 - \alpha_0)\exp(\beta_0 + \beta_1 x_{ji})} \right) - n \log n,
\]

where \(\beta = (\beta_1, \beta_0)^T\) and \(0 \leq \alpha_0 \leq 1\) are free arguments. This function has a duality relationship with the loglikelihood \(l(\beta_1, G_0)\) as shown in Proposition 1 (ii).

**Proposition 1.** (i) For a fixed \(\beta_1\), the loglikelihood \(l(\beta_1, G_0)\) achieves the maximum at \(\hat{G}_0 = \hat{G}_0(\beta_1)\) supported on \(\{x_{ji} : i = 1, \ldots, n_j, j = 1, \ldots, m\}\) with

\[
\hat{G}(\{x_{ji}\}) = \frac{n^{-1}}{\hat{\alpha}_0 + (1 - \hat{\alpha}_0)\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{ji})},
\]

where \(n = \sum_{j=1}^m n_j\), and \(\hat{\beta}_0 = \hat{\beta}_0(\beta_1)\) and \(\hat{\alpha}_0 = \hat{\alpha}_0(\beta_1)\) satisfy

\[
\alpha_0 = \frac{1}{n} \sum_{j=1}^m \sum_{i=1}^{n_j} \frac{\lambda_j}{\lambda_j + (1 - \lambda_j)\exp(\beta_0 + \beta_1 x_{ji})},
\]

(2)
The profile loglikelihood is \( pl(\beta_1) = \kappa(\beta_1, \hat{\beta}_0(\beta_1), \hat{\alpha}_0(\beta_1)) \).

(ii) Furthermore, for any \( G_0 \) and any \((\beta_0, \alpha_0)\) satisfying \( 0 \leq \alpha_0 \leq 1 \),

\[
\max_{G_0} l(\beta_1, G_0) \leq \min_{\alpha_0} \kappa(\beta, \alpha_0),
\]

where the equality holds when \( G_0 = \hat{\beta}_0(\beta_1) \) and \((\beta_0, \alpha_0) = (\hat{\beta}_0(\beta_1), \hat{\alpha}_0(\beta_1)) \).

The profile loglikelihood \( pl(\beta_1) \) is a function of \( \beta_1 \) alone, with \( \hat{\beta}_0(\beta_1) \) and \( \hat{\alpha}_0(\beta_1) \) implicitly defined by equations (2)–(4). Note that (3) and (4) are equivalent to each other, and both are equivalent to the equation \( 0 = \partial \kappa / \partial \alpha_0 \), and that (2), (3) and (4) jointly are equivalent to \( 0 = \partial \kappa / \partial \beta_0 \) and \( 0 = \partial \kappa / \partial \alpha_0 \) jointly. See the Appendix for the gradient and Hessian of \( \kappa(\beta, \alpha_0) \). By implicit differentiation, the gradient and Hessian of \( pl(\beta_1) \) are

\[
\begin{align*}
\frac{\partial pl}{\partial \beta_1} &= \frac{\partial \kappa}{\partial \beta_1} \bigg|_{\gamma=\hat{\gamma}(\beta_1)}, \\
\frac{\partial^2 pl}{\partial \beta_1^2} &= \left\{ \frac{\partial^2 \kappa}{\partial \beta_1^2} - \frac{\partial^2 \kappa}{\partial \beta_1 \partial \gamma} \left( \frac{\partial^2 \kappa}{\partial \gamma \partial \gamma} \right)^{-1} \frac{\partial^2 \kappa}{\partial \gamma \partial \beta_1} \right\} \bigg|_{\gamma=\hat{\gamma}(\beta_1)},
\end{align*}
\]

where \( \kappa(\beta, \alpha_0) \) is treated as \( \kappa(\beta_1, \gamma) \) with \( \gamma = (\beta_0, \alpha_0)^\top \), and \( \hat{\gamma}(\beta_1) = (\hat{\beta}_0(\beta_1), \hat{\alpha}_0(\beta_1))^\top \). These formulas can be used in the Newton–Raphson algorithm to numerically maximize \( pl(\beta_1) \) and find the maximum likelihood estimator \( \hat{\beta}_1 \).

