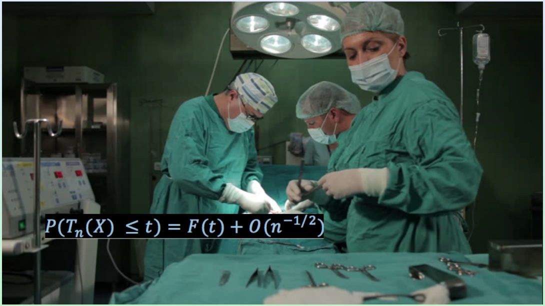


Higher-Order Asymptotics

Todd Kuffner
Washington University in St. Louis
WHOA-PSI 2016

First- and Higher-Order Asymptotics

Classical Asymptotics in Statistics: available sample size $n \to \infty$

First-Order Asymptotic Theory: asymptotic statements that are *correct* to order $O(n^{-1/2})$

Higher-Order Asymptotics: refinements to first-order results

	1^{st} order	2^{nd} order	3^{rd} order	k^{th} order
error	$O(n^{-1/2})$ or $o(1)$	$O(n^{-1})$ or $o(n^{-1/2})$	$O(n^{-3/2})$ or $o(n^{-1})$	$O(n^{-k/2})$ or $o(n^{-(k-1)/2})$

Why would anyone care?

- deeper understanding
- more accurate inference
- compare different approaches (which agree to first order)

Points of Emphasis

Convergence pointwise or uniform?

Error absolute or relative?

Deviation region moderate or large?

Common Goals

Refinements for better small-sample performance

Example Edgeworth expansion (absolute error)
Example Barndorff-Nielsen's R^*

Accurate Approximation

Example saddlepoint methods (relative error)
Example Laplace approximation

Comparative Asymptotics

Example probability matching priors
Example conditional vs. unconditional frequentist inference
Example comparing analytic and bootstrap procedures

Deeper Understanding

Example sources of inaccuracy in first-order theory
Example nuisance parameter effects

Is this relevant for high-dimensional statistical models?

The Classical asymptotic regime is when the parameter dimension p is fixed and the available sample size $n \to \infty$.

What if $p < n$ or p is close to n?

1. Find a meaningful non-asymptotic analysis of the statistical procedure which works for any n or p (concentration inequalities)

2. Allow both $n \to \infty$ <u>and</u> $p \to \infty$.

Some First-Order Theory

Univariate (classical) CLT: Assume $X_1, X_2, \ldots$ are i.i.d. with $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$.
Define $Y_i = (X_i - \mu)/\sigma$ and $S_n = \sqrt{n}\bar{Y}_n$. Let

$$G_n(s) = P(S_n \leq s).$$

CLT says that for any real s, $G_n(s) \to \Phi(s)$ as $n \to \infty$ where Φ is the standard normal c.d.f. Since $\Phi(s)$ is bounded and continuous, we know that this convergence is uniform, which means

$$\sup_{s \in \mathbb{R}} |G_n(s) - \Phi(s)| \to 0 \text{ as } n \to \infty.$$

This limit realist says nothing about how close the left and right sides must be for any fixed (finite) n.

Theorems discovered independently by Andrew Berry and Carl-Gustav Esseen in the early 1940s do this. In fact each proved a result which is slightly more general than the one that bears their names.

Theorem (Berry-Esseen)

There exists a constant c such that if $Y_1, Y_2, \ldots$ are independent and identically distributed random variables with mean 0 and variance 1, then

$$\sup_{s \in \mathbb{R}} |G_n(s) - \Phi(s)| \le \frac{cE|Y_1^3|}{\sqrt{n}}$$

for all n, where $G_n(s)$ is the cumulative distribution function of $\sqrt{n}\bar{Y}_n$.

Notice that this is uninterestingly true whenever $E|Y_1^3|$ is infinite. In terms of the original sequence $X_1, X_2, \ldots$ the theorem is sometimes stated by saying that when $\lambda = E|X_1^3| < \infty$,

$$\sup_{s \in \mathbb{R}} |G_n(s) - \Phi(s)| \le \frac{c\lambda}{\sigma^3 \sqrt{n}}.$$

Comment

For a proof, see:

- Stein (1972), Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2, pp. 583-602.
- Ho and Chen (1978). Annals of Probability, pp. 231-249.

Ho and Chen prove using Stein's results that it works for $c = 6.5$, but they don't prove that it is the smallest possible value of c. In fact the smallest possible value is not known. Shevtsova (2011, on arXiv) showed the inequality is valid for $c = 0.4748$. In fact, Esseen himself proved that c cannot be less than 0.4097.

