

Mean log-likelihood and higher order approximations

BY N. REID AND D.A.S. FRASER

Department of Statistics, University of Toronto, Toronto, Canada M5S 3G3

reid@utstat.toronto.edu, dfraser@utstat.toronto.edu

SUMMARY

Higher order approximations to p -values can be obtained from the log-likelihood function and a reparameterization that can be viewed as a canonical parameter in an exponential family approximation to the model. This approach clarifies the connection between Skovgaard (1996) and Fraser et al. (1999a), and shows that the Skovgaard approximation can be obtained directly using the mean log-likelihood function.

Some key words: approximate pivot; Fraser information; Kullback–Leibler distance; p^* approximation; tangent exponential model

1. INTRODUCTION

Parametric likelihood inference is often based on first order approximations to standard summary statistics from the likelihood, such as the likelihood ratio statistic or the standardized maximum likelihood estimate. For more accurate inference, refinements are needed to improve the first order approximations, to find the distribution of the relevant statistic, and to properly take account of nuisance parameters in the multi-parameter setting. A great many examples are available that illustrate the failure of first-order methods in models with large numbers of nuisance parameters, and there is a similar wealth of examples illustrating the accuracy of higher order approximations. Recent books include Barndorff-Nielsen & Cox (1994), Severini (2000), Butler (2007) and Brazzale et al. (2007).

In this note we establish a simple connection between the higher order approximation due to Skovgaard (1996) and that of Fraser et al. (1999a), and this in turn gives quite direct links to the approximations of Barndorff-Nielsen (1986) and Severini (1998). This connection is established by emphasizing the role of exponential family models as approximations to the original model in obtaining higher order approximations for parametric inference.

2. CANONICAL PARAMETRIZATION

We assume a model $f(y; \theta)$ for the observation $y = (y_1, \dots, y_n)$, where $\theta \in \mathbb{R}^d$, and $\theta = (\psi, \lambda)$ is typically partitioned into a parameter of interest $\psi \in \mathbb{R}^{d_1}$ and a nuisance parameter $\lambda \in \mathbb{R}^{d_0}$. We further assume that the components of y are continuous and independent; extensions are discussed briefly in §5. The log-likelihood function is $\ell(\theta; y) = \log f(y; \theta)$ with observed information function $j(\theta) = -\partial^2 \ell(\theta; y) / \partial \theta \partial \theta'$ and maximum likelihood estimate $\hat{\theta}$ assumed to be the solution of the score equation $\partial \ell(\theta; y) / \partial \theta = 0$.

First consider the case of a full exponential family model,

$$f(y; \theta) = \exp\{\varphi_1(\theta)t_1(y) + \dots + \varphi_d(\theta)t_d(y) - c(\theta) - h(y)\};$$

49 the canonical variable $t(y)$ is sufficient and has dimension d . Both the model $f(y; \theta)$ and the
 50 model for the marginal distribution of $t(y)$ have the same observed log-likelihood function
 51 $\ell^0(\theta) = \varphi^\top(\theta)t^0 - c(\theta)$ and the same canonical parameter $\varphi(\theta)$, that is, the same up to affine
 52 transformations $a\varphi + b$ where a and b are constants. This canonical parameter can be obtained
 53 as the derivative of the log-likelihood function with respect to t at any point, say the observed
 54 data t^0 . It could also be obtained by differentiation with respect to y at y^0 in any d linearly
 55 independent directions not tangent to the maximum likelihood surface $\hat{\theta} = \hat{\theta}^0$.

56 In more general models it is possible to find an approximating exponential family model, the
 57 tangent exponential model (Fraser, 1990), which is an exponential family model that has the
 58 same observed log-likelihood function as the original model and the same first derivative with
 59 respect to the data at the observed data point. The tangent exponential model at the data point y^0
 60 is defined from the model $f(y; \theta)$ as

$$61 \quad f_{\text{TEM}}(s; \theta) ds = \exp\{\varphi(\theta)^\top s + \ell(\theta; y^0)\} h(s) ds, \quad (1)$$

62 where s is a nominal variable that can be viewed as a score variable $s(y) = \ell_\theta(\hat{\theta}^0; y)$, and $\ell(\theta; y^0)$
 63 and $\varphi(\theta; y^0)$ are defined from the original model as

$$64 \quad \ell(\theta; y^0) = \log f(y^0; \theta), \quad \varphi(\theta; y^0)^\top = \ell_{;V}(\theta; y^0). \quad (2)$$

65 In (2) the notation $\ell_{;V}$ denotes differentiation in the sample space in directions given by the
 66 columns of a matrix V : the construction of V is described below at (3). It is convenient to use the
 67 score variable in (1), so that $s^0 = s(y^0) = 0$, but s is simply the argument of the density, which
 68 is defined on \mathbb{R}^d . The tangent exponential model was developed for the derivation of accurate
 69 approximations to tail probabilities, which is discussed in the next section. For the moment we
 70 note that it is completely determined by the pair of functions $\{\ell(\theta), \varphi(\theta)\}$.

71 It follows from (1) that $\log f_{\text{TEM}}(\theta; s^0) = \ell(\theta; y^0)$ and $(\partial/\partial s) \log f_{\text{TEM}}(\theta; s^0) =$
 72 $\varphi(\theta; y^0)^\top = \ell_{;V}(\theta; y^0)$. This shows that the tangent exponential model and the original model
 73 have the same log-likelihood function and the same sample space derivative of the log-likelihood
 74 function. The tangent exponential model implements conditioning on an approximately ancillary
 75 statistic, to construct an approximate model on \mathbb{R}^d from the original model on \mathbb{R}^n , so strictly
 76 should be written $f_{\text{TEM}}(s; \theta | a)$. This conditioning is implemented through the choice of V .
 77 Subject to this choice, the tangent exponential model is unique up to and including terms of
 78 $O(n^{-1})$.

79 Denote by $z = (z_1, \dots, z_n)$ a vector of pivotal statistics $z_i = z_i(y_i; \theta)$, which could be simply
 80 the vector of distribution functions $F(y_i; \theta)$. We use this vector of pivotals to define an $n \times d$
 81 matrix V , with rows V_i , by

$$82 \quad V = - \left(\frac{\partial z}{\partial y} \right)^{-1} \left(\frac{\partial z}{\partial \theta} \right) \Bigg|_{\theta=\hat{\theta}, y=y^0} = \frac{dy}{d\theta} \Bigg|_{\theta=\hat{\theta}, y=y^0}, \quad (3)$$

83 where the final expression in (3) is the derivative of y for fixed pivotal z .

