Web-based Supplementary Materials for "Adjusted Exponentially Tilted Likelihood with Applications to Brain Morphology"

Hongtu Zhu, Haibo Zhou, Jiahua Chen, Yimei Li, Jeffrey Lieberman, and Martin Styner
1. Web Appendix A.

**Algorithm for Computing \( \tilde{\beta}_{aET} \)**

The algorithm for computing \( \tilde{\beta}_{aET} \) consists of two steps. The first is to use the modified Newton-Raphson algorithm proposed by Chen et al. (2002) to compute the profile of \( \ell_{aET}(\lambda, \beta) \), denoted by \( L^*(\beta) = \inf_{\lambda} \ell_{aET}(\lambda, \beta) \). For given \( \beta \) and \( a_n \), \( L^*(\beta) \) is calculated as described in the following pseudo-code:

1. Compute \( g_i = g(z_i, \beta) \) for \( i = 1, \ldots, n \) and \( g_{n+1} = -a_n \bar{g}_n \).

2. Set the initial value for the Lagrange multiplier \( \lambda^0 = 0 \). Initialize iteration number \( k = 0 \), and let \( \gamma = 1 \) and \( \varepsilon = 10^{-8} \) be the step size in the iteration and the tolerance level respectively.

3. Compute the first and second partial derivatives of \( R_1(\lambda) = \sum_{i=1}^{n+1} \exp(\lambda^T g_i) \) with respect to \( \lambda \) evaluated at \( \lambda^k \). Let these be \( \hat{R}_1 \) and \( R_1 \) and further compute \( \Delta = -\hat{R}_1^{-1} \hat{R}_1 \).

   If \( \| \Delta \| < \varepsilon \) stop the iteration, report \( \lambda^k \), and go to Step (6).

4. Compute \( \delta = \gamma \Delta \). If \( R_1(\lambda^k + \delta) > R_1(\lambda^k) \), let \( \gamma = \gamma/2 \), and repeat this step. Otherwise, continue to the next step.

5. Let \( \lambda^{k+1} = \lambda^k + \delta \) and \( \gamma = (k+1)^{-1/2} \). Increase the count \( k \) by 1. Return to Step (3).

6. Report \( \lambda^k \) and the value of \( L^*(\beta) = R_1(\lambda^k) \).

The convergence of the above algorithm for calculating \( L^*(\beta) \) is guaranteed (Chen et al., 2002). The second step is to use the simplex method to optimize \( L^*(\beta) \). A set of good initial values of \( \beta \) will be helpful in this step. We use the least square estimate of \( \beta \) in this paper.

2. Web Appendix B.

To obtain the asymptotic results, we place the following conditions on \( g(z, \beta) \):

(a) The parameter space of \( \beta, \mathcal{B} \) is compact;

(b) \( \beta_0 \), an interior point of \( \mathcal{B} \), is the unique solution to \( E\{g(z, \beta)\} = 0 \);
(c) $g(\beta) = g(z, \beta)$ is continuous at each $\beta \in B$ with probability one;
(d) $E\{\sup_{\beta \in B} \|g(z, \beta)\|^{\alpha}\} < \infty$ for some $\alpha > 2$;
(e) $\Omega = E\{g_i(\beta_0)g_i(\beta_0)^T\}$ is finite and nonsingular.

**Lemma.** Under Conditions (a)-(e), (i) $\hat{\lambda} = O_p(n^{-1/2})$; (ii) $\hat{\beta} \rightarrow \beta_0$.

**Proof:** Introducing the adjustment function $g_{n+1}$ ensures all equations involved have proper solutions, particularly the existence of the sample point $(\hat{\beta}, \hat{\lambda})$. For simplicity, we have omitted the subindex $aET$. Since a rigorous proof can be tedious, and at the same time, our approach mirrors that of Newey and Smith (2004), we decide to provide only a scratch of the proof.

Define

$$P(\beta, \lambda) = \sum_{i=1}^{n+1} \exp\{\lambda^T g_i(\beta)\}. $$

The duality of the ET implies that $(\hat{\beta}, \hat{\lambda})$ is a saddle point of $P(\beta, \lambda)$. Hence, for any vector $\bar{\lambda}$,

$$P(\hat{\beta}, \bar{\lambda}) \geq P(\hat{\beta}, \hat{\lambda}) \geq P(\beta_0, \hat{\lambda}) \geq P(\beta_0, \lambda_0)$$

where,

$$\lambda_0 = \arg \min \, P(\beta_0, \lambda).$$

By Conditions (b) and (e), it is simple to show that $\lambda_0 = O_p(n^{-1/2})$. This result can be employed to expand $P(\beta_0, \lambda_0)$ and obtain

$$P(\beta_0, \lambda_0) = (n + 1) - \frac{1}{2} + o_p(1).$$

For any positive constant $\epsilon > 0$, put

$$\bar{\lambda} = -\frac{\epsilon \sum_{i=1}^{n+1} g(z_i; \hat{\beta})}{\sqrt{n\|\sum_{i=1}^{n+1} g(z_i; \beta)\|}}.$$