For the sake of simplicity, we may exploit these known results by taking $c = 1/2$ to state with certainty that

$$\sup_{s\in\mathbb{R}} |G_n(s) - \Phi(s)| \leq \frac{E|Y_1^3|}{2\sqrt{n}}.$$

Edgeworth Expansions

As in the last section, define $Y_i = (X_i - \mu)/\sigma$ and $S_n = \sqrt{n}\bar{Y}_n$. Furthermore, let

$$\gamma = EY_i^3 \quad \text{and} \quad \tau = EY_i^4$$

and suppose that $\tau < \infty$. The CLT says that for every real y,

$$P(S_n \leq y) = \Phi(y) + o(1)$$

as $n \to \infty$. But we want a better approximation to $P(S_n \leq y)$ than $\Phi(y)$, we we begin by constructing the characteristic function of S_n:

$$\psi_{S_n}(t) = E\exp\left\{(\mathrm{i}t/\sqrt{n})\sum_{i=1}^{n} Y_i\right\} = [\psi_Y(t/\sqrt{n})]^n, \tag{1}$$

where $\psi_Y(t) = E\exp\{\mathrm{i}tY\}$ is the characteristic function of Y_i.

Before proceeding with an examination of the characteristic function of S_n, we first establish four preliminary facts:

1. **Sharpening a well-known limit:** We already know that $(1 + a/n)^n \to e^a$. But how good is this approximation? The binomial theorem shows (after a lot of algebra) that for a fixed non-negative integer k,
$$\left(1+\frac{a}{n}\right)^{n-k} = e^a\left(1-\frac{a(a+2k)}{2n}\right)+o\left(\frac{1}{n}\right) \tag{2}$$
as $n \to \infty$.
2. **Hermite polynomials:** If $\phi(x)$ denotes the standard normal density function, then we define the Hermite polynomials $H_k(x)$ by the equation
$$(-1)^k\frac{d^k}{dx^k}\phi(x) = H_k(x)\phi(x). \tag{3}$$
Thus, by simply differentiating $\phi(x)$ repeatedly, we may verify that $H_1(x) = x$, $H_2(x) = x^2 - 1$, $H_3(x) = x^3 - 3x$, and so on. By differentiating the definition of Hermite polynomials (the displayed equation above) itself, we obtain the recursive formula
$$\frac{d}{dx}[H_k(x)\phi(x)] = -H_{k+1}(x)\phi(x). \tag{4}$$

3. **An inversion formula for characteristic functions:** Suppose $Z \sim G(z)$ and $\psi_Z(t)$ denotes the characteristic function of Z. If $\int_{-\infty}^{\infty} |\psi_Z(t)| dt < \infty$, then $g(z) = G'(z)$ exists and

$$g(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\mathrm{i}tz} \psi_Z(t) dt. \tag{5}$$

   You can find a proof in most theoretical probability books.

4. **An identity involving $\phi(x)$:** For any positive integer k,

$$\begin{aligned}
\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\mathrm{i}tx} e^{-t^2/2} (it)^k dt &= \frac{(-1)^k}{2\pi} \frac{d^k}{dx^k} \int_{-\infty}^{\infty} e^{-\mathrm{i}tx} e^{-t^2/2} dt && (6)\\
&= (-1)^k \frac{d^k}{dx^k} \phi(x) && (7)\\
&= H_k(x) \phi(x), && (8)
\end{aligned}$$

   where (7) follows from (5) since $e^{-t^2/2}$ is the characteristic function for a standard normal distribution and (8) follows from (3).

Now we return to equation (1). We next use a Taylor expansion of $\exp\{\mathrm{i}tY/\sqrt{n}\}$: As $n \to \infty$,

$$\begin{aligned} \psi_Y\left(\frac{t}{\sqrt{n}}\right) &= E\left\{1 + \frac{itY}{\sqrt{n}} + \frac{(it)^2Y^2}{2n} + \frac{(it)^3Y^3}{6n\sqrt{n}} + \frac{(it)^4Y^4}{24n^2}\right\} + o\left(\frac{1}{n^2}\right) \\ &= \left(1 - \frac{t^2}{2n}\right) + \frac{(it)^3\gamma}{6n\sqrt{n}} + \frac{(it)^4\tau}{24n^2} + o\left(\frac{1}{n^2}\right). \end{aligned}$$

If we raise this tetranomial to the nth power, most terms are $o(1/n)$:

$$\begin{aligned} \left[\psi_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n = \Big[&\left(1 - \frac{t^2}{2n}\right)^n + \left(1 - \frac{t^2}{2n}\right)^{n-1}\left(\frac{(it)^3\gamma}{6\sqrt{n}} + \frac{(it)^4\tau}{24n}\right) \\ &+ \left(1 - \frac{t^2}{2n}\right)^{n-2}\frac{(n-1)(it)^6\gamma^2}{72n^2}\Big] + o\left(\frac{1}{n}\right). \end{aligned} \tag{9}$$

By equations (2) and (9), we conclude that

$$\begin{aligned} \psi_{S_n}(t) &= e^{-t^2/2}\left[1 - \frac{t^4}{8n} + \frac{(it)^3\gamma}{6\sqrt{n}} + \frac{(it)^4\tau}{24n} + \frac{(it)^6\gamma^2}{72n}\right] + o\left(\frac{1}{n}\right) \\ &= e^{-t^2/2}\left[1 + \frac{(it)^3\gamma}{6\sqrt{n}} + \frac{(it)^4(\tau-3)}{24n} + \frac{(it)^6\gamma^2}{72n}\right] + o\left(\frac{1}{n}\right). \end{aligned} \tag{10}$$