84 Then in more detail, the canonical parameter in (1) is

$$85 \quad \varphi(\theta)^\top = \ell_{;V}(\theta; y) \Big|_{y=y^0} = \frac{d\ell(\theta; y)}{dV} \Bigg|_{y=y^0} = \sum_i \frac{\partial \ell(\theta; y^0)}{\partial y_i} V_i. \quad (4)$$

86 It is shown in Fraser & Reid (1995) that there is a statistic a that is ancillary to $O(n^{-1})$, that
 87 the subspace in \mathbb{R}^n determined by fixing the value of $a = a^0$ is spanned by the column vectors
 88 in V , and that $f_{\text{TEM}}(s; \theta | a^0)$ approximates the exact distribution of $f(s; \theta | a)$ to $O(n^{-1})$, in
 89 $O(n^{-1/2})$ -neighbourhoods of $\hat{\theta}$ and y^0 : details are summarized in the Appendix.

A different approach to higher order approximation was proposed by Skovgaard (1996), who obtained estimates of the directions of conditioning. We show in §3 that Skovgaard's approximation to p -values can be obtained by using the exponential family model (1), but with a different canonical parameter. Let $I(\theta; \theta_0)$ designate a mean log-likelihood function:

$$I(\theta; \theta_0) = E_{\theta_0} \{ \ell(\theta; y) \} = \int \ell(\theta; y) f(y; \theta_0) dy. \quad (5)$$

This function is related to the Kullback–Leibler distance, which is $I(\theta_0; \theta_0) - I(\theta; \theta_0)$. It also arises in studies of the robustness of likelihood inference, where it is called the Fraser information (Kent, 1982). A new version of φ , say $\bar{\varphi}$, of the canonical parameter for the model (1) is defined by differentiating the function $I(\theta; \hat{\theta})$ instead of $\ell(\theta; y)$;

$$\bar{\varphi}(\theta) = \left. \frac{\partial}{\partial \theta_0} I(\theta; \theta_0) \right|_{\theta_0 = \hat{\theta}} = \frac{\partial}{\partial \hat{\theta}} I(\theta; \hat{\theta}). \quad (6)$$

The exponential family model with canonical parameter $\bar{\varphi}(\theta)$ is an $O(n^{-1})$ approximation to the tangent exponential model (1). Averaging the log-likelihood in the calculation of $\bar{\varphi}$ eliminates dependence on the approximate ancillary, which in turn reduces the accuracy of tail area approximations based on $\bar{\varphi}$, discussed in the next section. On the other hand for many models the calculation of $\bar{\varphi}$ is simpler than the calculation of φ using (4).

In linear exponential families, with log-likelihood $\ell(\theta) = \theta^T s - c(\theta)$, we have $I(\theta; \theta_0) = \theta^T c'(\theta_0) - c(\theta)$, giving $\bar{\varphi}^T = \theta^T c''(\hat{\theta})$, where $c''(\theta)$ is a $d \times d$ matrix, so that both φ and $\bar{\varphi}$ are equal to the canonical parameter of the model (up to an affine transformation). Outside this special setting, broadly speaking the φ version is easier to compute in transformation families and the $\bar{\varphi}$ version is easier to compute in curved exponential families.

Example 1. Suppose y_i follows a one-parameter location model $f_0(y_i - \theta)$. A sample y_1, \dots, y_n admits an exact ancillary statistic, $a = (a_1, \dots, a_n) = (y_1 - \hat{\theta}, \dots, y_n - \hat{\theta})$, and the $n \times 1$ vector V from the pivotal $z_i = y_i - \theta$ is simply a vector of 1's. Thus

$$\varphi(\theta) = \ell_{;V}(\theta; y^0) = \sum \frac{\partial}{\partial y_i} \log f_0(y_i - \theta) = -\frac{\partial}{\partial \theta} \sum g_0(y_i - \theta) = -\sum g'_0(y_i - \theta), \quad (7)$$

writing $g_0(\cdot) = \log f_0(\cdot)$. From (5)

$$I(\theta; \theta_0) = E_{\theta_0} \sum \log f_0(y_i - \theta) = n \int g_0(z - \theta + \theta_0) f_0(z) dz,$$

showing as expected that $I(\theta; \theta_0) = I_0(\theta - \theta_0)$, say. Then

$$\bar{\varphi}(\theta) = n \int g'_0(y - \theta) f_0(y - \hat{\theta}) dy. \quad (8)$$

On dividing (7) and (8) by n we see that $\varphi(\theta)$ is the nonparametric bootstrap estimate of the expected value of $g'_0(Y - \theta)$ and $\bar{\varphi}(\theta)$ is the parametric bootstrap estimate of the same quantity.

Example 2. If the density of y_i is a $(d, 1)$ curved exponential family

$$f(y_i; \theta) = \exp[\alpha(\theta)^T t(y_i) - c\{\alpha(\theta)\} - d(y_i)], \quad (9)$$

where t and α are $d \times 1$ vectors, then writing $t. = \sum t(y_i)$,

$$I(\theta; \theta_0) = \alpha(\theta)^T E_{\theta_0}(t.) - nc\{\alpha(\theta)\} = n[\alpha(\theta)^T c'\{\alpha(\theta_0)\} - c\{\alpha(\theta)\}]$$

and $\bar{\varphi}(\theta) = \alpha(\theta)^T \bar{a}$, say, where

$$\bar{a}_i = n \sum_j \frac{\partial^2 c\{\alpha(\theta)\}}{\partial \alpha_i \partial \alpha_j} \frac{\partial \alpha_j}{\partial \theta} \Big|_{\theta=\hat{\theta}^0}.$$

The calculation of φ requires a vector of pivotal statistics which could be taken as $F(y_i; \theta)$, the cumulative distribution function for the i th observation, giving $\varphi(\theta) = \alpha(\theta)^T \hat{d}$, say, where

$$\hat{d}_j = - \sum_{i=1}^n \frac{\partial t_j(y_i)}{\partial y_i} \frac{\partial F(y_i; \theta) / \partial \theta}{f(y_i; \theta)} \Big|_{\theta=\hat{\theta}^0, y=y^0}.$$

Depending on the particular exponential family model, one or the other may be easier to calculate analytically or numerically.

The exponential family models, using φ or $\bar{\varphi}$ as the canonical parameters, provide approximate conditional densities on \mathbb{R}^d and are related to the p^* approximation of Barndorff-Nielsen (1986), given by

$$p^*(\hat{\theta}; \theta | a) = c|j(\hat{\theta})|^{1/2} \exp\{\ell(\theta; \hat{\theta}, a) - \ell(\hat{\theta}; \hat{\theta}, a)\} \quad (10)$$

for some exact or approximate ancillary statistic a . To use this formula requires details concerning the ancillary a , in particular what its contour $a = a^0$ looks like at y^0 . Let $y = y^0 + Vt$ be the first order Taylor series expansion of y given $a(y) = a^0$ in terms of coordinates t relative to tangents V to the surface $a(y) = a^0$. Then the tangent exponential model (1) with (2) re-expresses (10) to first derivative at y^0 using the V determined by the ancillary a in (10). This first derivative approximation to (10) gives third order inference at y^0 : see the Appendix.