Clearly, $\bar{\lambda} = O_p(n^{-1/2})$ so that $\bar{\lambda}^T g(z_i; \hat{\beta}) = o_p(1)$ uniformly including the case of $i = n + 1$. 


Thus, the following Taylor’s expansion is justified

\[ P(\tilde{\beta}, \tilde{\lambda}) = \sum_{i=1}^{n+1} \exp\{-\tilde{\lambda}^T g(z_i; \tilde{\beta})\} \]

\[ = (n + 1) - en^{-1/2} \left\| \sum_{i=1}^{n+1} g(z_i; \tilde{\beta}) \right\| + \frac{\epsilon^2}{2n} \left\{ \sum_{i=1}^{n+1} g(z_i; \tilde{\beta}) \}^T \{ \sum_{i=1}^{n+1} g(z_i; \tilde{\beta}) g(z_i; \tilde{\beta})^T \} \{ \sum_{i=1}^{n+1} g(z_i; \tilde{\beta}) \} \right\|^2 \]

By the uniform law of large numbers, and the condition (d), the third term, excluding the factor \(\epsilon^2\), in the above equation is \(O_p(1)\).

The fact that \(P(\tilde{\beta}, \tilde{\lambda}) > P(\beta_0, \lambda_0)\) implies that

\[ en^{-1/2} \left\| \sum_{i=1}^{n+1} g(z_i; \tilde{\beta}) \right\| \leq \frac{1}{2} + o_p(1) \]

for any sufficiently small \(\epsilon > 0\). This implies that

\[ \left\| \sum_{i=1}^{n+1} g(z_i; \tilde{\beta}) \right\| = O_p(n^{1/2}). \tag{1} \]

From Condition (b) that \(\beta_0\) is the only solution to \(E\{g(z, \beta)\} = 0\), and (a) that the parameter space is compact, we conclude that (1) implies \(\tilde{\beta} = \beta_0 + o_p(1)\). This result can in turn be used to expand \(P(\tilde{\beta}, \tilde{\lambda})\) to show that \(\tilde{\lambda} = O_p(n^{-1/2})\). Hence, the conclusions of the Lemma are obtained.

With this Lemma, we go a step further to find the limiting distributions.

**Theorem 1.** Under conditions (a)-(e), and assume further that \(D^T \Omega^{-1} D\) has full rank where \(D = E\{\partial_\beta g(z, \beta_0)\}\). Then

(i) \(\sqrt{n}(\tilde{\beta} - \beta_0)\) converges to \(N\{0, (D^T \Omega^{-1} D)^{-1}\}\) in distribution;

(ii) under the null hypothesis of \(R\beta = b_0\), \(LR_{\alpha ET}\) converges to a \(\chi^2(r)\) distribution in distribution, where \(r\) is the row rank of \(R\).

**Proof:** The so-called first order conditions (Newey and Smith, 2004) reflect the fact that \(\tilde{\beta}\)
and \( \lambda \) are stationary points of \( P(\beta, \lambda) \). In the case of aET, they translate into

\[
\sum_{i=1}^{n+1} g(z_i; \tilde{\beta}) \exp\{\lambda^T g(z_i; \tilde{\beta})\} = 0; \quad (2)
\]

\[
\sum_{i=1}^{n+1} \lambda^T \partial_\beta g(z_i; \tilde{\beta}) \exp\{\lambda^T g(z_i; \tilde{\beta})\} = 0. \quad (3)
\]

Expanding the terms in above equations at \( \beta = \beta_0, \lambda = 0 \), and ignore the high-order terms, we obtain

\[
\left\{ \sum_{i=1}^{n+1} \partial_\beta g(z_i; \beta_0) \right\} (\tilde{\beta} - \beta_0) + \left\{ \sum_{i=1}^{n+1} g(z_i; \beta_0)g(z_i; \beta_0)^T \right\} \tilde{\lambda} = -\sum_{i=1}^{n+1} g(z_i; \beta_0); \quad (4)
\]

\[
(\sum_{i=1}^{n+1} 0)^T (\tilde{\beta} - \beta_0) + \left\{ \sum_{i=1}^{n+1} \partial_\beta g(z_i; \beta_0) \right\}^T \tilde{\lambda} = 0. \quad (5)
\]

We tactically included \( \sum 0 \) so that the presentation is more balanced.