If we apply these three approximations to (9), we obtain

$$\left[\psi_X\left(\frac{t}{\sqrt{n}}\right)\right]^n = e^{-t^2/2}\left[1-\frac{t^4}{8n}+\frac{(it)^3\gamma}{6\sqrt{n}}+\frac{(it)^4\tau}{24n}+\frac{(it)^6\gamma^2}{72n}\right]+o\left(\frac{1}{n}\right)$$
$$= e^{-t^2/2}\left[1+\frac{(it)^3\gamma}{6\sqrt{n}}+\frac{(it)^4(\tau-3)}{24n}+\frac{(it)^6\gamma^2}{72n}\right]+o\left(\frac{1}{n}\right).$$

Putting (10) together with (5), we obtain the following density function as an approximation to the distribution of S_n:

$$g(y)=\frac{1}{2\pi}\left(\int_{-\infty}^{\infty}e^{ity}e^{-t^2/2}dt+\frac{\gamma}{6\sqrt{n}}\int_{-\infty}^{\infty}e^{-ity}e^{-t^2/2}(it)^3dt\right.$$
$$\left.+\frac{\tau-3}{24n}\int_{-\infty}^{\infty}e^{itx}e^{-t^2/2}(it)^4dt+\frac{\gamma^2}{72n}\int_{-\infty}^{\infty}e^{ity}e^{-t^2/2}(it)^6dt\right). \qquad (11)$$

Next, combine (11) with (8) to yield

$$g(y)=\phi(y)\left(1+\frac{\gamma H_3(y)}{6\sqrt{n}}+\frac{(\tau-3)H_4(y)}{24n}+\frac{\gamma^2H_6(y)}{72n}\right). \qquad (12)$$

By (4), the antiderivative of $g(y)$ equals

$$
\begin{aligned}
G(y) &= \Phi(y) - \phi(y)\left(\frac{\gamma H_2(y)}{6\sqrt{n}} + \frac{(\tau-3)H_3(y)}{24n} + \frac{\gamma^2 H_5(y)}{72n}\right) \\
&= \Phi(y) - \phi(y)\left(\frac{\gamma(y^2-1)}{6\sqrt{n}} + \frac{(\tau-3)(y^3-3y)}{24n} + \frac{\gamma^2(y^5-10y^3+15y)}{72n}\right).
\end{aligned}
$$

The expression above is called the second-order Edgeworth expansion. By carrying out the expansion in (10) to more terms, we may obtain higher-order Edgeworth expansions. On the other hand, the first-order Edgeworth expansion is

$$
G(y) = \Phi(y) - \phi(y)\left(\frac{\gamma(y^2-1)}{6\sqrt{n}}\right). \tag{13}
$$

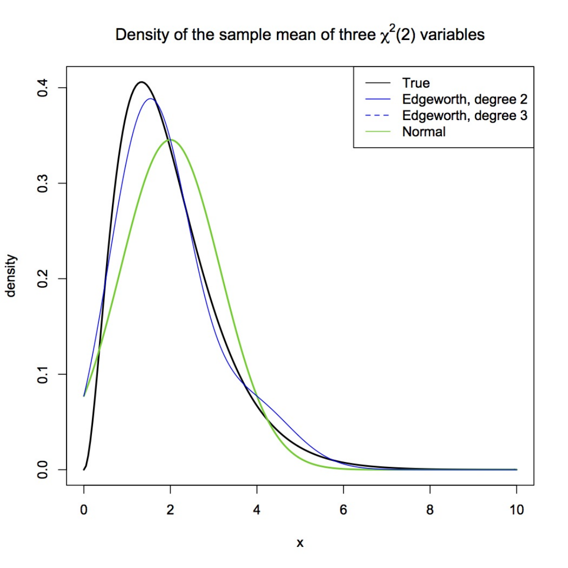
.

Thus if the distribution of Y is symmetric, we obtain $\gamma = 0$ and therefore in this case, the usual (zero-order) central-limit theorem given by $\Phi(y)$ is already first-order accurate. In fact, a better way to put it is: if X_1 has a symmetric distribution with $EX_1^4 < \infty$ and X_1 satisfies Cramer's condition,

$$\lim \sup_{|t| \to \infty} |E \exp(itX_1)| < 1$$

then the rate of normal approximation becomes $O(n^{-1})$.
(Incidentally, the second-order Edgeworth expansion explains why the standard definition of kurtosis of a distribution with mean 0 and variance 1 is the unusual-looking $\tau - 3$.)

Edgeworth Expansion Illustration



Motivating Saddlepoint Approximations

- Approximating probability densities
- Approximating tail-areas of distributions
- $\Longrightarrow$ more accurate statistical inference

Basic idea:

1. Write the quantity you want to approximate as an integral.
2. Expand the integral w.r.t. the dummy variable of integration.
3. Keep the first few terms and integrate.

Note: The integral can be over the complex plane, corresponding to the inversion formula of a Fourier transform.

