Asymptotics and the theory of inference

N. Reid
University of Toronto

May 29, 2002

Abstract

Asymptotic analysis has always been very useful for deriving distributions in statistics in cases where the exact distribution is unavailable. More importantly, asymptotic analysis can also provide insight into the inference process itself, suggesting what information is available and how this information may be extracted. The development of likelihood inference over the past twenty years provides an illustration of the interplay between techniques of approximation and statistical theory.

1 Introduction

The development of statistical theory has always relied on extensive use of the mathematics of asymptotic analysis, and indeed asymptotic arguments are an inevitable consequence of a frequency based theory of probability. This is so even in a Bayesian context, as all but the most specialized applications rely on some notion of long run average performance. Asymptotic analysis has also provided statistical methodology with approximations that have proved in many instances to be relatively robust. Most importantly, asymptotic arguments provide insight into statistical inference, by verifying that our procedures are moderately sensible, providing a framework for com-

paring competing procedures, and providing understanding of the structure of models.

One used to hear criticisms of asymptotic arguments on the grounds that in practice all sample sizes are finite, and often small, but this criticism addresses only the possibility that the approximations suggested by the analysis may turn out to be inaccurate; something that can be checked in applications of particular interest. The insights offered through asymptotics and the development of improved approximations using asymptotic expansions have effectively answered this criticism. Here are some simple examples.

A common technique in statistical consulting, often useful in helping the client to formulate the problem, is the "infinite data" thought experiment – what would the client expect to see with an arbitrarily large amount of data from the same experiment?

An early use of asymptotics for comparison of competing methodologies was Fisher's (1920) comparison of the variance of two competing estimators of scale,

$$\begin{aligned} s_1 &= \left\{ \frac{1}{(n-1)} \Sigma (x_i - \bar{x})^2 \right\}^{1/2} \\ s_2 &= c \frac{1}{n} \Sigma |x_i - \bar{x}| \end{aligned}$$

s_2 being scaled to have the same asymptotic mean as s_1. Fisher showed that s_1 has smaller asymptotic variance in independent, identically distributed sampling from the normal distribution, thereby in his view clinching the argument for its superiority. Interestingly, a further analysis of the model led Fisher to discover sufficiency, surely one of the most important insights into the structure of inference in the early development of theoretical statistics (Stigler, 1973).

A more recent example is the derivation of the minimax efficiency of

local linear or polynomial smoothing techniques of Stone (1980, 1982) and Fan (1993), reviewed in Hastie and Loader (1993), which led to such methods generally being preferred to competitors such as kernel estimates for problems of nonparametic density estimation and regression.

An early example of asymptotic theory providing an important insight into likelihood based inference was Neyman and Scott's (1948) paper showing that maximum likelihood estimators could be inconsistent or inefficient in problems with increasing numbers of nuisance parameters. More recently Smith (1985, 1989) showed that asymptotic behaviour of maximum likelihood estimators in nonregular models can be very different from that in regular models.

In the following we consider the insight offered by asymptotic analysis for inference based on the likelihood function. This is an area that has seen considerable development in the past twenty years, largely based on asymptotic expansions and improved approximation. Section 2 reviews the main asymptotic results in likelihood inference and mentions a number of other applications of asymptotics to areas of statistics of especial current interest. In Section 3 we provide additional detail on a particular type of approximation; that of approximating p-values in tests of significance. We emphasize here recent work of Barndorff-Nielsen and colleagues and of Fraser and Reid and colleagues. In Section 4 we discuss the gap between the theoretical development and applications with special emphasis on reviewing recent work that is aimed at narrowing this gap, and outlining work still needed.

2 Likelihood Asymptotics

2.1 First order theory

The main asymptotic results of likelihood based inference, presented in most graduate courses on statistical theory, can be summarized as follows:

i) the maximum likelihood estimator is consistent, asymptotically normal and asymptotically efficient

ii) the score statistic has mean zero and is asymptotically normally distributed

iii) the likelihood ratio statistic has an asymptotic chi-squared distribution

iv) the posterior distribution is asymptotically normal.

To make those statements slightly more precise, we introduce the following notation. We assume that we have a parametric model with density function $f(y;\theta)$, where θ takes values in a subset of $\mathbb{R}^k$ and $y = (y_1, \ldots, y_n)$ is an observed vector of observations, each component taking values in the same space, typically a subset of $\mathbb{R}$ or occasionally $\mathbb{R}^d$. The likelihood function $L(\theta; y) = c(y)f(y;\theta)$ is proportional to the density and $\ell(\theta; y) = \log L(\theta)$ is the log-likelihood. The maximum likelihood estimator $\hat{\theta} = \hat{\theta}(y)$ is the value of θ at which $L(\theta)$ or $\ell(\theta)$ reaches a maximum, the score function is $U(\theta) = \partial\ell(\theta)/\partial\theta$, the observed Fisher information function is $j(\theta) = -\partial^2\ell(\theta)/\partial\theta\partial\theta^T$, and the expected Fisher information is $i(\theta) = E\{j(\theta)\}$. The posterior distribution for θ based on a prior $\pi(\theta)$ is $\pi(\theta|y) \propto L(\theta)\pi(\theta)$. The asymptotic

results summarized above can then be expressed as

$$
\begin{aligned}
(\hat{\theta}-\theta)^T\{j(\hat{\theta})\}(\hat{\theta}-\theta) &\rightarrow \chi^2_k && (2.1)\\
\{U(\theta)\}^T\{j(\hat{\theta})\}^{-1}U(\theta) &\rightarrow \chi^2_k && (2.2)\\
2\{\ell(\hat{\theta})-\ell(\theta)\} &\rightarrow \chi^2_k && (2.3)\\
\int_{a_n}^{b_n}\pi(\theta|y)d\theta &\rightarrow \Phi(b)-\Phi(a)\ , && (2.4)
\end{aligned}
$$

where in (2.4) we have assumed $k=1$, $a_n=\hat{\theta}+a\{j(\hat{\theta})\}^{-1/2}$, $b_n=\hat{\theta}+b\{j(\hat{\theta})\}^{-1/2}$, and $\Phi(\cdot)$ is the cumulative distribution function for a standard normal random variable. The convergence is as $n\rightarrow\infty$ and is in distribution in (2.1)–(2.3) and in probability in (2.4). The vector version of (2.4) is given in Walker (1969, p.87).

Conditions are needed on the model $f(y;\theta)$ to ensure that these results are true, and some set of conditions is given in most textbook treatments, such as Lehmann and Casella (1998, Ch. 6). For example, if the components of y are independent and identically distributed, then assuming that $j(\hat{\theta})$ is positive definite, and the third derivative of $\ell(\theta)$ is bounded by an integrable function will be sufficient to apply a Taylor series expansion to the score equation $U(\theta)=0$. As long as the solution to this equation does indeed identify the maximum point of the likelihood function then a central limit theorem applied to $U(\theta)$ will lead to convergence in distribution results for both $\hat{\theta}$ and the likelihood ratio statistic.

When the n components of y are independent and identically distributed, the central limit theorem for the score statistic is usually easily established under relatively weak conditions on the model. A central limit theorem applied to the score statistic is the usual starting point for generalizations to more complex data structure. If the components are not identically distributed then the relevant asymptotic limit is one in which $i(\theta)\rightarrow\infty$, where

$i(\theta)$ is the expected Fisher information in $y = (y_1, \ldots, y_n)$. In the proportional hazards model, asymptotic normality of the partial likelihood estimator follows from a martingale central limit theorem for the partial likelihood score function (Cox, 1972). In a regression model $i(\theta)$ will depend on $X^T X$, and the condition that $i(\theta) \to \infty$ essentially means the explanatory variables do not concentrate on a finite number of distinct values as $n \to \infty$. Some types of dependence among the components of y can also be accommodated, as for example in the $AR(1)$ model with $\rho < 1$; again the crucial assumption is that a central limit theorem holds for the score vector. That conventional asymptotic theory can be quite misleading in the case of long range dependence is amply illustrated in Beran (1994).

There are many models where results (2.1) to (2.4) will not hold, and studying these often provides further insight. If the maximum likelihood estimator is not a root of the score equation then establishing (2.1) will typically require a specially tailored analysis.

One class of nonregular models, those with endpoint parameters, exhibits particularly interesting asymptotics. Suppose $\theta = (\psi, \lambda, \phi)$ and $f(y;\theta) = (y-\psi)^{\lambda-1} g(y-\psi, \phi)$, $y > \psi$ for a smooth function g. Smith (1985, 1989) shows that the behaviour of the maximum likelihood estimator of θ depends crucially on λ; very different asymptotic behaviour resulting as $\lambda > 2$, $\lambda = 2$, $1 < \lambda < 2$, and $0 < \lambda < 1$.

Another class of non-regular problems are Neyman-Scott problems, in which cross-classified or stratified observations have a common parameter of interest:

$$f(y_{ij}; \psi, \lambda_j) \ ; \ j = 1, \ldots, J \ ; \ i = 1, \ldots, I \ .$$

Neyman and Scott (1948) showed by example that as $J \to \infty$ for fixed I the maximum likelihood estimator of ψ may be consistent but not efficient, or

may be inconsistent. Recent work in two-index asymptotics, allowing $I \to \infty$ at a rate related to J, investigates this phenomena in more detail (Portnoy, 1984, 1985; Barndorff-Nielsen, 1996; Sartori, 2001).

The scaling in (2.1) and (2.2) is by observed Fisher information, evaluated at $\hat{\theta}$. The asymptotic statement would be true, under the same regularity conditions, if $j(\hat{\theta})$ were replaced by expected Fisher information $i(\theta)$ or by $i(\hat{\theta})$. Asymptotic theory has established the superiority of $j(\hat{\theta})$ on several grounds. Efron and Hinkley (1978) establish that $j(\hat{\theta})$ more nearly approximates the variance of $\hat{\theta}$ conditional on an ancillary statistic; this result turns out to be closely related to the p^* approximation discussed below. Numerical work has indicated that approximations using $j(\hat{\theta})$ are typically more accurate than those based on $i(\hat{\theta})$ or $i(\theta)$; this again is predicted by the p^* approximation. In models that incorporate strong dependence among the components of y such as often arise in stochastic processes; it is typically the case that $j(\hat{\theta})/i(\hat{\theta})$ does not converge in probability to 1 and scaling by $j(\hat{\theta})$ is required in order that (2.1) and (2.2) hold (Barndorff-Nielsen and Cox, 1994, Ch. 9).

Although (2.1) and (2.2) also holds if $j(\hat{\theta})$ were replaced by $j(\theta)$, this is not a natural substitution to make. In the language of the differential geometry of statistical models both $i(\theta)$ and $j(\hat{\theta})$ are Riemannian metrics for the statistical manifold defined by the model $f(y;\theta)$ and thus in a geometric sense provide the 'right' scaling; $j(\theta)$ does not (Barndorff-Nielsen, Cox and Reid, 1986).

Another interesting way in which the simplest first order asymptotic theory can fail is in models for which the score function is identically zero: in this case an asymptotic theory needs to be developed for the second derivative, and this can have quite different behaviour. The general case is discussed in

Rotnitzky et al. (2000).

Analogues to results (2.1) and (2.2) are often used in more general contexts. The theory of estimating functions replaces the score function by a general function $g(\theta; Y)$ required to have mean zero under the model. The estimator defined by equating $g(\theta; Y)$ to zero will typically converge to θ, and have asymptotic variance larger than that of $\hat{\theta}$. The advantage is that $g(\theta; Y)$ may be defined to have certain robustness properties, or to capture first order properties of an incompletely specified model. In other cases the estimating function may be a score function obtained from 'likelihood-like' function, such as the partial likelihood function from the proportional hazards model discussed above, or one of many so-called *adjusted profile likelihoods*, $\ell(\hat{\theta}_\psi) + B(\psi)$, where $\hat{\theta}_\psi$ is the maximum likelihood estimate of θ under the restriction that $\psi = \psi(\theta)$ is fixed, and $B(\psi)$ is an adjustment motivated usually by higher order asymptotic arguments.

Murphy and Van der Waart (2000) establish a first order asymptotic theory for general classes of semiparametric models. Their arguments rely on the construction of a least-favorable family, and the application of the usual asymptotic results to that parametric family.

Result (2.3), that the log-likelihood ratio statistic is asymptotically chi-squared, follows (again in regular models) by a simple Taylor series expansion establishing that to first order

$$\begin{aligned} 2\{\ell(\hat{\theta}) - \ell(\theta)\} &\simeq U^T(\theta)\{j(\hat{\theta})\}^{-1}U(\theta) \\ &\simeq (\hat{\theta} - \theta)^T\{j(\hat{\theta})\}(\hat{\theta} - \theta) \end{aligned}$$

so of course will not be in any of the cases discussed above where a central limit theorem is not available for $U(\theta)$, or where the maximum likelihood estimator is not a root of the score equation. Result (2.4) relies on similar regularity conditions on the models, and on the assumption that the prior

is of $O(1)$ in an asymptotic theory. Walker (1969) establishes the result for bounded (proper) priors, and Fraser and McDunnough (1984) extends it to improper priors. Freedman (1999) shows that the result cannot be true for nonparametric settings, and Wasserman (2000) provides insight into why this is the case.