In asymptotic considerations, assume that \( n_j/n \) tends to a constant \( 0 < \rho_j < 1 \) as \( n \to \infty \) \((j = 1, \ldots, m) \). Let \( \beta^* \) be the true value of \( \beta \), \( \alpha_0^\star = \sum_{j=1}^m \rho_j \lambda_j \), and \( \gamma^\star = (\beta_0^\star, \alpha_0^\star)^\top \). The following lemma provides basic calculations used in Proposition 2, due to the law of large numbers and the central limit theorem.

**Lemma 1.** Suppose that \( \beta \) and \( \alpha_0 \) are evaluated at the true values \( \beta^* \) and \( \alpha_0^\star \).

(i) As \( n \to \infty \),

\[
-\frac{1}{n} \left( \begin{array}{c}
\frac{\partial^2 \kappa}{\partial \beta_1^2} \\
\frac{\partial^2 \kappa}{\partial \beta_1 \partial \gamma_1} \\
\frac{\partial^2 \kappa}{\partial \beta_1 \partial \gamma_2} \\
\frac{\partial^2 \kappa}{\partial \gamma_1 \partial \gamma_1} \\
\frac{\partial^2 \kappa}{\partial \gamma_1 \partial \gamma_2} \\
\frac{\partial^2 \kappa}{\partial \gamma_2 \partial \gamma_2}
\end{array} \right) \to U^\top = \left( \begin{array}{cc}
S_{11} & S_{12} \\
S_{21} & S_{22}
\end{array} \right),
\]

in probability, where

\[
S_{11} = -\sum_{j=1}^m \rho_j \int \frac{\lambda_j (1 - \lambda_j) \exp(\beta^* x) z^\top}{\lambda_j + (1 - \lambda_j) \exp(\beta^* z)} \text{d}G_0 + \int \frac{\alpha_0^\star (1 - \alpha_0^\star) \exp(\beta^* z) z^\top}{\alpha_0^\star + (1 - \alpha_0^\star) \exp(\beta^* z)} \text{d}G_0,
\]

\[
S_{12} = S_{21} = -\int \frac{\exp(\beta^* z) z}{\alpha_0^\star + (1 - \alpha_0^\star) \exp(\beta^* z)} \text{d}G_0,
\]

\[
S_{22} = -\int \frac{(1 - \exp(\beta^* z))^2}{\alpha_0^\star + (1 - \alpha_0^\star) \exp(\beta^* z)} \text{d}G_0,
\]

and \( z = (x, 1)^\top \).
(ii) As \( n \to \infty \), \( n^{-1/2}(\partial \kappa/\partial \beta^T, \partial \kappa/\partial \alpha_0)^T \) converges in distribution to a multivariate normal variable with mean 0 and variance matrix

\[
V^\dagger = \begin{pmatrix}
S_{11} - \delta S_{12}S_{21} & -\delta S_{12}s_{22} \\
-\delta S_{21}s_{22} & -s_{22} - \delta s_{22}^2
\end{pmatrix},
\]

where \( \delta = \sum_{j=1}^m \rho_j \lambda_j^2 - \alpha_0 s^2 \).

We show in Proposition 2 (i) that (5) and (6) satisfy the second Bartlett identity: the variance of the gradient equals the expectation of the negative Hessian for a loglikelihood function, evaluated at the true value of the parameter. This leads directly to the negative Hessian representation for the asymptotic variance of the maximum likelihood estimator, and the chi-squared convergence of the loglikelihood ratio statistic in (ii) and (iii).

**PROPOSITION 2.** Rewrite \( U^\dagger \) and \( V^\dagger \) in Lemma 1 by the partition \((\beta_1, \gamma)\) instead of \((\beta, \alpha_0)\):

\[
U^\dagger = \begin{pmatrix}
\sigma_{11} & \Sigma_{10} \\
\Sigma_{01} & \Sigma_{00}
\end{pmatrix}, \quad V^\dagger = U^\dagger - \begin{pmatrix} 0 & S_{12} \\
S_{21} & 2s_{22} \end{pmatrix} - \delta \begin{pmatrix} S_{12} \\
S_{22} \end{pmatrix} (S_{21}, s_{22}).
\]

Assume that \( U^\dagger \) and \( \Sigma_{00} \) are nonsingular.