Background

One of the simplest ways to approximate a positive function $f(x)$: use first few terms of its Taylor series expansion:

- slight modification: let $h(x) \equiv \log f(x)$; write $f(x) = \exp h(x)$ and choose x_0 as the point of interest:

$$f(x) \approx \exp\Big\{h(x_0) + (x - x_0)h'(x_0) + \frac{(x - x_0)^2}{2}h''(x_0)\Big\}. \tag{14}$$

- Choose $x_0 = \hat{x}$ where $h'(\hat{x}) = 0$, yielding

$$f(x) \approx \exp\left\{h(\hat{x}) + \frac{(x - \hat{x})^2}{2}h''(\hat{x})\right\}. \tag{15}$$

- (15) is exact if $h(x)$ is quadratic; if not, obviously when x is far from $\hat{x}$, the higher-order terms will be important $\Rightarrow$ poor approximation

Using (15) for integrals

Consider the integral of a positive function, such as $\int f(x)dx$.

- Expand the integrand as in (15), yielding

$$\int f(x)dx \approx \int \exp\left\{h(\hat{x}) + \frac{(x-\hat{x})^2}{2}h''(\hat{x})\right\}dx. \tag{16}$$

- If $\hat{x}$ is maximum, h'' is negative and RHS of (17) can be computed explicitly: kernel of the integral is same as kernel of normal density with mean $\hat{x}$ and variance $-1/h''(\hat{x})$
- hence,

$$\int f(x)dx \approx \exp\{h(\hat{x})\}\left(-\frac{2\pi}{h''(\hat{x})}\right)^{1/2}. \tag{17}$$

- This is the **Laplace approximation**; implicitly requires integral over the whole line, but accurate provided the 'mass' of the approximating function is within integration limits

Next step: combine (15) and (17)–approximate a function by a Laplace-type approximation of an integral.

- Write f as

$$f(x) = \int m(x,t)dt, \tag{18}$$

  for some positive $m(x,t)$; always possible, e.g. $m(x,t) = f(x)m_0(t)$, where $m_0(t)$ integrates to one
- define $k(x,t) = \log m(x,t)$ we consider Laplace approximation of the integral of $\exp k(x,t)$ w.r.t. t; for fixed x write

$$f(x) \approx \int \exp\left\{k(x,\hat{t}(x)) + \frac{(t-\hat{t}(x))^2}{2}\frac{\partial^2 k(x,t)}{\partial t^2}\bigg|_{\hat{t}(x)}\right\}dt \tag{19}$$

$$= \exp\{k(x,\hat{t}(x))\}\left(-\frac{2\pi}{\frac{\partial^2 k(x,t)}{\partial t^2}\Big|_{\hat{t}(x)}}\right)^{1/2} \tag{20}$$

  where for each x, $\hat{t}(x)$ satisfies $\partial k(x,t)/\partial t = 0$ and $\partial^2 k(x,t)/\partial t^2 < 0$ (maximizes $k(x,t)$)

Observations

- the maximum $\hat{t}$ depends on x; thus (19) is a set of integrated Taylor expansions, one for each x, in contrast to (15) which is a single series around $\hat{x}$
- we are continually re-centering the approximation; thus we should hope it is more accurate than (15)
- The Cost: to get values of $f(x)$ at various points x, we must compute $\hat{t}(x)$, $k(x, \hat{t}(x))$ and $\partial^2 k(x,t)/\partial t^2$ each time
- clearly if $k(x,t)$ is quadratic, (19) is exact; also exact in other cases
- accuracy depends on value of x where we are approximating $f(x)$ since third- and higher-order derivatives w.r.t. t depend on x

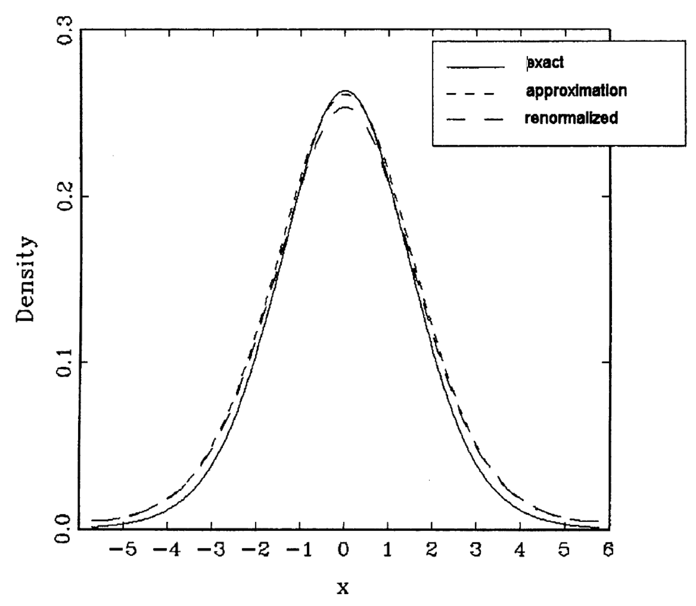
Example: Student's t distribution

Let $X \sim t_\nu$,

$$f(x) = C_\nu \frac{1}{(1 + x^2/\nu)^{\frac{\nu+1}{2}}}, \text{ where } C_\nu = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}}.$$

- let's approximate the sum $T = X_1 + X_2$, of independent t_ν random variables
- exact density (found by Roger Berger using Mathematica) for $\nu = 9$,

$$f_T(x) = \frac{C_\nu^2 5\pi\sqrt{9}}{64} \times \frac{622336 + 56576\frac{x^2}{9} + 48496\frac{x^4}{9^2} + 272\frac{x^6}{9^3} + 7\frac{x^8}{9^4}}{(x^2/9 + 4)^9}$$


exact
approximation
renormalized
Density
0.0
0.1
0.2
0.3
−5
−4
−3
−2
−1
0
1
2
3
4
5
6
x

Approximating Marginals

One important application, where a function is naturally represented by an integral, is the calculation of marginal densities.