In models for which (2.1) to (2.4) are valid, the following results are immediately available. First, the three likelihood-based test statistics are asymptotically equivalent, so any choice among them must be made on grounds other than their limiting distribution. Second, the influence of the prior vanishes in the limit, as is expected by requiring the prior to supply an amount of information equivalent to that contained in one observation. Using the normal approximation suggested by the limiting distributions will lead to the same p-values and confidence bounds in either a frequentist or Bayesian approach.

The classical approach to distinguishing among these test statistics is to investigate their power under alternatives to the model. This has not led very far, because there is typically no uniform domination under alternative models, although it can be shown for example that the score statistic (2.2) is locally most powerful, i.e. most powerful under alternatives of the form $f(y;\theta_0 + \delta/\sqrt{n})$. Amari (1985) shows that the asymptotic power of the likelihood ratio test is largest over a wide range of alternatives. Properties in addition to power, such as invariance or unbiasedness, have come to seem increasingly arbitrary. Of course in much applied work the choice is made on the basis of convenience, and in some quite complex models a version of the score statistic (2.2) using $i(\theta)$ for scaling is easiest to compute, as it doesn't involve computation of the maximum likelihood estimator.

An approach closer to that of the work surveyed here is to consider how good an approximation is provided by the limiting distribution. There are no definitive results available, but in problems with a scalar parameter of interest the signed square root of $w(\theta)$, which preserves information on the direction of departure from the assumed model value θ, has in a large number of numerical studies given the most accurate p-values. Higher order asymptotic criteria also favour the signed square root of $w(\theta)$, as outlined in the next section. It is difficult to make direct comparisons of (2.4) with (2.1) to (2.3) in a non-asymptotic context because of the presence of the arbitrary prior $\pi(\theta)$ in (2.4). However, higher order asymptotic analysis has lead to some direct comparisons of confidence limits, as outlined below in connection with matching priors.

2.2 Higher order asymptotics for likelihood

The main higher-order asymptotic results for likelihood based inference are:

(i) the p^* approximation to the density of the maximum likelihood estimator

(ii) the r^* approximation to the distribution function of the signed square root of the log likelihood ratio statistic

(iii) adjustments to profile likelihood to accommodate nuisance parameters

(iv) Laplace expansions for posterior distributions and matching priors.

We briefly discuss each of these in turn.

Barndorff-Nielsen (1980, 1983) emphasized the central importance of the following approximation to the density of the maximum likelihood estimator:

$$f(\hat{\theta};\theta|a) \doteq p^*(\hat{\theta};\theta|a) = \frac{c}{(2\pi)^{k/2}}|j(\hat{\theta})|^{1/2}\exp\{\ell(\theta)-\ell(\hat{\theta})\} \qquad (2.5)$$

where $c = c(\theta, a)$ is a renormalizing constant, $j(\hat{\theta})$ is the observed Fisher information, and $2\{\ell(\hat{\theta}) - \ell(\theta)\}$ is the log-likelihood ratio statistic. In the right hand side of (2.5), $\ell(\theta) = \ell(\theta; \hat{\theta}, a)$ and $j(\hat{\theta}) = j(\hat{\theta}; \hat{\theta}, a)$, where $(\hat{\theta}, a)$ is a one to one transformation of y, or equivalently the minimal sufficient statistic based on y. Note that the normal approximation obtained from (2.5) by Taylor expansion of the exponent has variance $j(\hat{\theta})$.

In applications of (2.5) a is exactly or approximately ancillary, i.e. distribution constant; if not then its marginal distribution will carry information for θ and (2.5) will not be very useful for inference. The right hand side of (2.5) can be obtained from $L(\theta)/\int L(\theta)d\theta$ by a Laplace approximation to the integral in the denominator, and thus generalizes Fisher's (1934) exact result in location models. Approximation (2.5) can also be derived in canonical exponential families from a saddlepoint approximation to the distribution of the minimal sufficient statistic.

The ancillary statistic a is best regarded as providing a dimension reduction (by conditioning) from the full dimension, n, of the data, to the dimension k of the parameter. What the approximation suggests is that conditioning on an ancillary statistic is an essential component of frequentist inference.

Approximation (2.5) is usually referred to as the p^* approximation, or Barndorff-Nielsen's p^* approximation. The connection to location models and exponential families was discussed in Reid (1988). Skovgaard (1990) outlined the main elements of the proof that (2.5) provides an approximation to the exact conditional density with relative error $O(n^{-3/2})$, although a completely general and rigorous proof is still lacking. The difficulty is that it is not clear in a very general setting what the transformation from y to $(\hat{\theta}, a)$ is. The easiest route is to embed the model in a full exponential family

in which θ is a restriction of the parameter to a curve. If the full model has a parameter with fixed dimension (free of n) then an approximately ancillary statistic is readily constructed. This is the approach outlined in Barndorff-Nielsen and Cox (1994, Ch.7). Skovgaard (2001) refers to (2.5) as a Laplace-type approximation, and notes the similarity to Wiener germs, defined by Dinges (1986).

In the special case that θ is a scalar parameter, (2.5) can be re-expressed as an approximation to the density of the likelihood root $r(\theta)$, defined by

$$\ell(\theta) - \ell(\hat{\theta}) = -\frac{1}{2}r^2(\theta) \tag{2.6}$$

and this is the basis for the approximation of the cumulative distribution function of r by

$$\begin{aligned} F(r;\theta|a) &\doteq \Phi(r^*) && (2.7)\\ &\doteq \Phi(r) + \left(\frac{1}{r} - \frac{1}{q}\right)\phi(r) && (2.8) \end{aligned}$$

where

$$\begin{aligned} r^* &= r + \frac{1}{r}\log\frac{q}{r} && (2.9)\\ q &= \{\ell_{;\hat{\theta}}(\theta) - \ell_{;\hat{\theta}}(\hat{\theta})\}\{j(\hat{\theta})\}^{-1/2} && (2.10)\\ \ell_{;\hat{\theta}}(\theta;\hat{\theta},a) &= \partial\ell(\theta;\hat{\theta},a)/\partial\hat{\theta}\ . && (2.11) \end{aligned}$$

This was derived in Barndorff-Nielsen (1986, 1990, 1991) by integrating the p^* approximation. A more direct version that avoids the transformation from y to $(\hat{\theta}, a)$ was outlined in Fraser (1990, 1991).

The insights provided by (2.7) and (2.8) are several, and arguably of more importance for inference then the p^* approximation, since in most cases we need the distribution function in order to compute p-values and confidence limits. In going from (2.5) to (2.7) or (2.8), the transformation to r is key,

which suggests that the likelihood root, r, is the intrinsic statistic with which to measure the departure of $\hat{\theta}$ from θ. Approximation (2.7), with (2.9), shows that r is nearly normally distributed, and a small adjustment makes it much closer to normally distributed. The 'smallness' of the adjustment can be quantified by Taylor series expansion of (2.10):

$$r = q + \frac{A}{\sqrt{n}}q^2 + \frac{B}{n}q^3 + O(n^{-3/2}) \tag{2.12}$$

where A and B are free of n. The appearance of $\partial\ell/\partial\hat{\theta}$ in q is an inevitable consequence of the change of variable from $\hat{\theta}$ to r: in other words it is essential to consider how the likelihood function changes with small changes to the data. These changes are in only certain directions, namely those in which the ancillary statistic a is held fixed.

Note that if we were interpreting the p^* approximation as an approximate posterior density with a flat prior, then the corresponding posterior distribution function would involve the change of variable $\partial r/\partial\theta$, instead of $\partial r/\partial\hat{\theta}$, and as a result the corresponding q would be a standardized score statistic.

The dependence of p^* on an exact or approximate ancillary statistic a has somewhat hampered its usefulness, as in general models it is not usually clear how to construct such a statistic, and an explicit expression is rarely available. However, in (2.7) and (2.8) the only dependence on a is on the first derivative of ℓ; Fraser (1988, 1990) exploited this fact to derive an alternate expression for q:

$$q = \{\ell_{;V}(\theta; y) - \ell_{;V}(\hat{\theta}; y)\}\{j(\hat{\theta})\}^{1/2}|\ell_{\theta;V}(\hat{\theta})|^{-1} \tag{2.13}$$

where $V = V(y)$ is an $n \times 1$ vector tangent to the ancillary statistic, and $\ell_{;V}(\theta)$ is the directional derivative $\frac{d}{dt}\ell(\theta; y^0 + tV)$ computed at a fixed data value y^0. A fuller description of q and V is given in Section 3.

To describe results in models with vector parameters some additional notation is needed. We assume that the parameter $\theta = (\psi, \lambda)$ is partitioned into a vector of parameters of interest, ψ and nuisance parameters λ. The more general case where the parameter of interest is expressed as a restriction on θ by $\psi = \psi(\theta)$, and the nuisance parameterization is not given explicitly, is discussed briefly in Section 3. We denote the restricted maximum likelihood estimator of λ by $\hat{\lambda}_\psi$, and write $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$. We assume $\hat{\lambda}_\psi$ is defined by $\partial \ell(\psi, \hat{\lambda}_\psi)/\partial \lambda = 0$. The observed information function $j(\theta)$ is partitioned as

$$j(\theta) = \begin{bmatrix} j_{\psi\psi}(\theta) & j_{\psi\lambda}(\theta) \\ j_{\lambda\psi}(\theta) & j_{\lambda\lambda}(\theta) \end{bmatrix}$$

with a similar partitioning of its inverse

$$j^{-1}(\theta) = \begin{bmatrix} j^{\psi\psi}(\theta) & j^{\psi\lambda}(\theta) \\ j^{\lambda\psi}(\theta) & j^{\lambda\lambda}(\theta) \end{bmatrix} .$$

The profile log-likelihood function is $\ell_p(\psi) = \ell(\psi, \hat{\lambda}_\psi)$, and the profile observed information is $j_p(\psi) = -\partial^2 \ell_p(\psi)/\partial\psi\partial\psi^T$. The result

$$|j_p(\psi)| = |j(\psi, \hat{\lambda}_\psi)|/|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)| \tag{2.14}$$

follows from the definition of $\hat{\lambda}_\psi$ and the formula for the determinant of a partitioned information matrix.

In problems with nuisance parameters, a central role is played by the adjusted profile log likelihood function

$$\ell_a(\psi) = \ell_p(\psi) - \frac{1}{2} \log |j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)| . \tag{2.15}$$

The easiest way to see why is to consider Laplace approximation of the posterior marginal density for ψ:

$$\begin{aligned}
\pi_m(\psi|y) &= \int \pi(\psi,\lambda|y)d\lambda \\
&= \frac{\int \exp\{\ell(\psi,\lambda)\}\pi(\psi,\lambda)d\lambda}{\iint \exp\{\ell(\psi,\lambda)\}\pi(\psi,\lambda)d\psi d\lambda} \\
&\doteq \frac{\exp\{\ell(\psi,\hat\lambda_\psi)\}|j_{\lambda\lambda}(\psi,\hat\lambda_\psi)|^{-1/2}\pi(\psi,\hat\lambda_\psi)}{\exp\{\ell(\hat\psi,\hat\lambda)\}|j(\hat\psi,\hat\lambda)|^{-1/2}\pi(\hat\psi,\hat\lambda)} \cdot \frac{1}{\sqrt{(2\pi)}} \\
&\doteq \frac{1}{\sqrt{(2\pi)}} \exp\{\ell_a(\psi)-\ell_a(\hat\psi)\}|j_a(\hat\psi)|^{1/2}\frac{\pi(\hat\theta_\psi)}{\pi(\hat\theta)} \qquad (2.16)
\end{aligned}$$

where the first approximation results from the leading term of a Laplace expansion of the integrals, and the second by using (2.14) and the results that $|j_a(\hat\psi)| = |j_p(\hat\psi)|\{1+O_p(n^{-1})\}$, and $\hat\psi_a - \hat\psi = O_p(n^{-1})$, where $\hat\psi_a$ is the solution of $\ell_a'(\psi)=0$.

The structure of (2.16) is very similar to that of the p^* approximation (2.5), with an additional factor due to the prior, and with the log-likelihood function replaced by the adjusted log-likelihood function. As a result, when ψ is a scalar this approximation can be integrated to give an r^*–type approximation to the marginal cumulative distribution function.

The approximation of the marginal posterior density can also be expressed in terms of the profile log-likelihood function. To the same order of approximation as in (2.16), i.e. $O(n^{-3/2})$, we have

$$\pi_m(\psi|y) \doteq \frac{1}{\sqrt{(2\pi)}} \exp\{\ell_p(\psi)-\ell_p(\hat\psi)\}|j_p(\hat\psi)|^{1/2}\left\{\frac{|j_{\lambda\lambda}(\hat\theta)|}{|j_{\lambda\lambda}(\hat\theta_\psi)|}\right\}^{1/2}\frac{\pi(\hat\theta_\psi)}{\pi(\hat\theta)} \qquad (2.17)$$

which leads to a tail area approximation involving the signed root of the profile log likelihood, as outlined in Section 3.

In a frequentist approach to inference it is more difficult to find a general prescription for eliminating the nuisance parameter. But in two classes

of models, exponential families and transformation families, elimination of nuisance parameters can be achieved by conditioning (in exponential families) or marginalizing (in transformation families), as long as the parameter of interest ψ is a component of the canonical parameter for the model. In these cases r^*–type approximations are derived from p^*–type approximations to the appropriate conditional or marginal density. The frequentist versions involve sample space derivatives, as in the one-dimensional case, and are reviewed in Reid (1996).