Under conditions of the theorem, \( n^{-1} \sum_{i=1}^{n+1} g(z_i; \beta_0)g(z_i; \beta_0)^T \to \Omega; \sum_{i=1}^{n+1} \partial_\beta g(z_i; \beta_0) \to D \) almost surely, and \( n^{-1/2} \sum_{i=1}^{n+1} g(z_i; \beta_0) \to N(0, \Omega) \) in distribution. Thus, ignoring the high order term, we get

\[
\tilde{\beta} - \beta_0 = -(n + 1)n^{-1}(D\Omega^{-1}D^T)^{-1}D\Omega^{-1} \sum_{i=1}^{n+1} g(z_i; \beta_0)
\]

\[
\tilde{\lambda} = -(n + 1)^{-1} \Omega^{-1} \{ \Omega - D^T(D\Omega^{-1}D^T)^{-1}D \} \Omega^{-1} \sum_{i=1}^{n+1} g(z_i; \beta_0).
\]

The conclusion (i) is hence obvious.

To prove (ii), we work on quadratic expansions of \( \ell_{aET}(\tilde{\lambda}, \tilde{\beta}) \) under the full model, and under the restriction \( R\beta = b_0 \), when the null model is right.
Under the full model, 
\[ \ell_{aET}(\tilde{\lambda}, \tilde{\beta}) = \log \left( (n + 1)^{-1} \sum_{i=1}^{n+1} \exp \{ \tilde{\lambda}^T g(z_i, \tilde{\beta}) \} \right) \]
\[ = \log \left[ 1 + (n + 1)^{-1} \tilde{\lambda}^T \sum_{i=1}^{n+1} g(z_i; \tilde{\beta}) + \frac{1}{2} \tilde{\lambda}^T \left\{ (n + 1)^{-1} \sum_{i=1}^{n+1} g^2(z_i; \tilde{\beta}) \right\} \tilde{\lambda} + o_p(n^{-1}) \right] \]
\[ = \log \left[ 1 + (n + 1)^{-1} \tilde{\lambda}^T \sum_{i=1}^{n+1} g(z_i; \beta_0) + \tilde{\lambda}^T \left\{ (n + 1)^{-1} \sum_{i=1}^{n+1} g^2(z_i; \beta_0) \right\} (\tilde{\beta} - \beta_0) \right. \]
\[ + \frac{1}{2} \tilde{\lambda}^T \left\{ (n + 1)^{-1} \sum_{i=1}^{n+1} g^2(z_i; \beta_0) \right\} \tilde{\lambda} + o_p(n^{-1}) \left. \right] \]
\[ = \log \left[ 1 + (n + 1)^{-1} \tilde{\lambda}^T \sum_{i=1}^{n+1} g(z_i; \beta_0) + \tilde{\lambda}^T D(\tilde{\beta} - \beta_0) + \frac{1}{2} \tilde{\lambda}^T \Omega \tilde{\lambda} + o_p(n^{-1}) \right]. \]

Substituting the expansion of \( \tilde{\beta} - \beta_0 \) and \( \tilde{\lambda} \), we get
\[ \ell_{aET}(\tilde{\lambda}, \tilde{\beta}) = \log \left[ 1 + \frac{1}{2(n + 1)^2} \left\{ \sum_{i=1}^{n+1} g(z_i; \beta_0) \right\}^T \{ \Omega - D^T(D^T \Omega^{-1} D)^{-1} D \} \left\{ \sum_{i=1}^{n+1} g(z_i; \beta_0) \right\} + o_p(n^{-1}) \right]. \]

Clearly, \( 2(n+1)\ell_{aET} \) is approximately a quadratic function of \( (n+1)^{-1/2} \sum_{i=1}^{n+1} g(z_i; \beta_0) \) which is asymptotically normal with variance matrix \( \Omega \). It is easy to verify that the quadratic form has chi-square distribution (Serfling, 1980.).