- if $(X, Y) \sim f(x, y)$, direct application of (19), we find the marginal density of X, $f_X(x)$ is approximated by

$$\begin{aligned} f_X(x) &= \int f(x,y)dy \\ &\approx \sqrt{2\pi} f(x,\hat{y}) \left(\left. \frac{\partial^2 \log f(x,y)}{\partial y^2} \right|_{\hat{y}} \right)^{-1/2} \end{aligned} \tag{21}$$

  where $\hat{y} = \hat{y}(x)$ satisfies $\partial \log f(x,y)/\partial y = 0$ and is a local maximum
- could repeat approximation to marginalize higher dimensional densities

Integrated Likelihood

Given a likelihood function $L(\theta, \lambda|x)$ where θ is interest parameter, λ is nuisance parameter, x is the data:

- an integrated likelihood L_I for θ obtained by integrating out λ, i.e. the approximation

$$\begin{aligned} L_I(\theta|x) &= \int L(\theta, \lambda|x) d\lambda \\ &\approx L(\theta, \hat{\lambda}|x) \left(\frac{\partial^2 \log L(\theta, \lambda|x)}{\partial \lambda^2} \Big|_{\hat{\lambda}} \right)^{-1/2} . \end{aligned} \tag{22}$$

- this is an *approximate conditional likelihood* of Cox and Reid and a form of the *modified profile likelihood* of Barndorff-Nielsen

Comment

The above examples are artificial, as exact methods exist. We now examine the method more formally and consider important statistical applications.

The Inversion Formula

Recall for a density $f(x)$, the moment generating function (mgf) is defined as

$$\phi_X(t) = \int_{-\infty}^{+\infty} \exp(tx) f(x) dx, \tag{23}$$

provided the integral is finite for t in some open neighborhood of zero

- from $\phi_X(t)$, $f(x)$ may be obtained using the inversion formula

$$\begin{aligned} f(x) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \phi_X(it) \exp(-itx) dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \exp\{K_X(it) - itx\} dt \end{aligned} \tag{24}$$

  where $K_X(t) = \log \phi_X(t)$
- common in statistics where $f(x)$ is density and $\phi_X(it)$ is characteristic function

The function $K_X(t) = \log \phi_X(t)$ is called the **cumulant generating function (cgf)** of X.

- mathematically, cgf and mgf are equivalent, but cgf generates mean and variance instead of uncentered moments—statistically more attractive; think of K'' as variance
- However, we don't have to think of random variables at all–this is valid even if $f(x)$ is negative or does not integrate to one
- notice that this looks similar to the ideas already introduced, suggesting we can use those methods here

Make a change of variable $t' = it$, and re-write as

$$f(x) = \frac{1}{2\pi i} \int_{\tau - i\infty}^{\tau + i\infty} \exp\{K_X(t) - tx\} dt \tag{25}$$

for τ in neighborhood of zero.

- theorem from complex analysis (closed curve theorem) implies that (25) is the same over all paths that are parallel to the imaginary axis in a neighborhood of zero where $\phi_X(t)$ exists; thus we may choose a value of τ over which to do the integration

Take $k(x,t) = K_X(t) - tx$ and find the point $\hat{t}(x)$ that satisfies

$$K'_X(t) = x. \tag{26}$$

- Expand the exponent in (25) around $\hat{t}(x)$, we have

$$K_X(t) - tx \approx K_X(\hat{t}(x)) - \hat{t}(x)x + \frac{(t - \hat{t}(x))^2}{2} K''_X(\hat{t}(x)). \tag{27}$$

- Now substitute in (25) and integrate w.r.t. t along the line parallel to the imaginary axis through the point $\hat{t}(x)$, i.e. choose the point τ in the limits of the integral to be $\hat{t}(x)$ (formal treatment of this maneuver is quite involved).
- Proceeding informally (as if it were a real integral), we see there is the kernel of a normal density.

Similar to (6), we obtain

$$f_X(x) \approx \left(\frac{1}{2\pi K_X''(\hat{t}(x))}\right)^{1/2} \exp\{K_X(\hat{t}(x)) - \hat{t}(x)x\}. \tag{28}$$

- Viewed as a point in the complex plane, $\hat{t}(x)$ is neither a maximum nor minimum, but a **saddlepoint** of $K_x(t) - tx$, as the function is constant in the imaginary direction and has an extrema in the real direction.
- Also, choosing $\tau = \hat{t}(x)$ in (25) is an application of the method called steepest descent; this takes advantage of the fact that since $\hat{t}(x)$ is an extreme point, the function is falling away rapidly as we move from the point.
- Thus the influence on the integral of neighboring points is diminished, making (28) seem more reasonable.