3. TAIL AREA APPROXIMATIONS

Exponential family models are particularly useful for approximating tail areas, as saddlepoint approximation arguments presented in Daniels (1954) and Barndorff-Nielsen (1986) can be used to construct an approximate p -value function for inference for a scalar parameter ψ with relative error $O(n^{-3/2})$. The approximation depends only on the canonical parameter and the observed log-likelihood function. More generally, in any asymptotic model $f(y; \theta)$, where y and θ have the same dimension, an approximate p -value function for a scalar parameter of interest ψ can be obtained from the observed log-likelihood function $\ell(\theta; y^0)$ and the derivative $\varphi(\theta) = \partial \ell(\theta; y^0) / \partial y$ (Fraser, 1990).

This saddlepoint argument applied to the tangent exponential model (1) leads to an approximate p -value function $\Phi(r^*)$, where $\Phi(\cdot)$ is the standard normal distribution function, $\phi(\cdot)$ is its density. The approximate pivot $r^* = r^*(\psi; y)$, defined as

$$r^* = r + \frac{1}{r} \log \left(\frac{q}{r} \right), \quad (11)$$

where

$$r = \pm [2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_\psi)\}]^{1/2}, \quad (12)$$

and

$$q = \frac{|\varphi(\hat{\theta}) - \varphi(\hat{\theta}_\psi) \quad \varphi_\lambda(\hat{\theta}_\psi)|}{|\varphi_\theta(\hat{\theta})|} \frac{|j(\hat{\theta})|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}. \quad (13)$$

In (12) $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$ is the constrained maximum likelihood estimator, and $\hat{\lambda}_\psi$ is assumed to be the solution of $\partial \ell(\theta) / \partial \lambda = 0$. If θ is a scalar then $r = \pm [2\{\ell(\hat{\theta}) - \ell(\theta)\}]^{1/2}$. The approximate pivot $r = r(\psi; y)$ is usually called the likelihood root, and to first order, i.e. with relative error $O(n^{-1/2})$, follows a standard normal distribution. The nuisance parameter λ is accommodated in this approximate pivot by the simple expedient of maximization. In (13) $\varphi(\cdot)$ is a $d \times 1$ vector, the matrix $\varphi_\lambda(\cdot)$ is the last $d - 1$ columns of $\varphi_\theta(\cdot)$, the matrix of derivatives of φ with respect to θ , and $j_{\lambda\lambda}(\theta)$ is the submatrix of the observed Fisher information function corresponding to the nuisance parameter components λ . Conditioning on an approximate ancillary and adjusting for nuisance parameters are both incorporated in the approximate pivot q . A detailed discussion of the r^* approximation and its numerical implementation in R (R Development Core Team, 2007), in the bundle `h0a`, is given in Brazzale & Davison (2008). Brazzale et al. (2007, Ch. 9) describe an algorithm for implementing r^* from $\ell(\theta)$ and a function to compute the vector array V ; the derivatives needed to compute r^* are then obtained numerically.

An important property of (11) is that it is completely determined by the pair of functions $\{\ell(\theta), \varphi(\theta)\}$, and their derivatives with respect to θ and for this the original model can be replaced by the tangent exponential model $f_{\text{TEM}}(s; \theta)$.

The original r^* approximation of Barndorff-Nielsen (1986) was derived from the p^* approximation to the conditional density of the maximum likelihood estimator assuming an explicit ancillary statistic was available. The corresponding expression for q is (13) with $\varphi(\theta)$ replaced by $\ell_{;\hat{\theta}}(\theta) = \partial \ell(\theta; \hat{\theta}, a) / \partial \hat{\theta}$:

$$q_{BN} = \frac{|\ell_{;\hat{\theta}}(\hat{\theta}; \hat{\theta}, a) - \ell_{;\hat{\theta}}(\hat{\theta}_\psi; \hat{\theta}, a) \quad \ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi)| \quad |j(\hat{\theta})|^{1/2}}{|\ell_{\theta;\hat{\theta}}(\hat{\theta}; \hat{\theta}, a)| \quad |j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}. \quad (14)$$

The normal approximation to r^* using this version of q also has relative error $O(n^{-3/2})$, provided that a is a second order ancillary statistic. As noted by a reviewer, φ and $\ell_{;\hat{\theta}}$ are affinely equivalent if the vectors V in (3) are tangent to the ancillary a in (10), in which case $q = q_{BN}$: see the Appendix and Fraser et al. (1999a, §2.3).

Skovgaard (1996) derived an approximation to (14) by showing that

$$\ell_{;\hat{\theta}}(\hat{\theta}; \hat{\theta}, a) - \ell_{;\hat{\theta}}(\hat{\theta}_\psi; \hat{\theta}, a) \doteq i^{-1}(\hat{\theta}) j(\hat{\theta}) \text{cov}_{\hat{\theta}}\{\ell_\theta(\hat{\theta}), \ell(\hat{\theta}) - \ell(\hat{\theta}_\psi)\}, \quad (15)$$

$$\ell_{\theta;\hat{\theta}}(\hat{\theta}_\psi) \doteq i^{-1}(\hat{\theta}) j(\hat{\theta}) \text{cov}_{\hat{\theta}}\{\ell_\theta(\hat{\theta}), \ell_\theta(\hat{\theta}_\psi)\}, \quad (16)$$

where the covariance terms are evaluated before the first argument of $\ell(\cdot; \cdot)$ is replaced: for example,

$$\text{cov}_{\theta_0}\{\ell_\theta(\theta_0; y), \ell(\theta_1; y)\} \Big|_{\theta_0=\hat{\theta}, \theta_1=\hat{\theta}_\psi}.$$

From the definition of $\bar{\varphi}(\theta)$ at (6) we see that Skovgaard's version of q is identical to (13) with φ replaced by $\bar{\varphi}$, as

$$\bar{\varphi}(\theta) = \text{cov}_{\hat{\theta}}\{\ell_\theta(\hat{\theta}), \ell(\theta)\}, \quad (17)$$

$$\bar{\varphi}_\theta(\theta) = \text{cov}_{\hat{\theta}}\{\ell_\theta(\hat{\theta}), \ell_\theta(\theta)\}, \quad (18)$$

$\bar{\varphi}_\theta(\hat{\theta}) = i(\hat{\theta})$, and $\ell_{\theta;\hat{\theta}}(\hat{\theta}) = j(\hat{\theta})$. Skovgaard (1996) also noted that the determinant in the numerator of (13) can be expressed as

$$|\varphi_\theta(\hat{\theta}_\psi)| [\varphi_\theta^{-1}(\hat{\theta}_\psi) \{\varphi(\hat{\theta}) - \varphi(\hat{\theta}_\psi)\}]_1,$$

241 where the choice of the first component of the vector in square brackets assumes that ψ is the
242 first component of θ .