What is striking about these approximations from a theoretical point of view is that to this order of approximation nuisance parameters are accommodated by a relatively simple adjustment based on the nuisance parameter information, and except for this adjustment the calculations are essentially the same as in the case of no nuisance parameters. A p^*–type density approximation is needed as a starting point for computing significance probabilities, which suggests in the general case that conditioning on an approximate ancillary is needed to reduce the dimension from n to k. As will be discussed in more detail in Section 3, the dimension reduction from that of θ to that of ψ is achieved by marginalizing.

The higher order expansion (2.16) for the posterior marginal density using Laplace approximation is due to Tierney and Kadane (1986). They also showed that the relative error in (2.16) is $O(n^{-3/2})$. A different approximation based on Edgeworth expansions was derived in Johnson (1970) and by Welch and Peers (1963) and Peers (1965). These latter two papers were particularly concerned with answering a question posed by Lindley: does a prior exist for which posterior probability limits have a frequentist intepretation as confidence limits? In location models it follows from Fisher (1934) that $p(\hat{\theta}|a;\theta) = \pi(\theta|y)$ under the flat prior $\pi(\theta) \propto 1$, and this was generalized

to transformation models in Fraser (1968) and Barndorff-Nielsen (1980): the flat prior is replaced by the right Haar measure.

Welch and Peers showed that when θ is a scalar, Jeffreys' prior

$$\pi(\theta) \propto \{i(\theta)\}^{1/2} \tag{2.18}$$

ensures the matching condition

$$\Pr_{Y|\theta}\{\theta \leq \theta^{(1-\alpha)}(\pi, Y)\} = 1 - \alpha + O(n^{-1}) \tag{2.19}$$

where $\theta^{(1-\alpha)}(\pi, y)$ is the $(1-\alpha)$ posterior limit:

$$\Pr_{\theta|Y}\{\theta \leq \theta^{(1-\alpha)}(\pi, y)|y\} = 1 - \alpha \ . \tag{2.20}$$

In the case that θ is a vector parameter there is no simple condition for matching posterior and sampling probabilities (Peers, 1965). If $\theta = (\psi, \lambda)$, and ψ is orthogonal to λ matching priors for inference on ψ are defined by

$$\pi(\theta) \propto \{i_{\psi\psi}(\theta)\}^{1/2} g(\lambda) \tag{2.21}$$

where $g(\lambda)$ is an arbitrary function (Tibshirani, 1989; Nicolau, 1993). The lack of a general solution has led to a large literature on finding matching priors using more restrictive matching conditions; see Ghosh and Mukerjee (1998) for a review. An alternative approach to matching based on r^*-type approximations is discussed briefly in Section 3.

A Bayesian approach can be used to derive frequentist asymptotic expansions by an elegant *shrinkage* argument due to J.K. Ghosh, and to Dawid (1991). An expository account is given in Mukerjee and Reid (2000). The main point is that expansions can be carried out in the parameter space, rather than the sample space, which as we have seen in the derivation of r^* is typically much easier.

2.3 Other applications

There are a number of related areas where asymptotic arguments have been crucial for the advancement of the theory of statistics: in fact it is difficult to find an area in the theory where asymptotic arguments do not play a central role. In this section we mention briefly particular areas that seem most closely related to statistical theory based on likelihood.

The most important of these is the asymptotic analysis of the bootstrap. As a methodological tool, the bootstrap is indispensable, but in terms of understanding when and how it is useful, its asymptotic properties are paramount and in this connection the result that the bootstrap provides second order correct confidence limits was key (Bickel and Freedman, 1981; Singh, 1981). In this sense bootstrap methods are also higher order asymptotic methods, and there are several close points of contact between results of the previous subsection and the parametric bootstrap. Although the connections are not yet completely clear, much progress is made in DiCiccio and Efron (1992) and Davison and Hinkley (1998).

Another development in higher order asymptotics is the development of Edgeworth expansions for U statistics; see Bentkus, Bötze and van Zwet (1997) and Bloznelis and Götze (2000). This enables the development of asymptotic theory for that accommodates certain types of dependence, which is important in several applications. From the point of view of establishing proofs, Edgeworth expansions are key, because the saddlepoint type expansions used in likelihood asymptotics are essentially specialized Edgeworth expansions.

Edgeworth expansions for martingales have been developed by Mykland (1995) and these provide the possibility of a type of higher order asymptotic analysis in settings like the proportional hazards model. A key component

of this is the verification that the likelihoods derived from martingales obey the Bartlett identities (Mykland, 2000).

As mentioned in Section 2.1, a different approach to higher order inference from that discussed here is to investigate the power properties of the standard test statistics. Edgeworth expansions for this are given in Amari (1985) and Pfanzagl (1985).

Empirical likelihood (Owen, 1988, 2001) has many properties of parametric likelihood, and second order properties have been investigated by, for example, DiCiccio, Hall and Romano (1991).

3 On p-values for a scalar parameter

3.1 Introduction

In this section we develop the approximation discussed briefly in Section 2 in more detail. In particular, we describe the construction of q used to compute p-values when testing a scalar parameter of interest, $\psi = \psi(\theta)$ in the presence of nuisance parameters. When computed as a function of ψ, $p(\psi)$, say, this provides approximate confidence bounds at any desired level of confidence, so the testing problem is not essentially different from the problem of interval estimation. This function $p(\psi)$ is called the significance function in Fraser (1991) and the confidence distribution function in Efron (1997). It provides a direct comparison to Bayesian inference, as the significance function is analogous to the marginal posterior cumulative distribution function.

The approximation to the p-value function is the same as that given by

(2.7) and (2.8):

$$p(\psi) \doteq \Phi(r) + \phi(r)\left(\frac{1}{r} - \frac{1}{q}\right) \tag{3.1}$$

$$\doteq \Phi\left(r + \frac{1}{r}\log\frac{1}{q}\right) \equiv \Phi(r^*) \tag{3.2}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal distribution function and density function. Here $r = r(\psi)$ is the log-likelihood root based on the profile log likelihood

$$\begin{aligned} r(\psi) &= \text{sign}(\hat{\psi} - \psi)[2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\}]^{1/2} \\ &= \text{sign}(\hat{\psi} - \psi)[2\{\ell_p(\hat{\psi}, \hat{\lambda}) - \ell(\psi, \hat{\lambda}_\psi)\}]^{1/2}\ , \end{aligned} \tag{3.3}$$

and $q = q(\psi)$ is a type of maximum likelihood or score statistic to be described below.

Other asymptotically equivalent approximations to (3.1) and (3.2) use as the starting point the likelihood root obtained from an adjusted profile log-likelihood. Numerical work in relatively simple models seems to indicate that approximations based on adjusted likelihood roots are better (Butler, Huzurbazar and Booth, 1992b; Pierce and Peters, 1992; Sartori, 2001) so the use of $\ell_p(\psi)$ in (3.3) may be a drawback of the approach developed here. An advantage though of this approach is the derivation of an expression for q that can be used in general models. Versions of the p-value approximation that use adjusted likelihood roots are currently available only for inference on a canonical parameter in full exponential models or a linear parameter in transformation models.

Approximations of the form (3.1) and (3.2), but accurate to $O(n^{-1})$ instead of $O(n^{-3/2})$ can be constructed without specification of an approximate ancillary statistic. Several such approximations have been suggested in the literature; see for example Barndorff-Nielsen and Chamberlin (1991, 1994)

and DiCiccio and Martin (1993). The version due due to Skovgaard (1996) seems most useful for both theoretical and applied work. Severini (1999) gives an alternate expression for Skovgaard's statistic and an alternate derivation of his result.

The steps in the derivation of (3.1) or (3.2) are the following:

1) Find a density or approximate density supported on $\mathbb{R}^k$ rather than $\mathbb{R}^n$, for example by using a p^* approximation for the conditional density given an approximate ancillary statistic.

2) Transform this density to that of a one-dimensional pivotal statistic for ψ, and a complementary statistic of dimension $k-1$, and find the marginal density of the pivotal.

3) Integrate this marginal density up to (or beyond) the observed data point.

Note that steps 1 and 3 are also needed for problems without nuisance parameters, so follow the pattern outlined in Section 2. As we will see, step 2 can be obtained using the same argument as in step 1. We now discuss steps 1 to 3 in more detail.

3.2 Reduction from n to k.

This reduction can be achieved using the p^* approximation to the density of the maximum likelihood estimator by conditioning on an approximate ancillary statistic. That is, we find a statistic $a = a(y)$ of dimension $n-k$ so that the transformation from y to $(\hat{\theta}, a)$ is one to one. In some models it may be convenient to first replace y by the minimal sufficient statistic for θ, if such exists, but the end result is the same.

We transform the joint density for y into that for $(\hat{\theta}, a)$, and express the result as

$$p(\hat{\theta}|a;\theta)p(a) \ , \tag{3.4}$$

where the requirement that a be ancillary ensures that all the information about θ is contained in the conditional density. As described in Section 2, $p(\hat{\theta}|a;\theta)$ can be approximated with relative error $O(n^{-3/2})$ by the p^* approximation (2.5) and to this order of approximation it is sufficient that a be ancillary only to $O(n^{-1})$, i.e. that $p(a;\theta_0 + \delta/\sqrt{n}) = p(a;\theta_0)\{1 + O(n^{-1})\}$.

The derivation of (3.4) is outlined most clearly in Skovgaard (1990), and from there the derivation of the p^* approximation proceeds by finding a suitable approximate ancillary statistic. This last step can be done explicitly if $f(y;\theta)$ can be embedded in a model with a $k+d-$dimensional minimal sufficient statistic and a $k+d-$dimensional parameter; i.e., in a $k+d$ full exponential family, thus making $f(y;\theta)$ a $(k+d, k)$ curved exponential family. Skovgaard's (1990) derivation assumes this to be the case, as do most papers by Barndorff-Nielsen and colleagues, see, for example, Barndorff-Nielsen and Wood (1998). In this case the approximate ancillary statistic can be constructed by a sequence of r^* statistics, defined as in (3.2), but treating each component of the nuisance parameter successively as a parameter of interest. An explicit construction along these lines is given in Barndorff-Nielsen and Wood (1998); see also Jensen (1992) and Skovgaard (1996).

However, since the end goal is to integrate the density approximation to obtain p-values, a simpler approach is possible that does not involve the explicit specification of $\hat{\theta}$ and a. As in the case with no nuisance parameters, the full dependence of the log-likelihood function on $(\hat{\theta}, a)$ is not needed. All that is needed is information on how the log-likelihood changes with changes in $\hat{\theta}$ that keep a fixed, and even this information is needed only to

first derivative at the observed data point: to first derivative because the integration of the density approximation involves a Jacobian for a change of variable, and at the observed data point because we are approximating a p-value.

This leads to a simplification of the p^* approximation, called in Fraser and Reid (1993, 1995) a tangent exponential model, and given by

$$p_{\mathrm{TEM}}(s|a;\theta) = c|j(\hat{\varphi})|^{-1/2} \cdot \exp[\ell(\theta;y^0) - \ell(\hat{\theta}^0;y^0) + \{\varphi(\theta) - \varphi(\hat{\theta}^0)\}^T s] \quad (3.5)$$

where y^0 is the observed data point, $\hat{\theta}^0 = \hat{\theta}(y^0)$ is the observed maximum likelihood estimate of θ,

$$\varphi(\theta) = \varphi(\theta;y^0) = \frac{\partial \ell(\theta;y)}{\partial V(y)}\Big|_{y=y^0} , \quad (3.6)$$

is a local reparametrization of the model, $V(y)$ is an $n \times k$ matrix whose columns are vectors tangent to the approximate ancillary statistic a, and

$$j(\varphi) = -\frac{\partial^2 \ell\{\varphi(\theta);y^0\}}{\partial\varphi\partial\varphi^T} \quad (3.7)$$

is the observed information in the new parameterization φ. In the left hand side of (3.5), (s,a) is a one-to-one transformation of y that is assumed to exist, but the precise form of the transformation is not needed. The variable s plays the role of a score variable, replacing $\hat{\theta}$ in the p^* approximation, and a is the approximate ancillary statistic. The existence of a is established in Fraser and Reid (1995) by using an local location model approximation to $f(y;\theta)$.

The right hand side of (3.5) has the structure of a full exponential model, with $\varphi = \varphi(\theta)$ the canonical parameter, $s = s(y)$ the minimal sufficient statistic, and $\ell(\theta) = \ell\{\varphi(\theta)\}$ playing the role of the cumulant generating function. In its dependence the log-likelihood function, we need only $\ell(\theta;y^0)$

and $\ell_{;V}(\theta; y^0)$, not $\ell(\theta; y)$ as is needed for p^*. The information determinant $|j(\hat{\varphi})|^{-1/2}$ is the volume element for the score variable s. The approximate ancillary a appears only through the matrix V used to define the canonical parameters. The k columns of V are vectors in $\mathbb{R}^n$ that are tangent to the subspace of $\mathbb{R}^n$ defined by holding a fixed. As in the scalar parameter case, these tangent vectors will determine the Jacobian for the change of variable to the likelihood root r which is needed to compute the p-value approximation.