Comment

(28) is what is usually thought of as the saddlepoint approximation to a density.

$$f_X(x) \approx \left(\frac{1}{2\pi K_X''(\hat{t}(x))} \right)^{1/2} \exp\{K_X(\hat{t}(x)) - \hat{t}(x)x\}.$$

- its error of approximation is much better than Taylor series approximation
- in classical 'first-order asymptotics', the error terms decrease at rate $n^{1/2}$, for sample of size n
- the saddlepoint is a 'second-order asymptotics' approach; error terms can decrease as fast as $n^{3/2}$, yielding huge improvement for small samples

Example: Non-central chi-squared

The non-central chi-squared density has no closed form, and is usually written

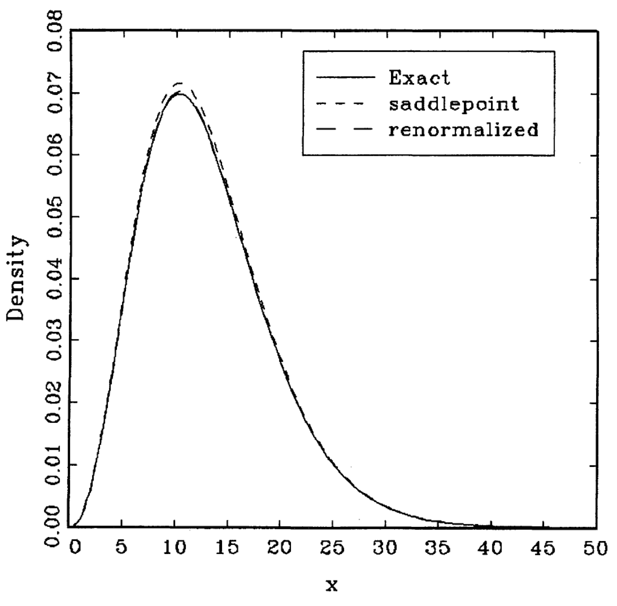
$$f(x|\lambda) = \sum_{k=0}^{\infty} \frac{x^{p/2+k-1} e^{-x/2}}{\Gamma(p/2+k) 2^{p/2+k}} \frac{\lambda^k e^{-\lambda}}{k!},$$

where p is d.o.f., λ noncentrality parameter.

- this density is an infinite mixture of central chi-squared densities, where weights are Poisson probabilities
- can actually find the mgf:

$$\phi_X(t) = \frac{e^{2\lambda t/(1-2t)}}{(1-2t)^{p/2}}.$$

- solving the saddlepoint equation $\partial \log \phi_X(t)/\partial t = x$ yields the saddlepoint; we plot for $p = 7$ and $\lambda = 5$



Exact
saddlepoint
renormalized
Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0
5
10
15
20
25
30
35
40
45
50
x

Application: Objective Bayes

Saddlepoint approximation to conditional distribution of adjusted signed root likelihood ratio statistic (Barndorff-Nielsen)

$$r^*(\psi) = r(\psi) + r(\psi)^{-1} \log\{u_F(\psi)/r(\psi)\}$$

where u_F is function of various likelihood quantities.

- standard normal approximation to the conditional distribution of above statistic has error of order $O(n^{-3/2})$
- Laplace approximation (Tierney-Kass-Kadane) for posterior distribution, with corresponding u_B has similar result
- identify PMPs by choosing prior such that $u_B = u_F + O_p(n^{-3/2})$ yielding higher-order matching

More on this theme later, time permitting!

Alternative Derivation

Original saddlepoint derivation of Daniels (1954) based on inversion of characteristic function.

- Alternative derivation, ‘more statistical’, described by Reid and others, allows us to make more precise what is the ‘order’ of the approximation.
- Recall an Edgeworth expansion of a distribution: expand the cumulant generating function in a Taylor series around zero, then apply inverse Fourier transform.

Let $X_1, \ldots, X_n$ i.i.d. with density f, mean μ, variance σ^2.

- useful form of Edgeworth expansion (see Hall 1992)

$$P\left(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \leq w\right) = \Phi(w) + \phi(w)\left[\frac{-1}{6\sqrt{n}}\kappa(w^2-1) + O(\frac{1}{n})\right], \quad (29)$$

  where Φ, ϕ are cdf, pdf of std. normal, $\kappa = E(X_1-\mu)^3$ is skewness.
- For Edgeworth expansion, the $O(\frac{1}{n})$ term is of the form $p(w)/n$, where p is a polynomial.
- Since it is multiplied by $\phi(w)$, then the derivatives (in w) maintain same order of approximation.
- Thus (29) may be differentiated to obtain density approximation with same order of accuracy.