243 The tangent exponential model approximates the conditional density at y^0 to $O(n^{-1})$, but
244 when the pair of functions $\{\ell(\theta; y^0), \varphi(\theta; y^0)\}$ from this model is used to derive the distribu-
245 tion function at y^0 , using (11), (12) and (13), the $O(n^{-1})$ term vanishes, and the error in the
246 tail area approximation is $O(n^{-3/2})$ (Fraser & Reid, 1993; Andrews et al., 2005). Skovgaard's
247 expressions do not require specification of the ancillary statistic a or its tangent vectors V ; the
248 resulting r^* approximation has relative error $O(n^{-1})$, for moderate deviations, and $O(n^{-1/2})$ for
249 large deviations (Skovgaard, 1996). The role of the ancillary conditioning on the relative error is
250 outlined in the Appendix. Severini (1999) proposed using empirical estimates of the covariances
251 in (17) and (18), although numerical work presented in Severini (2000, Ch. 7.5) suggests that
252 this may be numerically rather unstable.

253 254 255 4. EXAMPLES

256 *Example 3.* Suppose $(y_{1i}, y_{2i}), i = 1, \dots, n$ follow independent exponential distributions
257 with means $1/\theta$ and θ , respectively. The pivotal functions for computing V using (3) are
258 $z_{1i} = y_{1i}\theta$ and $z_{2i} = y_{2i}/\theta$, giving $V = \hat{\theta}^{-1}(-y_{11}, \dots, -y_{1n}, y_{21}, \dots, y_{2n})^T$ and

$$259 \quad \varphi(\theta) = \frac{n}{\hat{\theta}}(\theta\bar{y}_1 - \frac{1}{\hat{\theta}}\bar{y}_2) = na\theta \left(\frac{1}{\hat{\theta}^2} - \frac{1}{\theta^2} \right),$$

260 where $\hat{\theta} = (\bar{y}_2/\bar{y}_1)^{1/2}$, $a = (\bar{y}_1\bar{y}_2)^{1/2}$; in this model a is exactly ancillary.

261 The information function $I(\theta; \theta_0) = -n\theta/\theta_0 - n\theta_0/\theta$, leading to

$$262 \quad \bar{\varphi}(\theta) = n\theta \left(\frac{1}{\hat{\theta}^2} - \frac{1}{\theta^2} \right);$$

263 since $\varphi(\theta)$ and $\bar{\varphi}(\theta)$ differ only by a scalar multiple both give $q = (na/2)^{1/2}\{(\hat{\theta}/\theta) - (\theta/\hat{\theta})\}$,
264 so that the r^* approximation using $\bar{\varphi}$ is in this case accurate to $O(n^{-3/2})$.

265 Using Skovgaard's original expressions (17) and (18), the numerator and denominator of (13)
266 contain the scalar multiple a , which simply cancels, although ancillary information is incorpo-
267 rated into q through a scaling factor. Severini (2000, Ch. 6) shows that $\bar{\varphi}$ and (14) give the same
268 expression for q , which is expected as V is tangent to the curve $a = a^0$ at $\hat{\theta}^0$.

269 *Example 4.* Suppose now that (y_{1i}, y_{2i}) follow a bivariate normal distribution with means 0,
270 variances 1, and covariance θ . This example was used in Reid (2003) to illustrate the construction
271 of several approximate ancillary statistics, and in Reid (2005) to illustrate the accuracy of the
272 r^* approximation. The ancillary directions V are computed using the pivotal statistics $z_{1i} =$
273 $(y_{1i} + y_{2i})^2/\{2(1 + \theta)\}$ and $z_{2i} = (y_{1i} - y_{2i})^2/\{2(1 - \theta)\}$, leading to

$$274 \quad V_i = \left(\frac{y_{2i} - \hat{\theta}y_{1i}}{1 - \hat{\theta}^2}, \frac{y_{1i} - \hat{\theta}y_{2i}}{1 - \hat{\theta}^2} \right), \varphi(\theta) = \ell_{;V}(\theta) = \frac{n\{\theta(t - \hat{\theta}s) - (s - \hat{\theta}t)\}}{(1 - \theta^2)(1 - \hat{\theta}^2)},$$

275 where $t = \sum(y_{1i}^2 + y_{2i}^2)/(2n)$ and $s = \sum(y_{1i}y_{2i})/n$. Using (5) gives $\bar{\varphi}(\theta) = (n\theta)/(1 - \theta^2)$.
276 An expression for (14) in this example requires explicit expression of $\ell(\theta; y)$ as a function of
277 $\hat{\theta}$ and an approximately ancillary statistic. This can be obtained following Barndorff-Nielsen
278 & Wood (1998) by embedding the model in a full exponential family, but the details are quite
279 cumbersome. Table 1 presents a simulation showing that computing the tail area approximation
280 using φ or $\bar{\varphi}$ gives essentially the same numerical results.

Table 1. Normal correlation coefficient: approximate p -values (percent) computed using the r^* approximation, with φ and $\bar{\varphi}$ used in (13). Based on 10,000 simulations with sample size $n = 5$.

θ	Exact	left tail (%)					right tail (%)				
		0.1	0.5	1.0	2.5	5.0	5.0	2.5	1.0	0.5	0.1
0.9	r^*	0.10	0.51	0.98	2.51	5.03	5.02	2.44	0.97	0.47	0.10
	\tilde{r}^*	0.10	0.51	0.98	2.51	5.02	5.12	2.59	1.04	0.51	0.11
0.7	r^*	0.12	0.50	1.04	2.51	5.05	5.30	2.60	1.05	0.54	0.14
	\tilde{r}^*	0.12	0.50	1.04	2.51	5.04	6.61	3.09	1.26	0.67	0.19
0.5	r^*	0.12	0.49	1.01	2.56	5.14	5.64	2.85	1.18	0.61	0.19
	\tilde{r}^*	0.12	0.49	1.01	2.56	5.13	6.45	3.20	1.35	0.72	0.24
0.3	r^*	0.10	0.53	1.10	2.63	5.22	5.46	2.77	1.14	0.59	0.18
	\tilde{r}^*	0.10	0.53	1.10	2.61	5.21	5.86	2.94	1.20	0.64	0.21
0	r^*	0.13	0.56	1.07	2.63	5.15	5.07	2.53	1.04	0.53	0.11
	\tilde{r}^*	0.14	0.56	1.09	2.65	5.24	5.12	2.54	1.04	0.55	0.13