It is shown in Fraser and Reid (1995) that the vectors V can be constructed using a vector of pivotal statistics $z = \{z_1(y_1, \theta), \ldots, z_n(y_n, \theta)\}$, where each component $z_i(y_i, \theta)$ has a fixed distribution under the model. (This assumes that the components of y are independent.) Such a vector always exists in the form of the probability integral transformation $F(y_i; \theta)$, although simpler alternatives may be available. The vectors $v_1, \ldots, v_k$ are defined by

$$V = -\left\{ \left(\frac{\partial z}{\partial y} \right)^{-1} \left(\frac{\partial z}{\partial \theta} \right) \right\} \Big|_{(y^0, \hat{\theta}^0)} \tag{3.8}$$

and it can be shown that these are tangent to the surface in the sample space on which a second order ancillary statistic is held constant.

Example: Tilted logistic

We will illustrate the detailed calculations for a sample of size 2 from the model

$$f(y; \theta) = \frac{e^{(y-\theta)}}{\{1 + e^{(y-\theta)}\}^2} e^{\gamma(\theta)(y-\theta) - c\{\gamma(\theta)\}} , \quad -1 \leq \theta \leq 1 \tag{3.9}$$

where $\gamma(\theta) = 0.5 \tanh(\theta)$ and $c(\theta) = \log\{(\pi\theta)/\sin(\pi\theta)\}$. The joint density of (y_1, y_2) for fixed θ is shown in Figure 3.1.

To compute the p^* approximation to the joint density of $\hat{\theta}$ given a, we need to find coordinates $(\hat{\theta}, a)$ as a function of (y_1, y_2) for which $\ell'(\hat{\theta}) = 0$

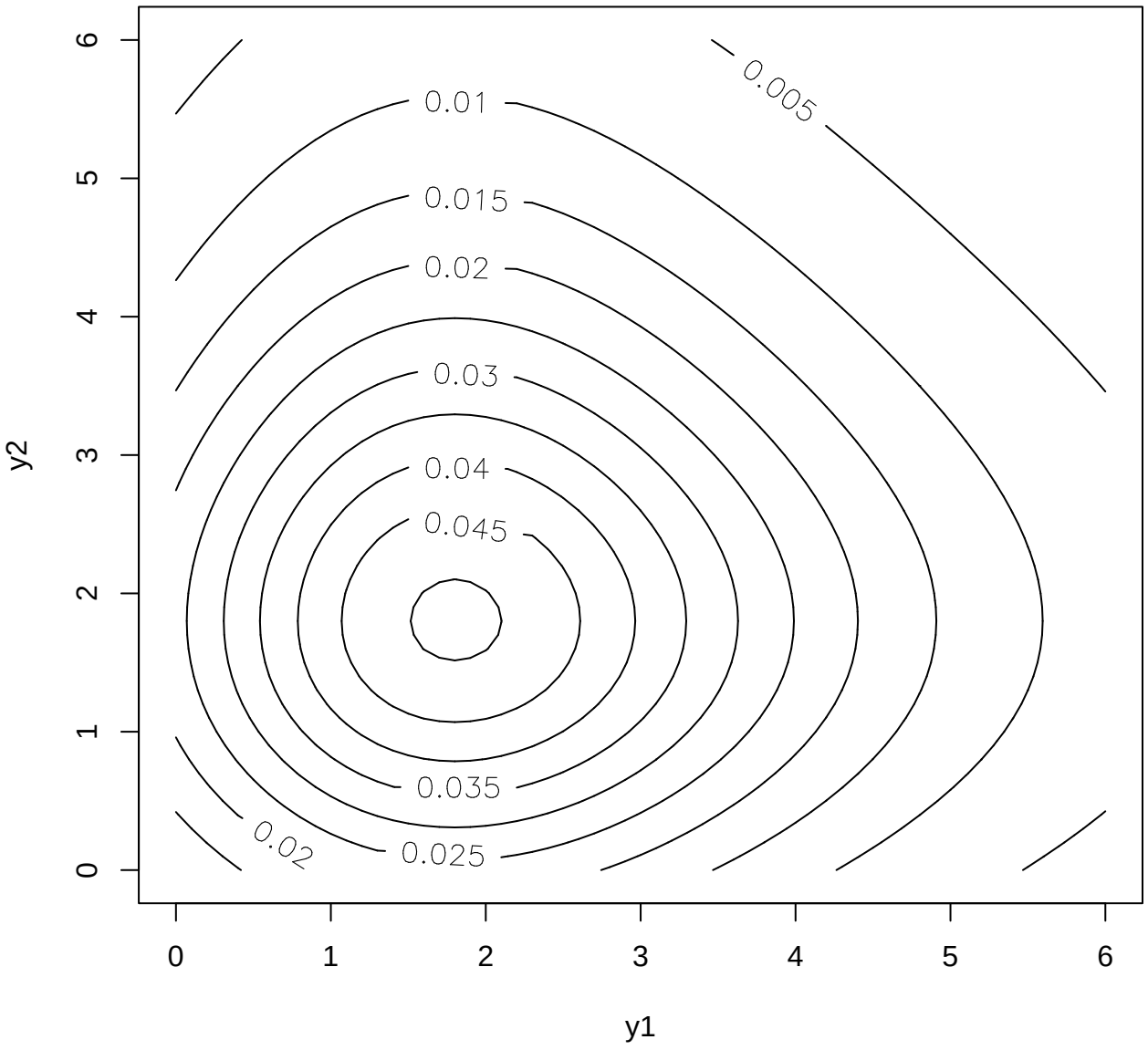

Figure 3.1: The joint density when $\theta = 1$, for the model given in (3.9).

and $f(a;\theta)$ is free of θ exactly or approximately. Fixing a will define a curve on the two-dimensional surface expressed in $\hat{\theta}$ coordinates with its associated measure $|j(\hat{\theta})|^{1/2}$.

To compute the tangent exponential model we start at the data point $y^0 = (y_1^0, y_2^0)$, and trace a curve in $\mathbb{R}^2$ by finding at the ith step

$$\begin{aligned} v_1^{(i)} &= -\frac{F_\theta(y_1,\theta)}{f(y_1,\theta)}\Big|_{y_1^{(i-1)},\hat{\theta}^{(i-1)}} \\ v_2^{(i)} &= -\frac{F_\theta(y_2,\theta)}{f(y_2,\theta)}\Big|_{y_2^{(i-1)},\hat{\theta}^{(i-1)}} \end{aligned} \tag{3.10}$$

where we are using the probability integral transformation to define the pivotals (z_1, z_2). The curve is then

$$y^0, y^{(1)} = y^0 + \delta v^{(1)}, \ldots, y^{(i)} = y^{(i-1)} + \delta v^{(i)}, \ldots$$

illustrated in Figure 3.2.

The tangent exponential model (3.5) gives the density function at each point along this curve, with relative error $O(n^{-3/2})$, in terms of a score variable $s = s(y)$ with associated measure $|j(\hat{\varphi})|^{-1/2}$.

3.3 Reduction from k to 1.

We now work within the model conditional on an approximate ancillary statistic; using either $p^*(\hat{\theta}|a)$ or $p_{\mathrm{TEM}}(s|a)$. The dimension reduction from k to 1 for inference about the scalar parameter ψ is achieved by finding a one dimensional pivotal statistic, i.e. a function of $\hat{\theta}$ and ψ or s and ψ that has a distribution free of the nuisance parameter λ: in other words a statistic that is ancillary for λ when ψ is fixed. Thus an application of Step 1 to p^* or p_{TEM} with ψ fixed will provide an ancillary statistic. This statistic, a_ψ, say,

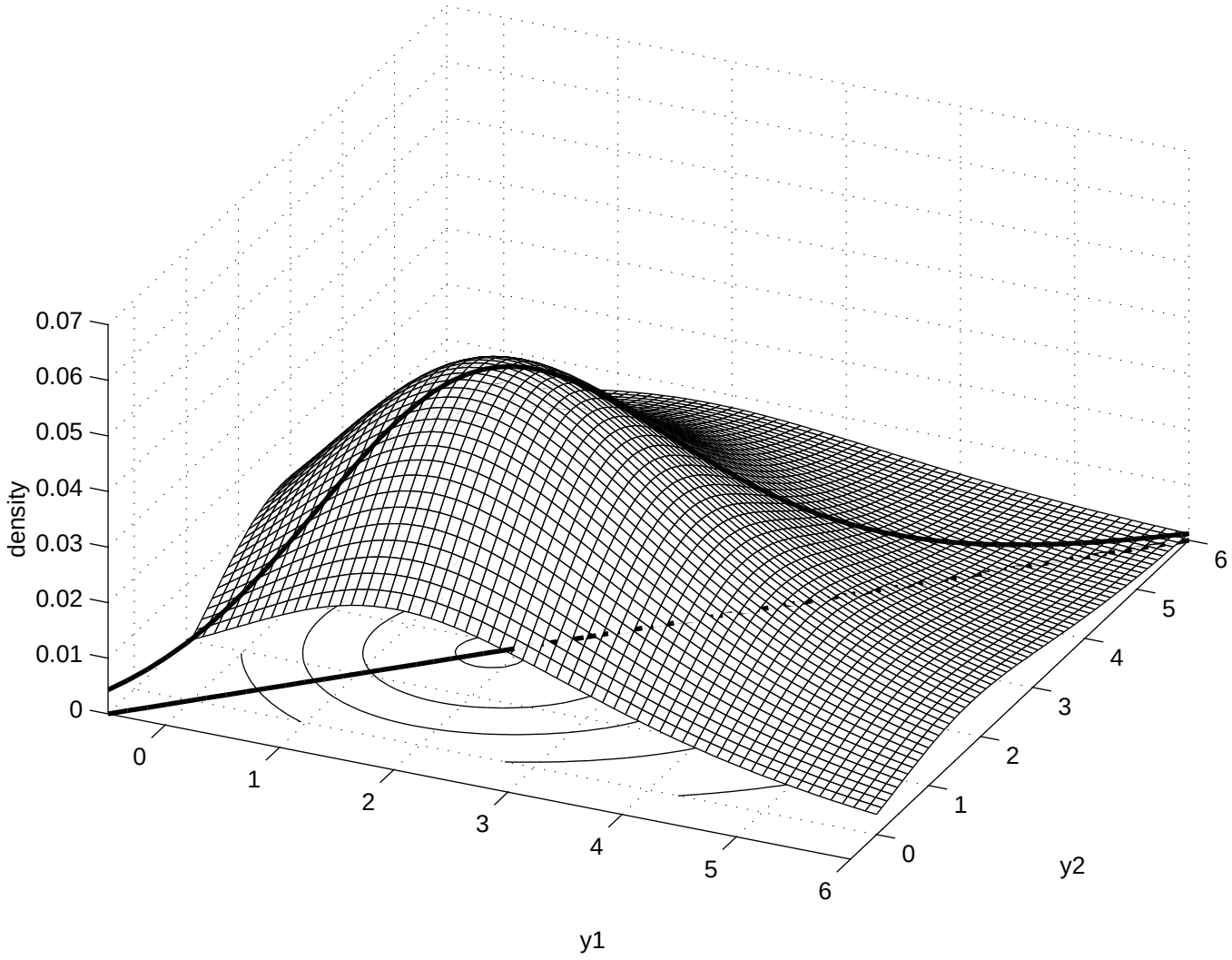

Figure 3.2: The curve along which the second order ancillary, with tangent vectors given by V, is constant.

will be only approximately ancillary, but as long as it is ancillary to $O(n^{-1})$ or better it will suffice to produce tail area approximations to $O(n^{-3/2})$.

In Barndorff-Nielsen's approach, starting with $p^*(\hat{\theta}|a)$, the pivotal is taken to be

$$r_\psi^* = r_\psi + \frac{1}{r_\psi} \log \frac{u_\psi}{r_\psi}. \tag{3.11}$$

The joint density of $\hat{\theta}$ is transformed to the joint density of $(r_\psi^*, \hat{\lambda}_\psi)$. An application of the p^* approximation (Step 1) to the conditional density of $\hat{\lambda}_\psi$ given r_ψ^* then gives the marginal density of r_ψ^* as $N(0,1)$ to $O(n^{-3/2})$. This marginal density is used to compute the p-value. This argument was first presented in Barndorff-Nielsen (1986), where a rather complicated expression for u_ψ is given. Subsequent work simplified the expression for u_ψ to that given in (3.15) below.

Nearly the same argument is used in Fraser and Reid (1995), starting from the tangent exponential model (3.5). As in Step 1, but now in the model for s given a with ψ fixed, there exists an ancillary statistic a_ψ for λ, and a score variable s_ψ with an approximating tangent exponential model; i.e.

$$p_{\mathrm{TEM}}(s|a;\theta) = p_1(s_\psi;\theta|a_\psi)p_2(a_\psi)$$

where p_1 also has exponential family form. The ratio of p_{TEM} to p_1 gives the marginal density of a_ψ as

$$p_2(a_\psi) = \frac{p_{\mathrm{TEM}}(s|a;\psi,\lambda)}{p_1(s_\psi|a_\psi;\psi,\lambda)} \ .$$

Since we know the result is free of both λ and s_ψ, we can evaluate the right hand side at convenient values of λ and s_ψ, which we take to be $\hat{\lambda}_\psi$ and 0. The resulting approximation has the form of a tangent exponential model with an adjustment factor involving the information matrix for the nuisance parameter.

3.4 Approximate tail areas.

In either approach the resulting marginal density is readily integrated to compute tail area approximations. In Barndorff-Nielsen's approach this is a trivial by product of the result in Step 2 that r^*_ψ has a standard normal distribution. In the Fraser-Reid approach the marginal density of the pivotal statistic lends itself to a Lugannani and Rice (1980) type approximation, as described in Cheah, Fraser, Reid (1995) and illustrated below at (3.xx). Skovgaard (2001) refers to these approximations as being a "Laplace type".