(i) Equations (2)–(4) for \( \beta_1 = \beta_1^* \) admit a solution \( \hat{\gamma}(\beta_1^*) \) that converges to \( \gamma^* \) in probability,

\[
\frac{1}{n^{1/2}} \frac{\partial p}{\partial \beta_1} \bigg|_{\beta_1 = \beta_1^*} \to N(0, V),
\]

in distribution, and

\[
-\frac{1}{n} \frac{\partial^2 p}{\partial \beta_1^2} \bigg|_{\beta_1 = \beta_1^*} \to U,
\]

in probability, where \( U = V = \sigma_{11} - \Sigma_{10} \Sigma_{01}^{-1} \Sigma_{01} \).

(ii) Assume that \( \int x^2 \, dG_0 < \infty \) and \( \int x^2 \, dG_1 < \infty \). Then \( 0 = \partial p/\partial \beta_1 \) admits a solution \( \hat{\beta}_1 \) that converges in probability to \( \beta_1^* \), and \( n^{1/2}(\hat{\beta}_1 - \beta_1^*) \) converges in distribution to \( N(0, V) \).

(iii) Furthermore, \(-2(p(\beta_1^*) - p(\hat{\beta}_1))\) converges in distribution to a chi-squared variable with one degree of freedom.

**COROLLARY 1.** Under the same conditions in Proposition 2 (ii), \( n^{1/2}\{\hat{\beta}_1 - \beta_1^*, \hat{\gamma}(\hat{\beta}_1) - \gamma^*\} \) converges in distribution to a multivariate normal variable with mean zero and variance matrix

\[
\begin{pmatrix}
S_{11} & 0 \\
0 & -s_{22} - \delta
\end{pmatrix},
\]

where \( s_{22} \) and \( S_{11} \) are defined by the partition of \( U^{1-1} \):

\[
U^{1-1} = \begin{pmatrix}
S_{11} & S_{12} \\
S_{21} & S_{22}
\end{pmatrix}.
\]

Several interesting conclusions can be drawn. First, the maximum likelihood estimator of \( \beta \) is given by \( \hat{\beta} = \{\hat{\beta}_1, \hat{\beta}_0(\hat{\beta}_1)\}^T \), and it agrees with that derived in Qin (1999). In fact, \( \{\hat{\beta}_1, \hat{\gamma}(\hat{\beta}_1)\} \) satisfies \( 0 = \partial \kappa/\partial \beta_1 \) and \( 0 = \partial \kappa/\partial \gamma \), which are equivalent to equations (2-2) and (2-5) in Qin (1999). By Corollary 1, \( n^{1/2}(\hat{\beta}_1 - \beta_1^*) \) converges in distribution to a multivariate normal variable with mean 0 and variance \( S_{11} \), which equals \( S_{11}^{1-1} + S_{11}^{1-1} S_{12} s_{22} S_{21} S_{11}^{1-1} \).

Second, \( \hat{\beta} \) is asymptotically at least as efficient as \( \hat{\beta}_e \), the maximum profile likelihood estimator of \( \beta \) in Zou et al. (2002), which is defined as a maximizer to \( \kappa(\beta, \hat{\alpha}_0) \) with \( \hat{\alpha}_0 = \sum_{j=1}^m (n_j/n) \lambda_j \). As shown by Zou et al. (2002), \( n^{1/2}(\hat{\beta}_e - \beta^*) \) converges in distribution to a multivariate normal variable with mean zero and variance \( S_{11}^{1-1} - \delta S_{11}^{1-1} S_{12} S_{21} S_{11}^{1-1} \), which is no smaller than \( S_{11} \) because \( -\delta \geq s_{22} \) by the positive semidefiniteness of (7) in Corollary 1.
Third, when $\beta^* = 0$, $\hat{\beta}$ falls along the line $\{\beta : (\mu, 1) \beta = 0\}$ asymptotically to the order of $n^{-1/2}$, where $\mu = \int x \, dG_0$. Let $\sigma^2 = \int x^2 \, dG_0 - \mu^2$. The asymptotic variance (7) becomes
\[
\begin{pmatrix}
\frac{1}{\delta \sigma^2} & -\frac{\mu}{\delta \sigma^2} & 0 \\
-\frac{\mu}{\delta \sigma^2} & \frac{\mu^2}{\delta \sigma^2} & 0 \\
0 & 0 & 0
\end{pmatrix}.
\] (8)
As a result, $\hat{\alpha}_0(\hat{\beta}_1) = \alpha_0^* + o_p(n^{-1/2})$, and $(\mu, 1) \hat{\beta} = o_p(n^{-1/2})$. Note that $s_{22}^2 = -\delta$ in this case. The asymptotic variance of $n^{1/2}(\hat{\beta}_1 - \beta^*)$ is identical to that of $n^{1/2}(\hat{\beta} - \beta^*)$, and therefore $\hat{\beta}_1$ falls along the line $\{\beta : (\mu, 1) \beta = 0\}$ in a similar manner as $\hat{\beta}$.