Example 5. We assume that we have n independent observations with $y_i = x_i(\beta) + \sigma\epsilon_i$, where $x_i(\beta)$ is a known nonlinear function of a d -dimensional parameter β and some known covariates, and ϵ_i follows a standard normal distribution. The canonical parameter $\varphi(\theta)$ computed using (4) is given in Fraser et al. (1999b) as

$$\varphi(\theta)^\top = \frac{1}{\sigma^2} \sum \{y_i - x_i(\beta)\} \{\hat{\epsilon}_i, X_i(\hat{\beta})\}$$

where $\hat{\epsilon}_i = \{y_i - x_i(\hat{\beta})\}/\hat{\sigma}$ is the standardized residual, and $X_i(\beta) = \partial x_i(\beta)/\partial \beta^\top$. The canonical parameter obtained using (6) is

$$\bar{\varphi}(\theta)^\top = \frac{1}{\sigma^2} [n\hat{\sigma}, \sum \{x_i(\hat{\beta}) - x_i(\beta)\} X_i(\hat{\beta})].$$

For more direct comparison the first component of $\varphi(\theta)$ is equal to

$$\frac{1}{\sigma^2} [n\hat{\sigma} + \sum \{y_i - x_i(\hat{\beta})\} \{x_i(\hat{\beta}) - x_i(\beta)\}] = n\hat{\sigma}/\sigma^2 + O(1)$$

after replacing $x_i(\hat{\beta}) - x_i(\beta)$ by the first term of its Taylor series expansion, $X_i(\hat{\beta})(\hat{\beta} - \beta)$. The vector that makes up the remaining components of φ is

$$\frac{1}{\sigma^2} \left\{ \sum y_i X_i(\hat{\beta}) - \sum x_i(\beta) X_i(\hat{\beta}) \right\}$$

so that $\varphi(\theta)$ and $\bar{\varphi}(\theta)$ are affinely equivalent, exactly for the β components, and approximately for the σ component.

More detailed formulas of for nonlinear regression are given in Brazzale et al. (2007, Ch. 8), and the R package `n1reg` implements both the Skovgaard and the Fraser-Reid versions of q .

In linear and nonlinear regression models with non-normal error, the existence of an explicit pivotal statistic for computing the array V ensures that $\varphi(\theta)$ is relatively easy to compute; formulae are given in Brazzale et al. (2007, Ch. 8.6) and the linear regression case is implemented in the R package `marq`. The computation of $\bar{\varphi}(\theta)$ requires a multi-dimensional integral that can rarely be evaluated explicitly.

337 *Example 6.* On the other hand in mixed effects linear regression models $\bar{\varphi}(\theta)$ is much easier
 338 to compute than $\varphi(\theta)$. We assume the marginal model is $y \sim N\{X\beta, V(\rho)\}$, where ρ indexes
 339 the parameters in the covariance matrix. If this model is obtained by integrating over the mixed
 340 effects linear model $y = X\beta + Zb + \epsilon$, the structure of V is $V(\rho) = Z\Omega Z^T + \Sigma$, where Ω is
 341 the covariance matrix for the random effects b and Σ is the covariance matrix for the errors; ρ
 342 would then include the unknown parameters in both Ω and Σ . The information function $I(\theta; \theta_0)$
 343 is readily evaluated:

$$\begin{aligned} 344 \quad I(\theta; \theta_0) &= E_{\theta_0} \left\{ -\frac{1}{2} (y - X\beta)^T V^{-1}(\rho) (y - X\beta) - \frac{1}{2} \log |V(\rho)| \right\} \\ 345 \quad &= -\frac{1}{2} \log |V(\rho)| - \frac{1}{2} \text{tr} \{V(\rho_0) V^{-1}(\rho)\} - \frac{1}{2} (\beta_0 - \beta)^T X^T V^{-1}(\rho) X (\beta_0 - \beta) \end{aligned}$$

349 from which we have the β and ρ components of $\bar{\varphi}(\theta)$

$$350 \quad \bar{\varphi}^{(\beta)}(\theta) = -X^T V^{-1}(\rho) X (\hat{\beta} - \beta), \quad \bar{\varphi}^{(\rho)}(\theta) = \frac{1}{2} \frac{\partial}{\partial \rho} \text{tr} \{V(\hat{\rho}) V^{-1}(\rho)\}.$$

353 These agree with Lyons & Peters (2000), who used the covariance versions (17) and (18).

354 To construct $\varphi(\theta)$ using the sample space derivative $\ell_{;V}$ requires specifying the pivotal statis-
 355 tics z . In the dependent data setting it is not obvious how to do this, but one approach is to
 356 construct the residual for each component y_i after regression on the preceding components
 357 (y_{i-1}, \dots, y_1) ; this is carried out in a University of Toronto dissertation by S. Iglesias-Gonzalez.
 358 Numerical investigation there confirms that the r^* approximation using φ is indeed more ac-
 359 curate than Skovgaard's version. However, the calculation of V depends on the ordering of the
 360 components of y , which is a somewhat unsatisfactory aspect of the φ version in this setting.

363 5. DISCUSSION

364 The asymptotic theory underlying the r^* approximation is most easily developed for indepen-
 365 dent and identically distributed observations, but extensions to regression settings are relatively
 366 straightforward, as long as the information function is proportional to the sample size: this will
 367 typically involve some conditions on the design matrix X for the regression. In the case of de-
 368 pendent observations, such as discussed in Example 4.4, detailed conditions on the nature of the
 369 dependence would be needed, to ensure that information continues to accumulate at a rate pro-
 370 portional to n . We are not aware of a thorough treatment of this case, but limited simulations in
 371 Lyons & Peters (2000), Guolo et al. (2006) and in the dissertation of Iglesias-Gonzalez suggest
 372 that the approximations are still useful in these dependent data settings.

373 The $\bar{\varphi}$ version of the approximating exponential model can be used in discrete models, whereas
 374 the derivation of φ depends on continuity. Davison et al. (2006) suggest an alternative to φ for
 375 discrete models obtained by replacing $dy/d\theta$ in (3) by $dE(s)/d\theta$, where s is the score variable
 376 if the discrete model is a curved (or full) exponential family model, and is a locally defined
 377 score variable in more general settings. In the curved exponential family case, this leads to the
 378 same canonical parameter as that defined by using $I(\theta; \theta_0)$, but in the more general setting the
 379 expressions are different.