Proceeding with this, make the transformation $x = \sigma w + \mu$, obtain the approximation to the density of $\bar{X}$ as

$$f_{\bar{X}}(x) = \frac{\sqrt{n}}{\sigma}\phi\left(\frac{x-\mu}{\sigma/\sqrt{n}}\right) \times \left[1 + \frac{\kappa}{6\sqrt{n}}\left\{\left(\frac{x-\mu}{\sigma/\sqrt{n}}\right)^3 - 3\left(\frac{x-\mu}{\sigma/\sqrt{n}}\right)\right\} + O(\frac{1}{n})\right].$$

- Ignoring the term in braces produces the usual normal approximation, which is accurate to $O(\frac{1}{\sqrt{n}})$.
- If we use (29) for value of x near μ, then the value of the expression in braces is close to zero, and the approximation will be accurate to $O(\frac{1}{n})$.
- The trick of the saddlepoint approximation is to make this always be the case.

References

- Barndorff-Nielsen, O. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika* **73** (2), 307-322.
- Butler, R.W. (2007). *Saddlepoint Approximations with Applications*. Cambridge.
- Goutis, C. and G. Casella (1999). Explaining the Saddlepoint Approximation. *The American Statistician*, **53** (3), 216-224.
- Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, Spring Series in Statistics.
- Jensen, J.L. (1986). *Saddlepoint Approximations*. Oxford University Press.
- Kolassa, J.E. (2006). *Series Approximation Methods in Statistics*.
- Pace, L. and A. Salvan (1997). *Principles of Statistical Inference From a Neo-Fisherian Perspective*.
- Severini, T.A. (2000) *Likelihood Methods in Statistics*
- Strawderman, R.L. (2000). Higher-Order Asymptotic Approximation: Laplace, Saddlepoint, and Related Methods. *JASA*, 95:452, 1358-1364.

Time Out

Cupuaçu and Octavia

Comparative Asymptotics: Probability Matching Priors

Motivations

1. Philosophical: unification of Bayesian and frequentist methods
2. Bayesians seek to appear 'objective'
3. Most users of Bayesian statistics are not statisticians; in need of a guide (e.g. Rubin 1984)
4. Alternative route to frequentist asymptotics

Examples: Noninformative Priors

- Jeffreys'-rule prior (Jeffreys 1961): **invariant under reparameterization of the parameter θ**

$$\pi(\theta) \propto \sqrt{\mathbf{det}\, I(\theta)}$$

- Reference priors (Bernardo 1979): **maximize Kullback-Liebler distance between prior and posterior**

- Probability matching priors (**PMPs**; Datta and Mukerjee 2004): **deliver posterior credible sets with approximately correct frequentist probability interpretation**

The inferential problem

Let $Y = \{Y_1, \ldots, Y_n\}$ be random sample from underlying distribution $F(y; \theta)$, indexed by d-dimensional parameter $\theta = (\psi, \phi)$,

- ψ scalar parameter of interest
- ϕ nuisance

Inference, notation

Let $\ell(\theta) \equiv \ell(\theta; Y)$ be log-likelihood, $\hat{\theta} = (\hat{\psi}, \hat{\phi})$ the overall MLE of θ, $\hat{\phi}_\psi$ the constrained MLE of ϕ, for fixed value of ψ.

Likelihood ratio statistic is $w(\psi) = 2\{\ell(\hat{\theta}) - \ell(\psi, \hat{\phi}_\psi)\}$.

Let

$$R(\psi) = \mathrm{sgn}(\hat{\psi} - \psi) w(\psi)^{1/2}.$$

This is the signed root likelihood ratio statistic.

First-order theory

Have $w(\psi)$ distributed as χ^2_1 to error of order $O(n^{-1})$.

Also, $R(\psi)$ distributed as $N(0,1)$, to error of order $O(n^{-1/2})$.

In small samples, need higher-order accuracy.

Frequentist Inference

For a standard normal pivot, such as $R(\psi)$, a confidence set of asymptotic coverage $1-\alpha$ for ψ is

$$\mathcal{I}(Y) \equiv \mathcal{I}_{1-\alpha}(Y) = \{\psi : \ u(Y, \psi) \leq 1-\alpha\},$$

with $u(Y, \psi) = \Phi\{R(\psi)\}$, in terms of the $N(0,1)$ distribution function $\Phi(\cdot)$.

Equivalently, the confidence set is

$$\mathcal{I}(Y) = \{\psi : \ R(\psi) \leq \Phi^{-1}(1-\alpha)\}.$$

The coverage error of the confidence set is $O(n^{-1/2})$: first-order accuracy.

Background on PMPs

In Bayesian inference on ψ, in absence of subjective prior information about θ, natural to use a prior which leads to posterior probability limits that are also frequentist limits, in the sense that

$$Pr_\theta\{\psi \leq \psi^{(1-\alpha)}(\pi, Y)\} = 1 - \alpha + O(n^{-r/2}),$$

for $r = 2$ or 3, each $0 < \alpha < 1$.

Here:

- n is sample size;
- $\psi^{(1-\alpha)}(\pi, Y)$ is $(1-\alpha)$ quantile of marginal posterior, given data Y, of ψ, under prior $\pi(\psi, \phi)$;
- Pr_θ denotes frequentist probability, under repeated sampling of Y, under parameter θ

PMPs

With $r = 2$: first-order probability matching prior.

With $r = 3$: second-order probability matching prior.