In the special case where $m = 2$, $\lambda_1 = 1$, and $\lambda_2 = 0$, corresponding to the two-sample exponential tilt model (Qin, 1998), the assumption that $U^\perp$ is nonsingular is violated for Proposition 2, because the $3 \times 3$ matrix
\[
U^\perp = \begin{pmatrix}
\delta s_{xx} & \delta s_{x1} & -s_{x1} \\
\delta s_{1x} & \delta s_{11} & -s_{11} \\
-s_{1x} & -s_{11} & \delta^{-1} s_{11}
\end{pmatrix}
\]
is singular, where
\[
s_{uv} = \int \frac{\exp (\beta_0^* + \beta_1^* x) \alpha_u \alpha_v}{\alpha_0^* + (1 - \alpha_0^*) \exp (\beta_0^* + \beta_1^* x)} \, dG_0
\]
for $(u, v) = (x, x), (x, 1),$ or $(1,1)$. This reduction occurs because equation (2) implies that $\hat{\alpha}_0(\beta_1) = n_1/n$, independent of $\beta_1$. Nevertheless, similar conclusions to Proposition 2 and Corollary 1 can be obtained. The asymptotic variance of $n^{1/2}(\hat{\beta}_1 - \beta_1^*)$ is $U^{-1}$, where $U = \delta(s_{xx} - s_{x1}^2/s_{11})$. The asymptotic variance of $n^{1/2}(\hat{\beta} - \beta^*)$ is
\[
\delta^{-1} \left\{ \begin{pmatrix} s_{xx} & s_{x1} \\ s_{1x} & s_{11} \end{pmatrix}^{-1} - \begin{pmatrix} 0 \\ 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix}^\top \right\},
\]
which was previously obtained in Prentice & Pyke (1979) and Qin & Zhang (1997). When $\beta^* = 0$, this matrix equals the first $2 \times 2$ principal submatrix of (8), and therefore $\hat{\beta}$ falls along the line $\{\beta : (\mu, 1) \beta = 0\}$ asymptotically to the order of $n^{-1/2}$.

3. Discussion

For model (1), the profile empirical likelihood of $\beta_1$ is well behaved with the usual properties of profile likelihoods whether or not $\beta^* = 0$, in contrast to that of $\hat{\beta}$, which has an irregularity when $\beta^* = 0$. The reason for this difference is that $\alpha_0(\beta_1)$ is a function of $(\beta_1, G_0)$ such that $\alpha_0 = 0$ whenever $\beta_1 = 0$. The profile empirical likelihood of $\beta$ is not defined on the line $\{(0, \beta_0)^\top : \beta_0 \neq 0\}$, which intersects each nonempty neighbourhood of $\beta^*$ when $\beta^* = 0$. Therefore, the profile empirical likelihood of $\beta$ does not have a proper Hessian at $\beta^*$ when $\beta^* = 0$.

The maximum likelihood estimator $\hat{\beta}$ is consistent and $n^{1/2}(\hat{\beta} - \beta^*)$ is asymptotically normal whether or not $\beta^* = 0$. This estimator can be derived in two ways, either by combining $\hat{\beta}_1$ and $\hat{\alpha}_0(\beta_1)$ from the profile empirical likelihood of $\beta_1$ in Section 2 or by maximizing the profile empirical likelihood of $\beta$ in Qin (1999). However, these two derivations are not equally useful for studying the asymptotic behavior of $\hat{\beta}$. Qin (1999) aimed to show that the profile empirical likelihood of $\beta$ is well-defined and has a maximizer in a small neighbourhood of $\beta^*$, but the first property fails when $\beta^* = 0$. In contrast, such arguments can be properly applied to the profile empirical likelihood of $\beta_1$ to obtain Proposition 2 and Corollary 1. Alternatively in our proofs, we focus on the estimating equations and apply the theory of M-estimators.