380 In models where the analytical calculation of $E_{\theta_0}\{\ell(\theta; y)\}$ is not possible, it may be possible to
 381 use a bootstrap approximation to estimate this mean, not only at θ_0 but at enough values of θ to be
 382 able to use smoothing to get the derivatives numerically. This is very close to the approach taken
 383 by Severini (1999), who estimated the derivatives empirically, but might have better numerical
 384

properties, as empirical estimates of means are usually more accurate than empirical estimates of covariances.

The simplicity of the expression for $\bar{\varphi}$ means it can easily be used with likelihood-like objects, such as partial likelihood or composite likelihood, although it is not clear what the asymptotic properties of this approach might be.

Skovgaard (2001) extended his approach to obtain a higher order approximation for vector parameters, w^* , by deriving a correction of the log-likelihood ratio statistic $w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\}$. This correction is a function of the likelihood ratio statistic, of $\bar{\varphi}(\hat{\theta}) - \bar{\varphi}(\hat{\theta}_\psi)$, and also of a standardized version of the score statistic $\ell_\psi(\hat{\theta}_\psi)$. The derivative $\partial I(\theta; \theta_0)/\partial \theta$ gives the mean value of the score under θ_0 , which might lead to an equivalent version of w^* . A quite different approach to the multivariate setting is to use a directional test; this can be derived from the pair $\{\ell(\theta), \bar{\varphi}(\theta)\}$ along the lines of that developed in Skovgaard (1988) and Fraser & Massam (1988).

ACKNOWLEDGEMENTS

We would like to thank the reviewers for helpful comments on an earlier draft. This work was partially supported by the Natural Sciences and Engineering Council of Canada. We are indebted to A. Chouldechova for the calculations presented in Table 1.

APPENDIX

Ancillary conditioning in the tangent exponential model.

The tangent exponential model is defined on \mathbb{R}^d , although it is computed from the original model on \mathbb{R}^n , through (2). Assume for the moment that we are able to find a one-to-one transformation $y = y(s, a)$, where $y = (y_1, \dots, y_n)$, $s = (s_1, \dots, s_d)$ and $a = (a_1, \dots, a_{n-d})$ is an exactly ancillary statistic, so that $f(y; \theta) = g_1(s | a; \theta)g_2(a) |J(s, a)|$. Fraser et al. (1999a, eq. (2.4) – (2.5)) show that

$$(\partial/\partial s) \log g_1(s | a^0; \theta)|_{s=s^0} = \ell_{;V}(\theta; y^0)$$

where $V = \partial y(s, a)/\partial s|_{s=s^0, a=a^0}$. Thus the sample space derivative used to define φ can be computed from the conditional model or more easily from the original model, provided the tangent directions can be determined. As noted by a reviewer, this shows that $\ell_{;\hat{\theta}}$ and $\ell_{;V}$ define the same tangent exponential model, where V is taken to be $\partial y/\partial \hat{\theta}$ for fixed a , although the derivative $\ell_{;\hat{\theta}}$ is not usually computable. In more detail, let $V = (v_1, \dots, v_d)$ be a linearly independent set of vectors tangent to the surface $a(y) = a^0$ at $\hat{\theta}^0$; and let $W = (w_1, \dots, w_{n-d})$ be a linearly independent set of vectors orthogonal to the tangent plane $y^0 + \mathcal{L}(V)$ to the ancillary contour. Then local coordinates near y^0 using these vectors as the basis are

$$y = y^0 + (V \ W) \begin{pmatrix} t \\ u \end{pmatrix},$$

where t gives coordinates with respect to V and u with respect to W . Denote by

$$H = \left. \frac{\partial \hat{\theta}}{\partial t} \right|_{t=0, u=0}$$

the local $d \times d$ Jacobian for the change from t to $\hat{\theta}$ given $a = a^0$. We can write

$$y = y^0 + (\bar{V} \ W) \begin{pmatrix} \bar{t} \\ u \end{pmatrix}$$

where $\bar{t} = Ht$, $\bar{V} = VH^{-1}$ and y is now expressed as a function of \bar{t} and u . It follows that

$$\left. \frac{\partial}{\partial \hat{\theta}} \ell(\theta; y) \right|_{y=y^0} = \left. \frac{\partial}{\partial \bar{t}} \ell(\theta; y) \right|_{\bar{t}=0, u=0} = \left. \frac{d}{d\bar{V}} \ell(\theta; y) \right|_{y=y^0}.$$

where \bar{V} is the array of tangent vectors to the ancillary at the data point y^0 relative to the $\hat{\theta}$ coordinates. From (11) and (13), we see that the formulas are unaffected if the vectors V are replaced by any linearly equivalent VH^{-1} for non-singular H , so there is no gain in calculating derivatives with respect to $\hat{\theta}$ coordinates. In fact in most cases a differentiation with respect to $\hat{\theta}$ would need computation of $\partial \ell(\theta; y) / \partial y_i$ for individual coordinates as indicated in (2).

For the density approximations (1) and (10) we only require a to be ancillary to $O(n^{-1})$. It is rare to be able to find an explicit expression even for such an approximate ancillary, but $\ell_{;V}(\theta; y^0)$ can be computed without making the transformation from y to (s, a) , using (3). It is shown in Fraser & Reid (1995, §5) for scalar θ and Fraser & Reid (2002, §6) for vector θ that there exists a statistic a which is ancillary to $O(n^{-1})$. This statistic is obtained by constructing a location model approximation to $f(y; \theta)$; the ancillary statistic for this approximate model is directly available from the residuals, and the vectors V defined in (3) are tangent to the surface defined by fixed a . This statistic is an exact first derivative ancillary, and thus is ancillary to $O(n^{-1/2})$, but can be modified to be ancillary to $O(n^{-1})$ without changing its tangent vectors V .

From tangent exponential models to p-values

The tangent exponential model approximates the original model to $O(n^{-1})$, in moderate deviation neighbourhoods of y^0 . If θ is a scalar, the saddlepoint approximation applied to (1) gives the r^* approximation with r as the signed log-likelihood root, and q as the standardized maximum likelihood estimator $(\hat{\varphi} - \varphi) \hat{j}_{\varphi\varphi}^{1/2}$. If θ is a vector, and the parameter of interest is a scalar component of θ , then an additional step is needed to eliminate the nuisance parameters. This step is outlined Reid (2003, §3.3), and it is very similar to the argument in Skovgaard (1996, §9.3); it requires integrating a density of what Skovgaard (2001, §5.4) calls ‘‘Laplace type’’.