The result is that the p-value is approximated by

$$\begin{aligned} \Phi(r^*_\psi) &= \Phi\left(r_\psi + \frac{1}{r_\psi}\log\frac{u_\psi}{r_\psi}\right) \\ &= \Phi(r_\psi) + \phi(r_\psi)\left(\frac{1}{r_\psi} - \frac{1}{u_\psi}\right) \end{aligned} \tag{3.12}$$

or

$$\begin{aligned} \Phi(r^*_\psi) &= \Phi\left(r_\psi + \frac{1}{r_\psi}\log\frac{q_\psi}{r_\psi}\right) \\ &= \Phi(r_\psi) + \phi(r_\psi)\left(\frac{1}{r_\psi} - \frac{1}{q_\psi}\right) \end{aligned} \tag{3.13}$$

where

$$r_\psi = \text{sign}(\hat\psi - \psi)[2\{\ell(\hat\theta) - \ell(\hat\theta_\psi)\}]^{1/2} \tag{3.14}$$

$$u_\psi = \frac{|\ell_{;\hat\theta}(\hat\theta) - \ell_{;\hat\theta}(\hat\theta_\psi) \qquad \ell_{\lambda;\hat\theta}(\hat\theta_\psi)|}{|\ell_{\theta;\hat\theta}(\hat\theta)|} \cdot \frac{|j_{\theta\theta}(\hat\theta)|^{1/2}}{|j_{\lambda\lambda}(\hat\theta_\psi)|^{1/2}} \tag{3.15}$$

$$q_\psi = \frac{|\ell_{;V}(\hat\theta) - \ell_{;V}(\hat\theta_\psi) \qquad \ell_{\lambda;V}(\hat\theta_\psi)|}{|\ell_{\theta;V}(\hat\theta)|} \cdot \frac{|j_{\theta\theta}(\hat\theta)|^{1/2}}{|j_{\lambda\lambda}(\hat\theta_\psi)|^{1/2}} \tag{3.16}$$

$$= \{\nu(\hat\theta) - \nu(\hat\theta_\psi)\}/\hat\sigma_\nu\ . \tag{3.17}$$

Another way of describing u_ψ is as a dual score statistic (Barndorff-Nielsen and Cox, 1994 §6.6). In contrast q_ψ is a type of dual maximum likelihood

statistic, as indicated at (3.17), for a derived scalar parameter ν. In (3.17) we have

$$\begin{aligned}
\nu(\theta) &= e_\psi^T \varphi(\theta) \ , \\
e_\psi &= \psi_{\varphi'}(\hat\theta_\psi)/|\psi_{\varphi'}(\hat\theta_\psi)| \ , \\
\hat\sigma_\nu^2 &= |j_{(\lambda\lambda)}(\hat\theta_\psi)|/|j_{(\theta\theta)}(\hat\theta)| \ , \qquad (3.18) \\
|j_{(\theta\theta)}(\hat\theta)| &= |j_{\theta\theta}(\hat\theta)||\varphi_{\theta'}(\hat\theta)|^{-2} \ , \qquad (3.19) \\
|j_{(\lambda\lambda)}(\hat\theta_\psi)| &= |j_{\lambda\lambda}(\hat\theta_\psi)||\varphi_{\lambda'}(\hat\theta_\psi)|^{-2} \ . \qquad (3.20)
\end{aligned}$$

Although the expression for q_ψ is relatively complicated, it is not difficult to implement algorithmically, starting from $\ell(\theta; y)$ and a specification for V. This is described in Fraser, Reid and Wu (1999) and implemented there in Maple. The most difficult aspect of making the algorithm robust is the computation of the restricted maximum likelihood estimator $\hat\lambda_\psi$.

It will sometimes be more convenient to have a version of q that is suitable when the nuisance parameter is available only implicitly. This can be obtained from (3.17) using a Lagrange multiplier argument; the details are given in Fraser, Reid and Wu (1999).

3.5 Bayesian asymptotics

Bayesian implementation of approximate inference for a scalar parameter is much simpler. First, the inferential basis is prescribed; one simply computes the marginal posterior for ψ. Second the asymptotic expansions are all computed in the parameter space for fixed data, with the result that the expansions are easier and the change of variable for integration is also easier. The resulting approximation is very similar to the frequentist version, and the precise nature of the difference helps to shed some light on the role of the prior.

Laplace approximation to the marginal posterior density was described in the previous section:

$$\begin{aligned}\pi_m(\psi|y) &\doteq \frac{1}{\sqrt{(2\pi)}}\exp\{\ell_p(\psi)-\ell_p(\hat\psi)\}|j_p(\hat\psi)|^{1/2}\cdot \\ &\qquad \frac{|j_{\lambda\lambda}(\hat\psi,\hat\lambda)|^{1/2}}{|j_{\lambda\lambda}(\psi,\hat\lambda_\psi)|^{1/2}}\frac{\pi(\psi,\hat\lambda_\psi)}{\pi(\hat\psi,\hat\lambda)} \qquad (3.21)\\ &= \frac{1}{\sqrt{(2\pi)}}\exp\{\ell_a(\psi)-\ell_a(\hat\psi_a)\}|j_a(\hat\psi)|^{1/2}\frac{\pi(\psi,\hat\lambda_\psi)}{\pi(\hat\psi,\hat\lambda)}. \qquad (3.22)\end{aligned}$$

Expression (3.22) has the same structure as the p^* approximation and the tangent exponential model approximation; i.e. it is a density of "Laplace type" which is readily integrated to give approximate posterior probabilities.

Expression (3.22) is integrated as follows

$$\begin{aligned}\int_\psi^\infty \pi_m(\psi|y)d\psi &= \int_\psi^\infty \frac{1}{\sqrt{2\pi}}\exp\{\ell_a(\psi)-\ell_a(\hat\psi_a)\}\{j_a(\hat\psi)\}^{1/2}\frac{\pi(\psi,\hat\lambda_\psi)}{\pi(\hat\psi,\hat\lambda)}d\psi \\ &= \int_{r_\psi}^\infty \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}r^2\right)\frac{r}{-\ell_a'(\psi)\{j_a(\hat\psi)\}^{-1/2}}\frac{\pi(\psi,\hat\lambda_\psi)}{\pi(\hat\psi,\hat\lambda)}dr \\ &= \int_{r_\psi}^\infty \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}r^2\right)\left(\frac{r}{q_B}+1-1\right)dr \\ &\doteq \Phi(r)+\int r\phi(r)\left(\frac{1}{q_B}-\frac{1}{r}\right)dr \\ &\doteq \Phi(r)+\left(\frac{1}{q_B}-\frac{1}{r}\right)\phi(r) \qquad (3.23)\end{aligned}$$

where

$$\begin{aligned} r &= \mathrm{sign}(q_B)[2\{\ell_a(\hat\psi)-\ell_a(\psi)\}]^{1/2} \\ q_B &= -\ell_a'(\psi)\{j_a(\hat\psi)\}^{-1/2}\frac{\pi(\hat\psi,\hat\lambda)}{\pi(\psi,\hat\lambda_\psi)}: \end{aligned}$$

compare (3.13), (3.14) and (3.17). The relative error in approximation (3.23) is $O(n^{-3/2})$, which is verified by obtaining the asymptotic expansion of r in

terms of q in the form $r = q + (a/\sqrt{n})q^2 + (b/n)q^3 + O(n^{-3})$, and noting that $d(1/r - 1/q) = O(n^{-1})$.

The result obtained starting from (3.21) has q_B and r replaced by

$$\begin{aligned} q_B &= -\ell_p'(\psi)\{j_p(\hat\psi)\}^{-1/2}\frac{\pi(\hat\psi,\hat\lambda)}{\pi(\psi,\hat\lambda_\psi)}\frac{|j_{\lambda\lambda}(\psi,\hat\lambda_\psi)|^{1/2}}{|j_{\lambda\lambda}(\hat\psi,\hat\lambda)^{1/2}} \\ r &= \text{sign}(q_B)[2\{\ell_p(\hat\psi) - \ell_p(\psi)\}]^{1/2} \end{aligned}$$

The derivation of the tail area approximation sketched above is of exactly the same form as the derivation of the tail area approximations discussed in Section 3.4. The details in Section 3.4 are more difficult, because it is more difficult to get workable expressions of the remainder term. The integration of Section 3.4 is in the sample space, so the change of variable to r in the first step of the integration involves a Jacobian on the sample space and this is where the ancillary directions are needed. It is much easier to work backwards from the Bayesian derivation to arrive at the non Bayesian version. A version of (3.23) was derived in DiCiccio, Field and Fraser (1990), and in full generality in DiCiccio and Martin (1993). Series expansions for r in terms of q are given in Cakmak et al (1998). The structure of the Bayesian significance function is examined in Sweeting (1995).

One conclusion available from comparing (3.23) and (3.13) is that the approximate tail areas will be identical if $q_B = q$. Since q_B involves the prior, this may be considered a definition of a "matching prior", i.e. a prior for which frequentist and Bayesian inference agree to some order of approximation. Such a prior necessarily depends on the data, as is evident from inspection of the expression for q given at (3.16). Such priors are called *strong matching* priors in Fraser and Reid (2001) where it is shown that in most cases a usable solution does not emerge unless we consider matching only to $O(n^{-1})$. In this case the strong matching prior is a version of the

observed Fisher information.

Data dependent priors are somewhat unusual objects from a Bayesian point of view, but they do arise in other contexts as well. Box and Cox (1964) proposed a data dependent prior for the transformed regression model as the only way to sensibly take account of the scale of the observations. Wasserman (2000) shows in a mixture model that matching priors must be data dependent at least in a weak sense. Pierce and Peters (1994) discuss the asymptotic frequentist-Bayesian connection from a slightly different point of view and come to the same conclusion: in going from $O(n^{-1})$ to $O(n^{-3/2})$ there is an essential dependence on the sample space in the frequentist version that has no counterpart in a Bayesian analysis with fixed prior. This dependence is seen in the derivation of (3.13) or more simply (2.8) as required by the integration of p^* or p_{TEM} in the sample space.

A more conventional approach to matching priors involves Edgeworth expansions for the posterior density. As described at the end of Section 2, Welch and Peers (1963) showed that in models with a single parameter of interest the Jeffreys' prior, $\pi(\theta) \propto \{i(\theta)\}^{1/2}$, is the unique prior for which the $(1-\alpha)$ posterior quantile has the property under the sampling model of giving a one-sided confidence bound with confidence coefficient $1-\alpha+O(n^{-1})$. Unfortunately, subsequent attempts to generalize this result to higher order and to models with nuisance parameters have generally not been successful. The higher order asymptotic results discussed here make clear in retrospect that this was inevitable. A survey of work on matching priors is given in Reid, Mukerjee and Fraser (2001).

4 Applications and Implementation

Implementation of higher order approximations has lagged behind the theoretical development, in part because of the complexity of expressions like (3.16), in part because the inferential basis is not well understood, and in part because the improvement offered by refined approximation is in many applications overshadowed by the sample size, the complexity of the sampling frames, the provisional nature of the model, and so on.

By far the most numerous applications of the asymptotic theory discussed in Section 3 available in the literature are illustrative examples, typically involving highly simplified models and extensive simulation. These are meant to highlight the accuracy of the approximations and the sometimes dramatic improvement in going from a first order approximation to a second or third order approximation. A fairly comprehensive list of published examples is given in Reid (1996).

An amusing but highly artificial example is a sample of size 1 from a Cauchy location model (Fraser, 1990; Barndorff-Nielsen, 1991). In this case the first order approximations to the distributions of the Wald, score and likelihood root statistic differ by hundreds of orders of magnitude, whereas the third order approximation is very nearly exact.

Our focus throughout continues to be accurate approximation of p-values for testing $H : \psi = \psi_0$. These may be computed for a fixed data value and a range of values for ψ_0, leading to confidence limits or more generally a confidence distribution function or significance function (Fraser, 1990; Efron, 1997).

Example 4.1: Normal coefficient of variation

Table 4.1 gives selected points of the significance function for the one parameter model $N(\theta, \theta^2)$ and a sample of size 5 generated randomly from the

Table 4.1: Comparison of exact and approximate p-values for selected values of θ in the model $N(\theta, \theta^2)$. Data vector is $y = (0.69, 1.56, 3.69, 2.09, 1.10)$.

θ	0.6	0.8	1.6	1.8	2.0
Method					
exact	0.9976	0.8863	0.0794	0.0416	0.0228
r^* using q or u (3.13)	0.9976	0.8855	0.0787	0.0412	0.0225
r^* using Skovgaard	0.9975	0.8837	0.0768	0.0400	0.0217
normal approximation	0.9960	0.8465	0.0513	0.0249	0.0128

$N(1, 1)$ distribution. The exact distribution is compared to the usual normal approximation for r, the third order approximation described in Section 3.3, and a second order version due to Skovgaard (1996). In this example, discussed in more detail in Fraser, Reid and Wu (1999), q and u are identical, as there is a unique ancillary.