An interesting point is that when $\beta^* = 0$, $\hat{\beta}$ lies away from the line $\{(0, \beta_0)^\top : \beta_0 \neq 0\}$, on which the profile empirical likelihood of $\beta$ is not defined. This separation allows $\beta$, obtained as a maximizer to the
profile empirical likelihood of $\beta$, to still behave asymptotically in the usual manner except with perfect correlation. Furthermore, the restriction of $\hat{\beta}$ to a straight line suggests that when $\beta^* = 0$, the loglikelihood ratio statistic for $\beta$ converges in distribution to a chi-squared variable with one degree of freedom. Formally, this result follows from Proposition 2 (iii), because the profile empirical likelihood of $\beta$ at $\hat{\beta}$, or $\beta^* = 0$, is identical to that of $\beta_1$ at $\hat{\beta}_1$, or $\beta_1^* = 0$.

**Acknowledgement**

The author thanks an associate editor and two referees for helpful comments. This research was supported by the U.S. National Science Foundation.

**Appendix**

**Technical details**

*Proof of Proposition 1.* (i) If $G_0(\{x\}) > 0$ for some point $x$ outside the sample $\{x_{ji} : i = 1, \ldots, n_j, j = 1, \ldots, m\}$, then $l(\beta_1, G_0) < l(\beta_1, G'_0)$, where $G'$ is a probability distribution such that $G'_0(\{x\}) = 0$ and $G'_0(\{x'\}) = G_0(\{x'\})/(1 - G_0(\{x\}))$ for $x' \neq x$. Therefore, we restrict $G_0$ to probability distributions supported on the sample, and write $w_{ji} = G_0(\{x_{ji}\})$. For a fixed $\beta_1$, we maximize the loglikelihood

$$
\sum_{j=1}^{m} \sum_{i=1}^{n_j} \left[ \log \lambda_j + (1 - \lambda_j) \exp(\beta_0 + \beta_1 x_{ji}) \right] + \log w_{ji},
$$

over $w_{ji} (i = 1, \ldots, n_j, j = 1, \ldots, m)$ and $\beta_0$ subject to the normalization conditions

$$
1 = \sum_{j=1}^{m} \sum_{i=1}^{n_j} w_{ji}, \quad 1 = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \exp(\beta_0 + \beta_1 x_{ji}) w_{ji}.
$$

By introducing Lagrange multipliers $n\alpha_0$ and $n\alpha_1$ and setting the derivatives with respect to $w_{ji}$ and $\beta_0$ to 0, we obtain

$$
0 = \frac{1}{w_{ji}} - n\alpha_0 - n\alpha_1 \exp(\beta_0 + \beta_1 x_{ji}), \quad \alpha_1 = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{(1 - \lambda_j) \exp(\beta_0 + \beta_1 x_{ji})}{\lambda_j + (1 - \lambda_j) \exp(\beta_0 + \beta_1 x_{ji})}.
$$

Multiplying the left-hand equation here by $w_{ji}$ and summing over the sample yields $\alpha_0 + \alpha_1 = 1$. Therefore, $w_{ji}s$ are of the claimed form, equation (2) results from the right-hand equation, and (3) and (4) correspond to the normalization conditions.

(ii) The claimed inequality follows because for any distribution $G_0$ supported on the sample, any $0 \leq \alpha_0 \leq 1$, and $\beta_0 = -\log(\int \exp(\beta_1 x) \, dG_0)$, we have $l(\beta_1, G_0) \leq \kappa(\beta, \alpha_0)$. In fact, by Jensen’s inequality,

$$
l(\beta_1, G_0) - \kappa(\beta, \alpha_0) = \sum_{j=1}^{n} \sum_{i=1}^{n_j} \left[ G_0(\{x_{ji}\}) n \{\alpha_0 + (1 - \alpha_0) \exp(\beta_0 + \beta_1 x_{ji})\} \right]
$$