Assume for notational simplicity that after conditioning on an approximate ancillary statistic, as described above, we have a model $f(y; \theta)$ on \mathbb{R} and a scalar parameter θ . A more explicit expression for the tangent exponential model can be obtained from Taylor series expansion of the log-likelihood function, expanded in both θ and y . Andrews et al. (2005, eq.(2.6)) give the coefficients for this expansion, and show that it has a particularly simple structure, after both y and θ have been centered at y^0 and $\hat{\theta}^0$ respectively, scaled to have unit second derivatives, and transformed to variables of the form $y + a_1 n^{-1/2} y^2 + a_2 n^{-1} y^3$, $\theta + b_1 n^{-1/2} \theta^2 + b_2 n^{-1} \theta^3$, where terms of $O(n^{-3/2})$ are ignored. The resulting expansion for the log-density of the transformed variable y with transformed parameter ϑ , where $y^0 = 0$ and $\vartheta^0 = 0$ is

$$c + P_{1n}(y) + P_{2n}(\vartheta) + \vartheta y + \frac{\gamma}{4n} \vartheta^2 y^2, \tag{A1}$$

where c is the normalizing constant,

$$P_{1n}(y) = -\frac{\alpha_3 y}{2n^{1/2}} - \left(\frac{1}{2} - \frac{\alpha_4 - 2\alpha_3^2 - 5\gamma}{4n} \right) y^2 + \frac{\alpha_3}{6n^{1/2}} y^3 + \frac{\alpha_4 - 3\alpha_3^2 - 6\gamma}{24n} y^4,$$

is determined by requiring the density to integrate to 1, and

$$P_{2n}(\vartheta) = -\frac{1}{2} \vartheta^2 - \frac{\alpha_3}{6n^{1/2}} \vartheta^3 - \frac{\alpha_4}{24n} \vartheta^4,$$

where α_3 and α_4 and γ are $O(1)$. This model has the property that its cumulative distribution function, evaluated at 0, does not depend on γ , so for approximating the p -value to $O(n^{-3/2})$ we can use the simpler version of (A1)

$$c + P_{1n}(y) + P_{2n}(\vartheta) + \vartheta y$$

481 and this is the tangent exponential model with $\varphi(\vartheta) = \vartheta$ and $\ell(\vartheta) = P_{2n}(\vartheta)$.

482 The role of the ancillary is suppressed in this argument, but the coefficients α_3 , α_4 and γ depend on
 483 the approximate ancillary statistic. Thus even though the tangent exponential model with φ replaced by
 484 Skovgaard's version $\bar{\varphi}$ is also free of the non-exponential term $\gamma\vartheta^2 y^2/(4n)$, it can only approximate
 485 the true conditional model to $O(n^{-1})$, and hence give approximate conditional p -values to that order. It
 486 has been suggested by a reviewer that the unconditional p -values from Skovgaard's version might yet be
 487 accurate to $O(n^{-3/2})$, but it is not clear to us how this might be established.

488 However we now verify, for scalar θ , that $\bar{\varphi} = \varphi\{1 + O(n^{-1})\}$ via a Taylor series expansion, to
 489 $O(n^{-1})$ only, of the joint model for $\hat{\theta}$ and a . The dimension reduction from n to 1 in principle requires
 490 an ancillary statistic a of dimension $n - 1$, but it follows from §3 of Fraser & Reid (1995) that only a
 491 finite number of coordinates of a are needed for third order inference, and that the joint density can be
 492 approximated at $(\hat{\theta}^0, a^0)$ by an expansion of the form

$$493 f(y; \theta) = \exp\left\{-\frac{1}{2}(\hat{\vartheta} - \vartheta)^2 + \frac{\alpha_3}{6n^{1/2}}(\hat{\vartheta} - \vartheta)^3 + \frac{\delta}{n^{1/2}}a(\hat{\vartheta} - \vartheta)^2 - \frac{1}{2}a^2 + \frac{\gamma}{6n^{1/2}}a^3 + O(n^{-1})\right\}. \quad (A2)$$

495 In (A2) we assume the dimension of a is 1, although the general structure for a finite dimensional a is
 496 of the same form. We again use the notation ϑ in place of θ , but in (A2) the series expansion assumes
 497 that the original parameter θ has been reexpressed and standardized so that the model follows a location
 498 family form to second order, with observed Fisher information equal to 1. In (A1) the reexpression and
 499 standardization is to exponential family form. Both types of expansions are detailed in Andrews et al.
 500 (2005, §2), although in the approximate location case they give only the conditional model given a .

501 From (A2) we obtain $\varphi(\vartheta; y^0) = \varphi(\vartheta; \hat{\vartheta}^0, a^0)$ as

$$502 \frac{\partial}{\partial \hat{\vartheta}} \log f(y; \theta)|_{\hat{\vartheta}=\hat{\vartheta}^0} = \vartheta - \hat{\vartheta}^0 + \frac{\alpha_3}{2n^{1/2}}(\vartheta - \hat{\vartheta}^0)^2 - \frac{2a^0\delta}{n^{1/2}}(\vartheta - \hat{\vartheta}^0),$$

503 and dividing by $1 - 2a^0\delta/n^{1/2}$ we have that φ as a function of ϑ is affinely equivalent to

$$504 \varphi(\vartheta) = \vartheta - \hat{\vartheta}^0 + \frac{\alpha_3}{2n^{1/2}}(\vartheta - \hat{\vartheta}^0)^2 + O(n^{-1}).$$

505 This also shows that ignoring terms of $O(n^{-1})$ and higher, $\varphi(\vartheta, y^0)$ is free of the ancillary value a^0 .

506 We now use (A2) to calculate $I(\vartheta; \vartheta_0)$; for example

$$507 E_{\vartheta_0}(\hat{\vartheta} - \vartheta)^2 = 1 + (\vartheta_0 - \vartheta)^2 + \frac{\alpha_3}{n^{1/2}}(\vartheta_0 - \vartheta) + O(n^{-1}),$$

$$508 E_{\vartheta_0}(\hat{\vartheta} - \vartheta)^3 = 3(\vartheta_0 - \vartheta) + (\vartheta_0 - \vartheta)^3 + O(n^{-1}),$$

$$509 I(\vartheta; \vartheta_0) = -\frac{1}{2}(\vartheta_0 - \vartheta)^2 - \frac{1}{2} + \frac{\alpha_3}{6n^{1/2}}(\vartheta_0 - \vartheta)^3 + O(n^{-1})$$

510 and hence

$$511 \bar{\varphi}(\vartheta) = \vartheta - \hat{\vartheta}^0 + \frac{\alpha_3}{2n^{1/2}}(\vartheta - \hat{\vartheta}^0)^2 + O(n^{-1}).$$

512 If we used a more accurate $O(n^{-3/2})$ expansion in place of (A2) we would find dependence on a in the
 513 $O(n^{-1})$ term for φ , whereas $\bar{\varphi}$ can of course only depend on $\hat{\vartheta}$.