Example 4.2: Log-normal distribution

The illustrative data set given in Table 4.2 is is a small subset of five observations (on Volkswagen cars) from the car crash test data from the Data and Story Library. In the full data set the response Y is a measure of injury, and appears to be approximately lognormally distributed. There are a number of covariates relating to properties of the various cars. We use the model $\log Y \sim N(\mu, \sigma^2)$ and define the parameter of interest to be the mean of Y. Figure 4.1 illustrates the significance function for the log of the mean: $\psi = \mu + \sigma^2/2$. The significance function using the usual normal approximation to r is compared to the third order approximations (3.12) and (3.13); the latter two cannot be distinguished on the graph.

Example 4.3: Bivariate normal

Our next example provides detailed calculation for the general formula

Figure 4.1

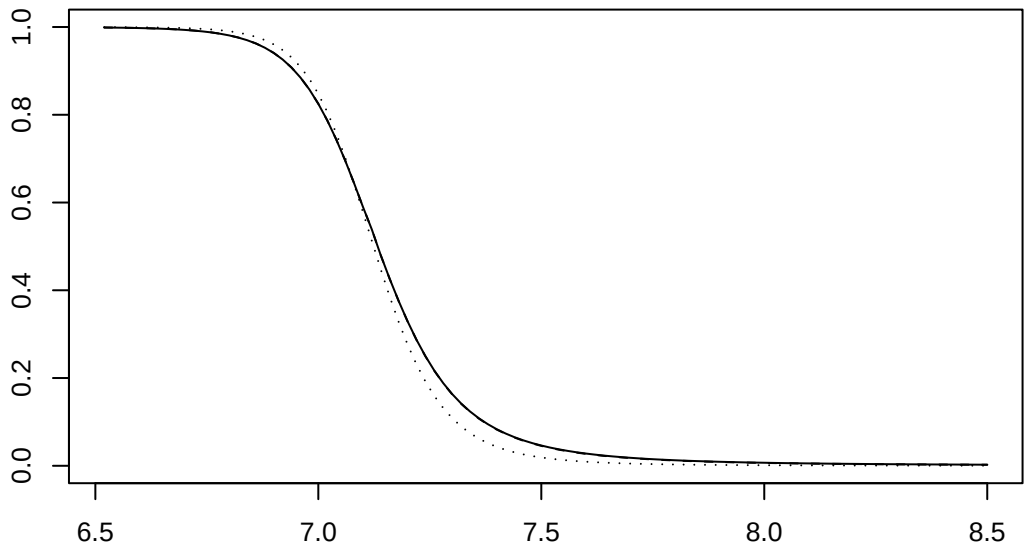

Figure 4.3: The significance function for $\psi = \mu + \sigma^2/2$, based on the data in Table 4.2 and a log-normal model. Solid line is the third order approximation using (3.12) and dashed line is the standard normal approximation to the distribution of the likelihood root r. The r^* approximation is indistinguishable from (3.12).

in the case of no nuisance parameters. The emphasis is on the calculation of the approximate ancillary using a pivotal statistic. The model is

$$f(y_1, y_2; \theta) = \frac{1}{2\pi(1-\theta^2)^{1/2}} \exp\left\{-\frac{1}{2(1-\theta^2)}(y_1^2 + y_2^2 - \theta y_1 y_2)\right\}, \qquad (4.1)$$

a (2,1) curved exponential family. The minimal sufficient statistic based on a sample of size n is (S, T), where

$$S = \Sigma Y_{1i} Y_{2i}/n, \qquad T = \Sigma(Y_{1i}^2 + Y_{2i}^2)/(2n). \qquad (4.2)$$

Working in the sample space defined by (S, T) we have the pivotal statistics

$$\begin{aligned} Z_1 &= \frac{T+S}{1+\theta} = \frac{\Sigma(Y_{1i}+Y_{2i})^2}{2n(1+\theta)} \\ Z_2 &= \frac{T-S}{1-\theta} = \frac{\Sigma(Y_{1i}-Y_{2i})^2}{2n(1-\theta)} \end{aligned}$$

which are independently distributed as χ_n^2/n. Using these in (3.8) to define the vector V gives

$$V = \frac{1}{1-\hat{\theta}^2}\begin{pmatrix} t - \hat{\theta}s \\ s - \hat{\theta}t \end{pmatrix} \qquad (4.3)$$

where $\hat{\theta}$ is the real root of the cubic equation

$$\hat{\theta}^3 - s\hat{\theta}^2 + (2t-1)\hat{\theta} - s = 0.$$

The ancillary curve defined by the components of V is plotted in Figure 4.2, where the axes are rotated from the S, T plane to the $T-S, T+S$ plane. Also shown in Figure 4.2 is the curve where the approximately ancillary statistic $A = (T-1)/\sqrt{S^2+1}$ is constant. This statistic was used by Wang (1993), following a suggestion in Cox and Hinkley (1974, Ch.2). Note that A has mean zero and variance 1, although its higher cumulants depend of course on θ.

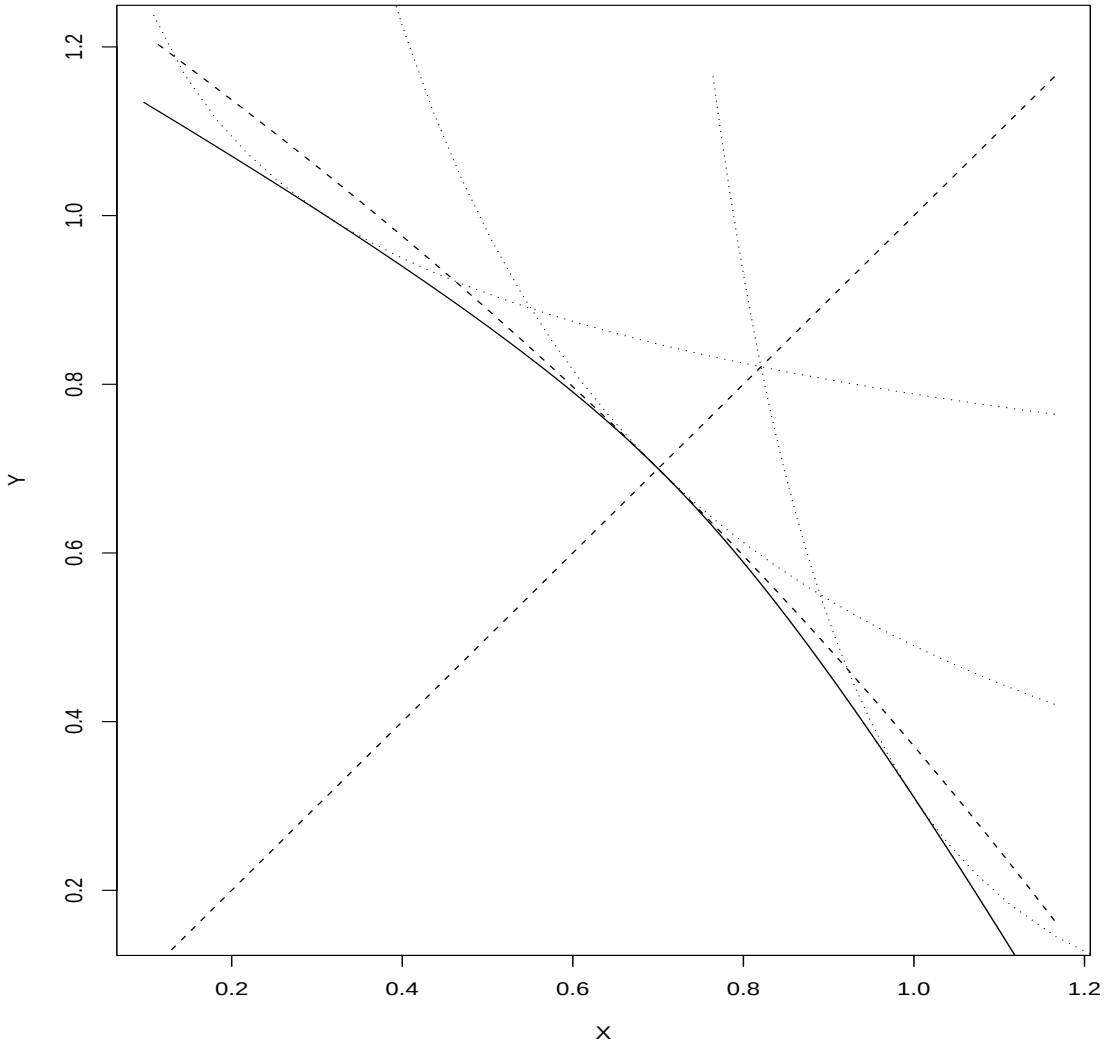

Figure 4.4: The curve defined by the tangent directions to the second order ancillary (solid line), and the curve defined by a fixed value of Wang's ancillary statistic (dashed line). The line $Y = X$ corresponds to $\hat{\theta} = 0$; where $X = T - S$ and $Y = T + S$. The dotted lines show the local ancillary defined by integrating the vector V at a fixed value of ρ.

From (4.1), the likelihood function is

$$\ell(\theta; s, t) = -\frac{n}{2}\log(1-\theta^2) - \frac{n}{1-\theta^2}(t - \theta s) \tag{4.4}$$

giving

$$\begin{aligned} \ell_{;V}(\theta; s, t) &= \frac{\partial \ell}{\partial s}V_1 + \frac{\partial \ell}{\partial t}V_2 \\ &= \frac{n\theta}{1-\theta^2}\frac{t - \hat{\theta}s}{1-\hat{\theta}^2} - \frac{n}{1-\theta^2}\frac{s - \hat{\theta}t}{1-\hat{\theta}^2} \end{aligned}$$

and

$$\ell_{\theta;V}(\hat{\theta}) = n(\hat{\theta}s + t)/(1-\hat{\theta}^2)^2.$$

This, combined with $j(\hat{\theta})$ in (2.13) gives an explicit expression for q, and hence the approximate significiance function for θ.

Note that no calculation of an explicit approximate ancillary is involved. In this case it is possible to find an approximate ancillary, by embedding the model in a two-parameter exponential family, for example by treating $\alpha_1 Z_1$ and $\alpha_2 Z_2$ as independent chi-squared random variables, where $\alpha_1 = \alpha/(1-\theta)$, $\alpha_2 = \alpha/(1+\theta)$, thus recovering (4.4) when $\alpha = 1$. It is possible to calculate the components of the r^*-type statistic for testing $\alpha = 1$ in this full model, thus giving an explicit expression for an approximate ancillary, although the detailed expressions are in this case unenlightening.

Example 4.4: Gamma hyperbola

Another $(2,1)$ exponential family where the computation of the ancillary statistic is particularly straightforward is the model

$$f(y;\theta) = \exp\left(-\frac{e^{-\theta}}{\theta}y_1 - \theta y_2\right), \quad \theta > 0; y_1 > 0, y_2 > 1. \tag{4.5}$$

This example is considered in this context in Barndorff-Nielsen and Chamberlin (1994) and DiCiccio, Field and Fraser (1990). An explicit expression

for an approximately ancillary statistic is derived in Barndorff-Nielsen and Wood (1998). Because a simple scale model underlies (4.5), pivotal statistics are readily available as

$$z_1 = \frac{1}{\theta} e^{-\theta} y_1 \ , \quad z_2 = \theta(y_2 - 1)$$

from which we have

$$V = -\Big(1 + \frac{1}{\hat\theta}\Big) y_1 \ , \ -\frac{1}{\hat\theta}(y_2 - 1) \ ,$$

$$\varphi(\theta) = \ell_{;V}(\theta) \ = \ \frac{-e^{-\theta}}{\theta}\Big(1 + \frac{1}{\hat\theta}\Big) y_1 + \frac{\theta}{\hat\theta}(y_2 - 1)$$

and

$$q = \left\{ \left(\frac{e^{-\theta}/\theta}{e^{-\hat\theta}/\hat\theta} - \frac{\theta}{\hat\theta} \right) y_2 + \frac{\theta}{\hat\theta} \right\} \left(y_2 + \frac{2y_2}{\hat\theta} - \frac{1}{\hat\theta} \right)^{-1} \left\{ \frac{y_2(2/\hat\theta + 2 + \hat\theta)}{1 + \hat\theta} \right\}^{1/2}$$

where y_1, y_2 and $\hat\theta$ are related via the likelihood equation $\ell'(\hat\theta) = 0$ as

$$\frac{e^{-\hat\theta}}{\hat\theta} \left(1 + \frac{1}{\hat\theta} \right) y_1 = y_2 \ .$$

Numerical comparisons of (2.7) using q or u from Barndorff-Nielsen and Chamberlin (1991) are given in Fraser, Reid and Wu (1999).