$$
\leq n \log \left[ \sum_{j=1}^{n} \sum_{i=1}^{n_j} G_0(\{x_{ji}\}) \{\alpha_0 + (1 - \alpha_0) \exp(\beta_0 + \beta_1 x_{ji})\} \right] = 0. \quad \square
To find the Gradient and Hessian of $\kappa(\beta, \alpha_0)$, write $z_{ji} = (x_{ji}, 1)^T$ $(i = 1, \ldots, n_j, j = 1, \ldots, m)$. By taking derivatives, we find

$$
\frac{1}{n} \frac{\partial \kappa}{\partial \alpha_0} = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{-1 + \exp(\beta^T z_{ji})}{\alpha_0 + (1 - \alpha_0) \exp(\beta^T z_{ji})},
$$

$$
\frac{1}{n} \frac{\partial \kappa}{\partial \beta} = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left\{ \frac{(1 - \lambda_j) \exp(\beta^T z_{ji}) z_{ji}^T}{\lambda_j + (1 - \lambda_j) \exp(\beta^T z_{ji})} - \frac{(1 - \alpha_0) \exp(\beta^T z_{ji}) z_{ji}^T}{\alpha_0 + (1 - \alpha_0) \exp(\beta^T z_{ji})} \right\}.
$$

By taking second derivatives, we find

$$
\frac{1}{n} \frac{\partial^2 \kappa}{\partial \alpha_0^2} = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{\{(1 - \exp(\beta^T z_{ji}))^2}{\alpha_0 + (1 - \alpha_0) \exp(\beta^T z_{ji})}^2.
$$

$$
\frac{1}{n} \frac{\partial^2 \kappa}{\partial \beta \partial \beta^T} = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left\{ \frac{\lambda_j (1 - \lambda_j) \exp(\beta^T z_{ji}) z_{ji} z_{ji}^T}{\lambda_j + (1 - \lambda_j) \exp(\beta^T z_{ji})^2} - \frac{\alpha_0 (1 - \alpha_0) \exp(\beta^T z_{ji}) z_{ji} z_{ji}^T}{\alpha_0 + (1 - \alpha_0) \exp(\beta^T z_{ji})^2} \right\}.
$$

$$
\frac{1}{n} \frac{\partial^2 \kappa}{\partial \beta \partial \alpha_0} = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{\exp(\beta^T z_{ji}) z_{ji}}{\alpha_0 + (1 - \alpha_0) \exp(\beta^T z_{ji})^2}.
$$

Proof of Proposition 2. (i) Fix $\beta_1 = \beta_1^\ast$. Equations (2)–(4) are equivalent to $0 = \partial \kappa / \partial \gamma|_{\beta_1=\beta_1^\ast}$. The individual terms in $\partial \kappa / \partial \gamma$ and $\partial^2 \kappa / (\partial \gamma \partial \gamma^T)$ are uniformly bounded by constants for $\gamma$ in a neighbourhood of $\gamma^\ast$. By the asymptotic theory of M-estimators, the equation $0 = \partial \kappa / \partial \gamma|_{\beta_1=\beta_1^\ast}$ admits a solution $\hat{\gamma}(\beta_1^\ast) = \gamma^\ast + O_p(n^{-1/2})$. More specifically,

$$
\hat{\gamma}(\beta_1^\ast) - \gamma^\ast = \left( \frac{\partial^2 \kappa}{\partial \gamma \partial \gamma^T} \right)^{-1} \frac{\partial \kappa}{\partial \gamma} \bigg|_{\beta_1=\beta_1^\ast, \gamma=\gamma^\ast} + o_p(n^{-1/2}).
$$

By a Taylor expansion of $(\partial \beta_1 / \partial \beta_1)(\beta_1^\ast)$, given by (5), with $\hat{\gamma}(\beta_1^\ast)$ around $\gamma^\ast$, we find

$$
\left. \frac{1}{n} \frac{\partial \beta_1}{\partial \beta_1} \right|_{\beta_1=\beta_1^\ast} = \frac{1}{n} \left\{ \frac{\partial \kappa}{\partial \beta_1} - \frac{\partial^2 \kappa}{\partial \beta_1 \partial \gamma^T} \left( \frac{\partial^2 \kappa}{\partial \gamma \partial \gamma^T} \right)^{-1} \frac{\partial \kappa}{\partial \gamma} \bigg|_{\beta_1=\beta_1^\ast, \gamma=\gamma^\ast} \right\} + o_p(n^{-1/2}).
$$