514 The extension to vector θ follows from expansions given in Cakmak et al. (1994) but the notation
 515 becomes rather cumbersome.

516 Finally, we note that the (10) can also be expressed in the form of a tangent exponential model, by
 517 expanding about $\hat{\theta}^0$ as follows:

$$518 p^*(\hat{\theta}; \theta | a^0) = c|j(\hat{\theta})|^{1/2} \exp\{\ell(\theta; \hat{\theta}, a^0) - \ell(\hat{\theta}; \hat{\theta}, a^0)\}$$

$$519 = c|j(\hat{\theta})|^{1/2} \exp\{\ell(\theta; \hat{\theta}^0, a^0) - \ell(\hat{\theta}^0; \hat{\theta}^0, a^0) + (\hat{\theta} - \hat{\theta}^0)^\top \{\ell_{;\hat{\theta}}(\theta; \hat{\theta}^0, a^0) - \ell_{;\hat{\theta}}(\hat{\theta}^0; \hat{\theta}^0, a^0)\}$$

$$520 = \exp\{(\hat{\theta} - \hat{\theta}^0)^\top \ell_{;\hat{\theta}}(\theta) + \ell(\theta)\} k(\hat{\theta}),$$

529 which compares directly with (1) with the score variable s replaced by $\hat{\theta} - \hat{\theta}^0$ and the canonical parameter
 530 replaced by $\ell_{,\hat{\theta}}$.

532
 533 REFERENCES

- 534 ANDREWS, D. A., FRASER, D. A. S. & WONG, A. (2005). Computation of distribution functions from likelihood
 535 information near observed data. *Journal of Statistical Planning and Inference* **134**, 180–193.
- 536 BARNDORFF-NIELSEN, O. & WOOD, A. (1998). On large deviations and choice of ancillary for p^* and r^* . *Bernoulli*
 537 **4**, 35–63.
- 538 BARNDORFF-NIELSEN, O. E. (1986). Inference on full or partial parameters based on the standardized signed log
 539 likelihood ratio. *Biometrika* **73**, 307–322.
- 540 BRAZZALE, A. R., DAVISON, A. & REID, N. (2007). *Applied Asymptotics*. Cambridge: Cambridge University
 541 Press.
- 542 BRAZZALE, A. R. & DAVISON, A. C. (2008). Accurate parametric inference for small samples. *Statistical Science*
 543 To appear.
- 544 BUTLER, R. W. (2007). *Saddlepoint Approximations with Applications*. Cambridge: Cambridge University Press.
- 545 CAKMAK, S., FRASER, D. & REID, N. (1994). Multivariate asymptotic model: exponential and location approxi-
 546 mations. *Utilitas Mathematica* **46**, 21–31.
- 547 DANIELS, H. E. (1954). Saddlepoint approximations in statistics. *Annals of Mathematical Statistics* **25**, 631–650.
- 548 DAVISON, A., FRASER, D. A. S. & REID, N. (2006). Improved likelihood inference for discrete data. *Journal of*
 549 *the Royal Statistical Society, Series B* **68**, 495–508.
- 550 FRASER, D. & REID, N. (2002). Ancillary information for statistical inference. In *Empirical Bayes and Likelihood*
 551 *Inference*, D. Ahmed & N. Reid, eds. New York: Springer-Verlag, pp. 185–210.
- 552 FRASER, D. A. S. (1990). Tail probabilities from observed likelihoods. *Biometrika* **77**, 65–76.
- 553 FRASER, D. A. S. & MASSAM, H. (1988). Conical tests: Observed levels of significance and confidence regions.
 554 *Statistische Hefte* **26**, 1–18.
- 555 FRASER, D. A. S. & REID, N. (1993). Third order asymptotic models: Likelihood functions leading to accurate
 556 approximations to distribution functions. *Statistica Sinica* **3**, 67–82.
- 557 FRASER, D. A. S. & REID, N. (1995). Ancillaries and third order significance. *Utilitas Mathematica* **47**, 33–53.
- 558 FRASER, D. A. S., REID, N. & WU, J. (1999a). A simple general formula for tail probabilities for frequentist and
 559 Bayesian inference. *Biometrika* **86**, 249–264.
- 560 FRASER, D. A. S., WONG, A. C. M. & WU, J. (1999b). Regression analysis, nonlinear or nonnormal: Simple and
 561 accurate p values from likelihood analysis. *Journal of the American Statistical Association* **94**, 1286–1295.
- 562 GUOLO, A., BRAZZALE, A. R. & SALVAN, A. (2006). Improved inference on a scalar fixed effect of interest in
 563 nonlinear mixed-effects models. *Computational Statistics and Data Analysis* **51**, 1602–1613.
- 564 KENT, J. T. (1982). Robust properties of likelihood ratio test. *Biometrika* **69**, 19–27.
- 565 LYONS, B. & PETERS, D. (2000). Applying Skovgaard’s modified directed likelihood statistics to mixed linear
 566 models. *Journal of Statistical Computation and Simulation* **65**, 225–242.
- 567 R DEVELOPMENT CORE TEAM (2007). *A Language and Environment for Statistical Computing*. Vienna, Austria:
 568 R Foundation for Statistical Computing.
- 569 REID, N. (2003). Asymptotics and the theory of inference. *Annals of Statistics* **31**, 1695–1731.
- 570 REID, N. (2005). Asymptotics and the theory of statistics. In *Celebrating Statistics: Papers in Honour of D.R. Cox*,
 571 A. Davison, Y. Dodge & N. Wermuth, eds. Oxford: Oxford University Press, pp. 73–88.
- 572 SEVERINI, T. A. (1998). An approximation to the modified profile likelihood function. *Biometrika* **85**, 403–411.
- 573 SEVERINI, T. A. (1999). An empirical adjustment to the likelihood ratio statistic. *Biometrika* **86**, 235–248.
- 574 SEVERINI, T. A. (2000). *Likelihood Methods in Statistics*. Oxford: Oxford University Press.
- 575 SKOVGAARD, I. M. (1988). Saddlepoint expansions for directional test probabilities. *Journal of the Royal Statistical*
 576 *Society, Series B* **50**, 3–32.
- SKOVGAARD, I. M. (1996). An explicit large-deviation approximation to one-parameter tests. *Bernoulli* **2**, 145–165.
- SKOVGAARD, I. M. (2001). Likelihood asymptotics. *Scandinavian Journal of Statistics* **28**, 3–32.