Example 4.1 continued: log-normal

We now illustrate the step-by-step calculations for the third order approximation in the case of nuisance parameters, using the log-normal model of Figure 4.1. We suppose $X_1, \ldots, X_n$ are independent and follow the $N(\mu, \sigma^2)$ distribution, and that the parameter of interest is $\psi = \mu + (1/2)\sigma^2$, the log of the mean of the associated log-normal distribution. Using the pivotal $z_i = (x_i - \mu)/\sigma$, the ith row of the $n \times 2$ matrix V is $(1, (x_i - \hat\mu)/\hat\sigma)$. From

this we have $\varphi(\theta) = \ell_{;V}(\theta) = \sum_{i=1}^{n} (\partial\ell/\partial x_i)\cdot v_i = (n(\hat{\mu}-\mu)/\sigma^2, -n\hat{\sigma}/\sigma^2)$. Note that φ is an affine transformation of the canonical parameter, as it must be, by construction. We now extract from φ the component corresponds to our parameter of interest ψ, using (3.19) and (3.20)–(3.22). It is simpler to work with $\varphi = (\mu/\sigma^2, -1/2\sigma^2)$, from which we can express ψ in terms of φ_1 and φ_2 as $\psi = -(2\varphi_1 + 1)/2\varphi_2$, and

$$\psi'_{\varphi}(\theta) = \left(\frac{1}{\sqrt{(1+\psi^2)}}, \frac{\psi}{\sqrt{(1+\psi^2)}}\right)$$

which in this case does not depend on μ or σ^2 except through ψ. The resulting expression for $\nu(\hat{\theta}) - \nu(\hat{\theta}_\psi)$ is

$$\frac{(\bar{y}-\psi)/\hat{\sigma}^2 + (1/2)}{\sqrt{(1+\psi^2)}}$$

which is then standardized by the ratio of information functions, using (3.19) and (3.20).

Example 4.5: exponential family

In special classes of models the expressions for u and q given in Section 3.3 simplify substantially. First, in the scalar parameter case discussed in Section 2.2, where q is given by (2.13), suppose we have an exponential family model $f(y;\theta) = \exp\{\theta y - c(\theta) - d(y)\}$. In i.i.d. sampling from this model the sample sum Σy_i is minimal sufficient and a one-to-one function of $\hat{\theta}$. Thus $\ell_{;V}$ is replaced by $\ell_{;y} = \theta$, and the expression for q simplifies to

$$q = (\hat{\theta} - \theta)\{j(\hat{\theta})\}^{1/2} ,$$

the standardized maximum likelihood estimate. Similarly, if the model is $f(y;\psi,\lambda) = \exp\{\psi y_1 + \lambda^T y_2 - c(\psi,\lambda) - d(y)\}$ then expression (3.16) simplifies to

$$q = (\hat{\psi} - \psi)\{j_p(\hat{\psi})\}^{1/2}\{\rho(\hat{\psi},\psi)\}^{-1}$$

where

$$\rho(\hat{\psi},\psi) = \frac{|j_{\lambda\lambda}(\psi,\hat{\lambda}_\psi)|^{1/2}}{|j_{\lambda\lambda}(\hat{\psi},\hat{\lambda})|^{1/2}} \tag{4.6}$$

is an adjustment to allow for estimation of the nuisance parameters.

The asymptotically equivalent approximation using the adjusted likelihood can be obtained by saddlepoint approximation to the exact conditional density $f(y_1|y_2)$, which is of exponential family form and free of λ.

Example 4.6: Location models

In the one parameter location model $f(y-\theta)$, the exact ancillary from n independent samples is $(a_1,\ldots,a_n)$ where $a_i = y_i - \hat{\theta}$. The exact conditional density of $\hat{\theta}$ given a is obtained by renormalizing the likelihood function, and the approximation (2.10) or (2.13) gives

$$q = \ell'(\theta)\{j(\hat{\theta})\}^{-1/2}$$

the standardized score statistic.

In the location regression setting where $y_i = \mu + \psi x_{1i} + \lambda^T \mathbf{x}_{2i} + e_i$, say where $e_i \sim f(e)$, the general expression for q in (4.16) reduces to

$$q = -\ell_p'(\psi)\{j_p(\hat{\psi})\}^{-1/2}\ \rho(\hat{\psi},\psi).$$

To see this note that using the pivotal $z_i = y_i - \psi x_{1i} - \lambda^T \mathbf{x}_{2i}$ gives the ith row of V as $(1\ x_{1i}\ \mathbf{x}_{2i})$, and hence

$$\varphi(\theta) = \ell_{;V} = \big(\Sigma g(e_i), \Sigma x_{1i} g(e_i), \Sigma \mathbf{x}_{2i}^T g(e_i)\big)$$

where $g(e) = d\log f(e)/de$ and $e_i = e_i(\theta) = y_i - \mu - \psi x_{1i} - \lambda^T \mathbf{x}_{2i}$. The components of φ are simply the score functions $(\partial\ell/\partial\mu, \partial\ell/\partial\psi, \partial\ell/\partial\lambda^T)$. Similarly $\ell_{\theta;V}(\hat{\theta}) = j_{\theta\theta}(\hat{\theta})$ and $\ell_{\lambda;V}(\hat{\theta}_\psi) = j_{\lambda\lambda}(\hat{\theta}_\psi)$, leading to

$$\begin{aligned} q &= -\ell_\psi(\hat{\theta}_\psi)\frac{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}{|j(\hat{\theta})|} \\ &= -\ell_\psi(\hat{\theta}_\psi)|j_p(\hat{\theta}_\psi)|^{-1/2}\{\rho(\hat{\psi},\psi)\}, \end{aligned}$$

using (2.14) and (4.6).

The improvements achieved by relatively simple adjustments in a number of examples suggest that if the approximations could be made simple from the point of view of the user, they would be well worth incorporating in routine analyses. In some cases the approximations could reduce or avoid the need for lengthy simulations, for example. In this connection applications the might be described as case studies are very useful. These involve moderately realistic models, although the data sets are often fairly stylized examples from the statistical literature. The focus is typically on illustrating the use of higher order methods in models of potential interest for applications, and on comparing first order and third order methods. Usually simulations are too cumbersome to enable comparison of the approximations to the "exact" answer to be computed.

A substantial number of such case studies have been carried out by Butler and co-authors; see Reid (1996) for a description of this work and Butler, Huzurbazar and Booth (1992ab) and Butler, Booth and Huzurbazar (1993). Several of their applications involve the construction of exponential families for which the related approximations are particularly straightforward.

Fraser, Wong and Wu (1999) study a class of linear regression models of the form $y = x(\beta)+\sigma e$, where $x(\beta)$ is a known and possibly nonlinear function of x and unknown parameter β, and e is either a normal or t distribution. In this class of models the pivotal statistic $z = \{y-x(\beta)\}/\sigma$ leads to an explicit expression for φ which generalizes in a natural way the location model version given above. Expressions for φ, nu and q are given in (35), (38) and (41) of their paper. The formulae are illustrated on four simulation studies and four sample datasets from the statistical literature on nonlinear regression.

Bellio (1999, Ch.1) gives an overview of r^*–type approximations with

particular emphasis on the version proposed by Skovgaard (1996). In Chapter 2 he gives a systematic account of higher order likelihood inference for generalized linear models of the form

$$y_{ij} = \mu(x_i, \beta) + e_{ij} \tag{4.7}$$

where $e_{ij} \sim N(0, g^2(x_i, \beta, \rho))$ and the forms of $\mu(\cdot)$ and $g^2(\cdot)$ are assumed known. His work thus illustrates the use of higher order likelihood based methods in models of practical importance. Of particular interest is the use of enhanced profile plots to compare the usual normal approximation to the higher order version. Bellio (2000) illustrates the use of Skovgaard's approximation in inverse regression problems. An important contribution of Bellio's work is to consider as well inference for vector parameters of interest, based on an r^*–type extension developed in Skovgaard (2001). Several aspects of higher order approximations that can be applied to the study mixed linear models are described in Bellio (1999, Ch.4).

The potentially most useful category of applications is what might be called "nearly automatic" software. An attempt at this was made in Fraser, Reid and Wu (1999) using Maple. The user provides the model function and the pivotal statistic, and the software computes the necessary sample space derivatives. The computations of the profile log likelihood and its curvature needs a fair bit of input from the user.

The most important advances in nearly automatic software are provided by Brazzale (1999, 2000). Splus libraries, available at `statlib`, require from the user only a specification of the model function and the parameter of interest. Three classes of models are currently incorporated into the software:

1. Conditional inference for logistic regression and log linear models

2. Marginal inference for linear regression with non-normal errors

3. Conditional inference for model (4.2).

Again the most difficult aspect of the computation is the evaluation of the profile log likelihood and its curvature. Another, more minor, difficulty arises in computing the entire significance function: at the sample point $\hat{\theta} = \theta$, both r and q are zero and the r^* and Lugannani and Rice type approximations are no longer valid. A simple expedient incorporated in Brazzale's software is to fit a spline curve through the middle of the significance function. However, automatic selection of the "middle" is not straightforward, as the range of numerical instability is generally problem specific. Some details on the implementation of Brazzale's approach are described in Bellio and Brazzale (2001).

There is limited experience with much more general models. Butler (2000, 2001) has investigated saddlepoint approximations for first passage time distributions in very complex reliability trees, Huzurbazar (2000) has applied similar ideas for models of cellular telephone nextworks, and Yau and Huzurbazar (2002) have applied similar ideas in multi-stage survival models. This work is not as yet very closely related to likelihood inference, but rather closer to numerical approximation theory. Kolassa and Tanner (1994) combined third order approximations to conditional densities in exponential families with Markov chain Monte Carlo methods. Limited experience with higher order approximations for empirical likelihood indicates that theoretically accurate formulae do not have good finite simple behaviour (Davison, Corcoran and Spady, 1999) indicating that continued development of case studies is needed.

Although implementation of Bayesian approximations is quite straightforward, the approximations do not appear to be much used in the literature on applications of Bayesian methods, preference being given to Markov chain

Monte Carlo computation of 'exact' posteriors. Notable exceptions are Hsu (1995), and Tierney (1990). This seems somewhat surprising, as the approximations are very easily computed and the inferential basis is straightforward. Using approximations to compare the effect of different priors and as a check on the accuracy of MCMC methods was suggested in Kass, Tierney and Kadane (1988).

In addition to the need for both more widely available automatic software and a larger library of case studies, there are several other aspects of application and implementation to be investigated. By far the most important is robustness: what happens to the approximations if the model is incorrect, and does the comparison of first order and higher order approximation provide any information on the adequacy of the model. The work by Bellio is an important first step in the use of the higher order approximations for improved diagnostics in regression. In the extension to more complex models, some of which may have a nonparametric component, comparison to bootstrap inference is necessary. Some theoretical discussion is given in DiCiccio and Efron (1996) but the main practical comparisons are due to Bellio (1999).

Establishing connections, if possible, between the type of likelihood inference described here and the inference obtained by Markov chain Monte Carlo methods, typically applied in complex hierarchical models, seems a very valuable next step. The inference of the prior on the results obtained from MCMC methods seems very unclear in most applications, and it would be very useful if the asymptotic theory for likelihood inference could shed some light on this. In a related development, Brazzale (2000, Ch.7) describes the use of MCMC methods for conditional inference in regression models, thus implementing higher order asymptotic solutions by a different route.

Whether or not higher order approximations are used in "real" applications should depend on a considerable extent to the context of the problem. In a large prospective study of a relatively common disease, aspects of study design and sampling bias will be much more important than the type of inference made, which in any case is likely to involve little more than simple summary statistics. In contrast, a comparison of brain activation rates by functional magnetic resonance imaging may well involve a very small number of subjects, even though a large database may be created for each subject. In this setting modelling and precise inference may be expected to be relatively more important.

There are also a number of theoretical areas for potential further development. Higher order asymptotic results for dependent data may potentially be useful in time series and spatial data models. It is not clear if a higher order theory may be available for heavy-tailed distributions. The current interest in several application areas in models more variables than observations suggests that it is important to understand how best to deal with large numbers of nuisance parameters. It also seems likely that a higher order asymptotic theory can be developed for some types of semi-parametric models.

Higher order asymptotics has a very large literature, and it has not been possible to survey all the results here. The most useful general references for the particular emphasis here on tail area approximation and likelihood methods are Skovgaard (1990, 1996, 2001), Barndorff-Nielsen and Cox (1994, Ch. 6), Barndorff-Nielsen and Wood (1998), Fraser, Reid and Wu (1999), and Reid (1996). A web site on higher order asymptotic theory is maintained at the University of Padua (`www.stat.unipd.it/LIKASY`).

Acknowledgments

This paper was presented at the Fifth World Congress of the Bernoulli Society and the Fifty-third Annual Meeting of the Institute of Mathematical Statistics, in Guanajuato, Mexico, May 2000. I would like to thank Martha Cerilla and Victor Perez-Abreu of the local organizing committee for their assistance. Much of the work described here has been developed jointly with Don Fraser over may years. I am grateful to Ib Skovgaard, Ruggero Bellio and Alessandra Brazzale for providing preprints of their work, and to Alessandra Brazzale for providing Figures 3.1 and 3.2. It is a pleasure to acknowledge helpful discussion with David Andrews, David Cox, Andrey Feuerverger, Radford Neal and Augustine Wong.

References

Amari, S.-I. (1985). *Differential-geometric Methods in Statistics.* Lecture Notes in Statistics. Springer-Verlag, New York.

Barndorff-Nielsen, O.E. (1980). Conditionality resolutions. *Biometrika* **67**, 293–310.

Barndorff-Nielsen, O.E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70**, 343–365.

Barndorff-Nielsen, O.E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika* **73**, 307–322.

Barndorff-Nielsen, O.E. (1991). Modified signed log likelihood ratio. *Biometrika* **78**, 557–563.

Barndorff-Nielsen, O.E. (1996). Two index asymptotics. in *Frontiers in Pure and Applied Probability II.* Proceedings of the 4th Russian-Finnish

SYmposium in Probability Theory and Mathematical Statistics. Ed. I. Melnikov, pp. 9–20. TVP Science, Moscow.

Barndorff-Nielsen, O.E. and Chamberlin, S. (1991). An ancillary invariant modification of the signed log likelihood ratio. *Scand. J. Statist.* **18**, 341–352.