By Lemma 1, as $n \to \infty$, $n^{-1/2}(\partial \beta_1 / \partial \beta_1)(\beta_1^\ast)$ converges in distribution to a multivariate normal variable with mean zero and variance matrix

$$
V = \left( 1, -\Sigma_{10} \Sigma_{00}^{-1} \right) V' \left( \begin{array}{c} -\Sigma_{00}^{-1} \Sigma_{01} \\ 1 \end{array} \right) = \sigma_{11} - \Sigma_{10} \Sigma_{00}^{-1} \Sigma_{01},
$$

where the simplification is due to $(1, -\Sigma_{10} \Sigma_{00}^{-1})(S_{12}, s_{22})^T = 0$. Next, by Lemma 1 and (6), $-n^{-1}(\partial^2 \beta_1 / \partial \beta_1^2)(\beta_1^\ast)$ converges in probability to $U = \sigma_{11} - \Sigma_{10} \Sigma_{00}^{-1} \Sigma_{01}$. It follows immediately that $U = V$.

(ii) Note that $\hat{\beta}_1$ satisfies $0 = \partial \beta_1 / \partial \beta_1$ if and only if $[\hat{\beta}_1, \hat{\gamma}(\hat{\beta}_1)]$ satisfies $0 = \partial \kappa / \partial \beta_1$ and $0 = \partial \kappa / \partial \gamma$. The individual terms in $\partial \kappa / \partial \beta_1$ and $\partial \kappa / \partial \gamma$ and the second-order derivatives are uniformly bounded by quadratic functions of observations from $G_0$ and $G_1$ for $(\beta_1, \gamma)$ in a neighborhood of $(\beta_1^\ast, \gamma^\ast)$. By the asymptotic theory of M-estimators, there exists a solution $[\beta_1, \hat{\gamma}(\beta_1)] = (\beta_1^\ast, \gamma^\ast) + O_p(n^{-1/2})$. By a Taylor expansion of $(\partial \beta_1 / \partial \beta_1)(\beta_1^\ast)$, we obtain

$$
\hat{\beta}_1 - \beta_1^\ast = -\left( \frac{\partial^2 \beta_1}{\partial \beta_1^2} \right)^{-1} \left. \frac{\partial \beta_1}{\partial \beta_1} \right|_{\beta_1=\beta_1^\ast} + o_p(n^{-1/2}),
$$

which together with (i) implies that $n^{1/2}(\hat{\beta}_1 - \beta_1^\ast)$ converges in distribution to $N(0, V)$. 
(iii) By a second-order Taylor expansion of \( pl(\beta_1^*) \), we obtain
\[
pl(\hat{\beta}_1) - pl(\beta_1^*) = -\frac{1}{2} \left. \frac{\partial^2 pl}{\partial \beta_1^2} \right|_{\beta_1 = \beta_1^*} (\hat{\beta}_1 - \beta_1^*)^2 + o_p(1),
\]
which together with (i) implies that \(-2[pl(\beta_1^*) - pl(\hat{\beta}_1)]\) converges in distribution to a chi-squared distribution with one degree of freedom, as \( n \to \infty \).

**Proof of Corollary 1.** The proof of Proposition 2 (ii) shows that \( n^{1/2} \{\hat{\beta}_1 - \beta_1^*, \hat{\gamma}(\hat{\beta}_1) - \gamma^*\} \) converges in distribution to a multivariate normal variable with mean 0 and variance \( U^{-1}VU^{-1} \). The matrix \( U^{-1}VU^{-1} \) reduces to the claimed form, because \( U^{-1}(S_{12}, s_{22})U^{-1} = (0, 0, 1)^T \) and hence
\[
U^{-1} \begin{pmatrix} 0 & S_{12} \\ 0 & s_{22} \end{pmatrix} U^{-1} = \begin{pmatrix} 0 & 0 \\ S_{21} & s_{22} \end{pmatrix}.
\]

**References**


[Received March 2008. Revised September 2008]