Next, we study the null model specified by \( R\beta = b_0 \), which can be viewed as a special case of the full model: the number of free parameters in \( \beta \) is reduced. More specifically, we can identify a set of free parameters \( \alpha \) such that \( \beta = \varphi(\alpha) \) for some linear function \( \varphi \), and hence
\[ g(z; \beta) = g(z; \varphi(\alpha)). \]

The rest of derivations remains the same, except the matrix \( D \) must be re-defined in terms of \( \alpha \). Thus, by denoting the Lagrange multiplier and the aET estimator as \( \tilde{\lambda}_\alpha \) and \( \tilde{\alpha} \), we conclude that \( 2(n+1)\ell_{aET}(\tilde{\lambda}_\alpha, \tilde{\alpha}) \) is also approximated by a quadratic function of \((n + 1)^{-1/2} \sum_{i=1}^{n+1} g(z_i; \beta_0) \) which is asymptotically chi-square distributed.

In summary, the aET likelihood ratio statistic
\[ 2(n + 1) \{ \ell_{aET}(\tilde{\lambda}, \tilde{\beta}) - \ell_{aET}(\tilde{\lambda}_\alpha, \tilde{\alpha}) \} \]
is approximated by a non-negative definite quadratic form in 

\[(n + 1)^{-1/2} \sum_{i=1}^{n+1} g(z_i; \beta_0),\]

\[= 2 \sum_{i=1}^{n+1} g(z_i; \beta_0),\]

can be written as the difference of two quadratic forms that are both asymptotically chisquare distributed. Therefore, it is also approximately chisquare distributed. The degrees of freedom is given by the row rank of \(R\).

3. Web Appendix C.

3.1 First Set of Simulation Results

The first simulated data are from

\[y_i = \beta + \varepsilon_i,\] (6)

for \(i = 1, \cdots, n\), where \(\varepsilon_i\) was a random error with zero mean. For model (6), the Type I errors rates of \(L R_{aE T}\) and the \(t\)-test were found accurate for all sample sizes \((n = 20, 40, \text{or } 60)\) considered and for all different distributions of error terms at the 1\% significant levels (Fig. 1). We observed that Type II error rates for \(L R_{aE T}\) and the \(t\)-test were similar under symmetric errors and for all sample sizes (Fig. 1). However, the power of \(L R_{aE T}\) to reject the null hypothesis increased modestly when the distribution of error terms followed skewed distribution \(\chi^2(3) - 3\) (Fig. 1 (b)). This decline in Type I error was caused by the fact that the distribution of the \(t\)-test was not in fact \(t\)-distributed.

[Figure 1 about here.]

3.2 Second Set of Simulation Results

The second set of simulations generated data from

\[y_i = x_i^T \beta + \varepsilon_i\] (7)

for \(i = 1, \cdots, n\), where \(\varepsilon_i\) was a random error with zero mean and \(\beta = (\beta_1, \beta_2, \beta_3)^T\) is a \(3 \times 1\) unknown parameter vector. For model (7), except that \(L R_{aE T}\) had slightly inflated Type I error under Gaussian distribution with severe heterogeneous variances, the Type I
errors rates for the $t$-test and $LR_{aET}$ were accurate for all sample sizes ($n = 20$, 40, or 60) considered and for almost all different distributions of error terms at the 1% significant level (Fig. 2). The type II error rates of $LR_{aET}$ and the $t$-test were similar under symmetric errors and for all sample sizes (Fig. 2). However, for skewed distribution $\chi^2(3) - 3$ and Gaussian distribution with severe variance heterogeneity, the power of $LR_{aET}$ to reject the null hypothesis increased modestly (Figs. 2b and 2c). Consistent with our expectations, the statistical power for rejecting the null hypothesis increased with the sample size $n$.

[Figure 2 about here.]

References


Figure 1. Rejection rates of the t test and $LR_{aET}$ for different sample sizes 20, 40, and 60 under four different error distributions at significance level 0.01. All panels show estimated Type I and II error rates. Four error distributions are $N(0,1)$, $\chi^2(3) - 3$, $0.5N(2,1) + 0.5N(-2,1)$, and $t(3)$. 
Figure 2. Rejection rates of the $t_1$ test and $LR_{aET}$ for different sample sizes 20, 40, and 60 under six different error distributions at significance level 0.01. All panels show estimated Type I and II error rates. Six error distributions are $N(0,1)$, $\chi^2(3) - 3$, heter1, heter2, $0.5N(2,1) + 0.5N(-2,1)$, and $t(3)$. Moreover, heter1 and heter2 denote Gaussian distributions with heterogeneous variance $\sigma(i)$. For heter1, $\sigma(i) = \exp(u)$ when $x_{i2} = 0$ and $\sigma(i) = \exp(u+1)$ when $x_{i2} = 1$, where $u \sim N(0,1)$; however, for heter2, $\sigma(i) = 1$ for $x_{i2} = 0$ and $\sigma(i) = 4$ for $x_{i2} = 1$. This figure appears in color in the electronic version of this article.