Barndorff-Nielsen, O.E. and Chamberlin, S. (1994). Stable and invariant adjusted directed likelihoods. *Biometrika* **81**, 485–499.

Barndorff-Nielsen, O.E. and Cox, D.R. (1994). *Inference and Asymptotics.* Chapman & Hall, London.

Barndorff-Nielsen, O.E., Cox, D.R. and Reid, N. (1986). The role of differential geometry in statistical theory. *Inter. Stat. Inst. Rev.* **54**, 83-96.

Barndorff-Nielsen, O.E., Wood, A.J.T. (1998). On large deviations and choice of ancillary for p^* and r^*. *Bernoulli* **4**, 35–63.

Bellio, R. (1999). *Likelihood asymptotics: applications in biostatistics.* PhD Thesis, University of Padua, Italy.

Bellio, R. (2000) Likelihood methods for controlled calibration. preprint.

Bellio, R. and Brazzale, A.R. (2001). Higher-order asymptotics unleashed: software design for nonlinear heteroscedastic regression. preprint

Bellio, R., Jensen, J.E. and Seiden, P. (2000) Applications of likelihood asymptotics for nonlinear regression in herbicide bioassays. *Biometrics* **56**, 1204–1212.

Bentkus, V., Götze, F. and vanZwet, W.R. (1997). An Edgeworth expansion for symmetric statistics. *Ann. Statist.* **25**, 851–896.

Beran, J. (1994). *Statistics for Long Memory Processes.* Chapman & Hall, London.

Bickel, P. and Freedman, D. (1981). Some asymptotic theory for the boot-

strap. *Ann. Statist.* **9**, 1196–1217.

Bloznelis, M. and Götze, F. (2000). An Edgeworth expansion for finite population U-statistics. *Bernoulli*, **6**, 629–760.

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations (with discussion). *J. Roy. Statist. Soc.* B

Brazzale, A. (1999). Approximate conditional inference in logistic and loglinear models. *J. Comp. Graph. Statist.* **8**, 653–661.

Brazzale, A. (2000). *Practical Small-Sample Parametric Inference.* PhD Thesis Department of Mathematics, Swiss Federal Institute of Technology Lausanne.

Butler, R. (2001). First passage time disctributions in semi-Markov processes and their saddlepoint approximation. in *Data Analysis and Statistical Foundations.*, Ed. E. Saleh. Nova Science, New York.

Butler, R. (2000). Reliabilities for feedback systems and their saddlepoint approximation. *Statist. Sci.* **15**, 279–298.

Butler, R., Booth, J., and Huzurbazar, S. (1993). Saddlepoint approximations for tests of block independence, sphericity and equal variances and covariances. *J. Roy. Statist. Soc.* B **55**, 171–184.

Butler, R., Huzurbazar, S. and Booth, J. (1992a). Saddlepoint approximations for generalized variance and Wilks' statistic. *Biometrika* **79**, 157–170.

Butler, R., Huzurbazar, S. and Booth, J. (1992b). Saddlepoint approximation for the Bartlett-Nanda-Pillai trace statistic. *Biometrika* **79**, 705–716.

Cakmak, S., Fraser, D.A.S., McDunnough, P., Reid, N. and Yuan, X. (1998). LIkelihood centered asymptotic model: exponential and location model versions. *J. Statist. Plann. Infer.* **66**, 211–222.

Cheah, P.K., Fraser, D.A.S. and Reid, N. (1995). Adjustments to likelihood and densities; calculating signficance. *J. Statist. Res.* **29**, 1–13.

Cox, D.R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc.* B **34**, 187–220.

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

Davison, A.C., Corcoran, S. and Spady, R. (1999). Reliable inference from empirical likelihood. preprint.

Davison, A.F. and Hinkley, D.V. (1998) *Bootstrap Methods and Their Application.* Springer-Verlag, New York.

Dawid, A.P. (1991). Fisherian inference in likelihood and prequential frames of reference (with discussion). *J. Roy. Statist. Soc.* B **53**, 79–109.

DiCiccio, T.J. and Efron, B. (1992). More accurate confidence intervals in exponential families. *Biometrika* **79**, 231–245.

DiCiccio, T.J. and Efron, B. (1996). Bootstrap confidence intervals. (with discussion). *Statist. Sci.* **11**, 189–228.

DiCiccio, T.J., Field, C.A. and Fraser, D.A.S. (1990). Approximations of marginal tail probabilities and inference for scalar parameters. *Biometrika* **77**, 77–95.

DiCiccio, T.J., Hall, P. and Romano, J. (1991). Empirical likelihood is Bartlett-correctable. *Ann. Statist.* **19**, 1053–1061.

DiCiccio, T.J. and Martin, M.A. (1993). Simple modifications for signed roots of likelihood ratio statistics. *J. Roy. Statist. Soc.* B **55**, 305–316.

Dinges, H. (1986). Asymptotic normality and large deviations. *Proceedings of the 10th Prague Conference on information theory, statistical decision functions, and random processes.* Academia, Prague.

Efron, B. (1997) Bayes and likelihood calculations from confidence intervals. *Biometrika* **80**, 3–26.

Efron, B. and Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected information (with discussion). *Biometrika* **65**, 457–482.

Fan, J.Q. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21**, 196–216.

Fisher, R.A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. *Monthly Notices. R. Astronomical Soc.* **80**, 758–770. Reprinted (1950) in *Fisher's Contributions to Mathematical Statistics*. New York, Wiley.

Fisher, R.A. (1934). Two new properties of mathematical likelihood. *Proc. Roy. Soc.* A **144**, 285–304.

Fraser, D.A.S. (1968). *The Structure of Inference*. John Wiley & Sons, New York.

Fraser, D.A.S. (1988). Normed likelihood as saddlepoint approximation. *J. Mult. Anal.* **27**, 181–193.

Fraser, D.A.S. (1990). Tail probabilities from observed likelihoods. *Biometrika* **77**, 65–76.

Fraser, D.A.S. (1991). Statistical inference: likelihood to significance. *J. Amer. Statist. Assoc.* **86**, 258–265.

Fraser, D.A.S. and McDunnough, P. (1984). Further remarks on the normality of likelihood and conditional analyses. *Canad. J. Statist.* **12**, 183–190.

Fraser, D.A.S. and Reid, N. (1993). Simple asymptotic connections between densities and cumulant generating functions leading to accurate ap-

proximations for distribution functions. *Statist. Sinica* **3** 67–82.

Fraser, D.A.S. and Reid, N. (1995). Ancillaries and third order significance. *Util. Math.* **47**, 33–53.

Fraser, D.A.S. and Reid, N. (2001). Strong matching of frequentist and Bayesian parametric inference. *J. Statist. Plann. Inf.*, to appear.

Fraser, D.A.S., Reid, N. and Wu, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86**, 249–264.

Fraser, D.A.S., Wong, A. and Wu, J. (1999). Regression analysis, nonlinear or nonnormal: simple and accurate p-values from likelihood analysis *J. Amer. Statist. Ass.* **94**, 1286–1295.

Freedman, D.L. (1999). On the Bernstein-vonMises theorem with infinite-dimensional parameters. *Ann. Statist.* **4**, 1119–1140.

Ghosh, M. and Mukerjee, R. (1998). Recent developments in matching priors. in *Applied Statistical Science III.* Eds. S.E. Ahmed et al. Nova Science, New York. 227–252.

Hastie, T. and Loader, C. (1993). Local regression: automatic kernel carpentry. (with discussion). *Statist. Sci.* **8**, 120–143.

Hsu, J. (1995) Generalized Laplace approximation in Bayesian inference. *Canad. J. Statist.* **23**, 399–410.

Huzurbazar, A.V. (2000). Modelling and analysis of engineering systems data using flowgraph models. *Technometrics* **42**, 300–306.

Jensen, J.L. (1992). Modified signed likelihood and saddlepoint approximations. *Biometrika* **79**, 693–703.

Johnson, R.A. (1970). Asymptotic expansions associated with posterior distributions. *Ann. Math. Statist.* **41**, 851–864.

Kass, R., Tierney, L.J. and Kadane, J. (1988) Asymptotics in Bayesian com-

putation. in *Bayesian Statistics III*. Eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith. Clarendon Press, Oxford.

Kolassa, J. and Tanner, M. (1994). Approximate conditional inference in exponential families with the Gibbs sampler. *J. Amer. Statist. Assoc.* **89**, 697–702.

Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation*. John Wiley & Sons, New York.

Lugannani, R. and Rice, S.O. (1980). Saddlepoint approximation for the distribution of the sum of independent random variables. *Adv. Appl. Probab.* **12**, 475–490.

Mukerjee, R. and Reid, N. (2000). On the Bayesian approach for frequentist computations. *Braz. J. Statist.* **14**, 159–166.

Murphy, S.A. and van der Vaart, A.W. (2000). On profile likelihood (with discussion). *J. Amer. Statist. Assoc.* **95**, 449–485.

Mykland, P.K. (1995). Martingale expansions and second order inference. *Ann. Statist.* **23**, 707–731.

Mykland, P.K. (2000). Bartlett identities and large deviations in likelihood theory. *Ann. Statist.* **27**, 1105–1117.

Neyman, J. and Scott, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32.

Nicolau, A. (1993). Bayesian intervals with good frequentist behaviour in the presence of nuisance parameters. *J. Roy. Statist. Soc.* B **55**, 377–390.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–249.

Owen, A. (2001). *Empirical Likelihood*. Champman & Hall/CRC Press, Boca Raton.

Peers, H.W. (1965). On confidence points and Bayesian probability points in the case of several parameters. *J. Roy. Statist. Soc.* B **27**, 9–16.

Pfanzagl, J. (1985). *Asymptotic Expansions for General Statistical Models.* Lecture Notes in Statistics 31. Springer-Verlag, New York.

Pierce, D.A. and Peters, D.L. (1992). Practical use of higher order asymptotics for multiparameter exponential families (with discussion). *J. Roy. Statist. Soc.* B **54**, 701–737.

Pierce, D.A. and Peters, D.L. (1994). Higher-order asymptotics and the likelihood principle: one-parameter models. *Biometrika* **81**, 1–10.

Portnoy, S.L. (1984) Asymptotic behaviour of M-estimates of p regression parameters when p^2/n is large I. *Ann. Statist.* **12**, 1298–1309.

Portnoy, S.L. (1985) Asymptotic behaviour of M-estimates of p regression parameters when p^2/n is large II. *Ann. Statist.* **13**, 1403–1417.

Reid, N. (1988). Saddlepoint methods and statistical inference. (with discussion). *Statist. Sci.* **3** 213–238.

Reid, N. (1996). Likelihood and higher-order approximations to tail areas: A review and annotated bibliography. *Canad. J. Statist.* **24**, 141–166.

Reid, N., Mukerjee, R. and Fraser, D.A.S. (2001). Some aspects of matching priors. preprint.

Rotnizky, A., Cox, D.R., Bottai, M. and Robins, J. (2000). Likelihood based inference with a singular information matrix. *Bernoulli* **6**, 243–284.

Severini, T.A. (1999). An empirical adjustment to the likelihood ratio statistic. *Biometrika* **86**, 235–248.

Sartori, N. (2001). Modifications to the profile likelihood in models with nuisance parameters. preprint.

Sartori, N., Bellio, R., Salvan, A. and Pace, L. (1999). The directed modified profile likelihood in models with many nuisance parameters.

Biometrika **86**, 735–742.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9**, 1187–1195.

Skovgaard, I.M. (1990). On the density of minimum contrast estimators. *Ann. Statist.* **18**, 779–789.

Skovgaard, I.M. (1996). An explicit large-deviation approximation to one-parameter tests. *Bernoulli* **2**, 145–165.

Skovgaard, I.M. (2001). Likelihood asymptotics. *Scand. J. Statist.* **28**, 3–32.

Smith, R.L. (1985). Maximum likelihood estimation in a class of non-regular cases. *Biometrika* **72**, 67–90.

Smith, R.L. (1989). A survey of nonregular problems. *Bull. Intern. Statist. Inst.* **53**, 353–372.

Stigler, S. (1973). Studies in the history of probability and statistics. XXXII. Laplace, Fisher, and the discovery of the concept of sufficiency. *Biometrika* **60**, 439–445.

Stone, C.J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8**, 1348–1360.

Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040–1053.

Sweeting, T.J. (1995). A framework for Bayesian and likelihood approximations in statistics. *Biometrika* **8**, 1-23.

Tibshirani, R.J. (1989). Noninformative priors for one parameter of many. *Biometrika* **76**, 604–608.

Tierney, L.J. (1990). *Lispstat.* John Wiley & Sons, New York.

Tierney, L.J. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Ass.* **81**, 82–87.

Walker, A.M. (1969). On the asymptotic behaviour of posterior distributions. J. Roy. Statist. Soc. b **31**, 80–88.

Wang, S.J. (1993). Saddlepoint approximations in conditional inference. *J. Appl. Probab.* **30**, 397–404.

Wasserman, L. (2000). Asymptotic inference for mixture models using data dependent priors. *J. Roy. Statist. Soc.* B **25**, 318–329.

Welch, B. and Peers, H.W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc.* B **25**, 318–329.

Yau, L. and Huzurbazar, A.V. (2002). Analysis of censored and incomplete data using flowgraph models. *Statist. Medic.*, to appear.