- note the convention in this literature is at odds with the usual definition of 'first', 'second' order

Comments

- standard asymptotic theory tells us that all priors are probability matching to order $O(n^{-1/2})$
- Jeffreys' prior, which is the square root of the expected information, is 'first' order probability matching (Welch & Peers 1963)
- Many authors, e.g. Tibshirani (1989):
  - PMPs feasible for scalar interest parameters ☺
  - for vector interest parameters life is tough ☹
- PMPs may be improper (in fact they often are)
- small sample sizes (5, 10, 15, etc.) are of interest

Existing Routes to Identify Matching Priors

Welch and Peers (1963) Find a standard normal pivot which is a function of the prior and which admits an asymptotic expansion

Bickel and Ghosh (1990) Follow 'shrinkage argument' idea of Dawid (1991). Posterior distribution of LR statistic admits posterior Bartlett correction—χ^2 approximation now has error of order $O(n^{-2})$; subsequently, J.K. Ghosh, R. Mukerjee and coauthors 'match' Bayesian and (unconditional) frequentist Bartlett correction terms

Shinkage Argument Basics

Suppose it is intended to find expression for $E_\theta\{h(Y,\theta)\}$ where h is measurable function, with finite expectation (in particular, h may be indicator function so that $E_\theta\{h(Y,\theta)\}$ is a probability)

Bayesian approach to evaluate $E_\theta\{h(Y,\theta)\}$:

1. Obtain posterior density of θ under prior $\xi(\cdot)$ for θ. Hence obtain $E^\xi\{h(Y,\theta)|Y\}$, which is expectation of $h(Y,\theta)$ in posterior setup.
2. Find $E_\theta E^\xi\{h(Y,\theta)|Y\}$ ($=\lambda(\theta)$, say)
3. Integrate $\lambda(\cdot)$ w.r.t. $\xi(\cdot)$ and allow $\xi(\cdot)$ to converge weakly to degenerate prior at θ. This yields $E_\theta\{h(Y,\theta)\}$.

End result: Magic! Directly calculate frequentist coverage of Bayesian quantiles.

Some theory, (unconditional) PMPs

Write $\theta = (\theta^1, \ldots, \theta^d)$, with $\theta^1 = \psi$. Write $D_j = \partial/\partial\theta^j$, let $I = [I_{ij}]$ be Fisher information matrix, with inverse $I^{-1} = [I^{ij}]$.

Further, let $L_{jrs} = E_\theta(V_{jrs})$, with $V_{jrs} = D_j D_r D_s \log f(Y_1; \theta)$,
$\tau^{jr} = I^{j1} I^{r1} / I^{11}$, $\sigma^{jr} = I^{jr} - \tau^{jr}$.

Adopt summation convention, with summation of an expression being implied, over all repeated superscripts or subscripts.

Key Results

Result 1

A prior $\pi(\cdot)$ is first-order probability matching iff

$$D_j\{\pi(\theta)I^{j1}(I^{11})^{-1/2}\} = 0. \tag{D1}$$

Result 2

A prior $\pi(\cdot)$ is second-order probability matching iff, in addition,

$$\frac{1}{3}D_u\{\pi(\theta)\tau^{jr}L_{jrs}(3\sigma^{su} + \tau^{su})\} - D_jD_r\{\pi(\theta)\tau^{jr}\} = 0. \tag{D2}$$

Levine & Casella (2003) and Sweeting (2005): numerical solution of these partial differential equations; identification of matching priors via these conditions not practical for everyone

DKY, Bka, 2012

Intuition of these conditions arises from alternative route.

Let $\mu_F \equiv \mu_F(\theta)$ be frequentist mean of $R(\psi)$, $\mu_B \equiv \mu_B(Y)$ be Bayesian posterior mean.

Let σ_F^2 be frequentist variance of $R(\psi) - \mu_B$, σ_B^2 be posterior variance of $R(\psi) - \mu_B$. [Note, very specific mean adjustment.]

Key Results

(1) If the prior $\pi(\theta)$ is specified so that the frequentist and Bayesian means match to $O_p(n^{-1})$, $\mu_B = \mu_F + O_p(n^{-1})$, then (D1) is satisfied, and the prior is first-order probability matching.

(2) If in addition, prior $\pi(\theta)$ is specified so that $\sigma_B^2 = \sigma_F^2 + O_p(n^{-3/2})$, (D2) is satisfied, and the prior is second-order probability matching.

Roughly, second-order probability matching arises from matching of frequentist and Bayesian variances of a very particular mean-adjusted version of $R(\psi)$.

Idea of Proof: Similar to conditional setting (coming soon to a theatre near you)

Enter Conditional Inference

Definition: A statistic $T = T(Y)$ is *sufficient* for θ if the distribution of Y, conditional on $T(Y) = t$, is independent of θ (T is *min. sufficient* if a function of every other sufficient statistic).

Definition: Suppose the minimal sufficient statistic T is partitioned as $T = (S, C)$ where:

(a) the distribution of C depends on λ but not on ψ

(b) the conditional distribution of S given $C = c$ depends on ψ, but not on λ, for each c

Then C is an **ancillary statistic**, and S is conditionally sufficient for ψ given C.

Conditionality Principle: Inference about ψ should be based on the conditional distribution of S given C.

In cases where MLE is not sufficient, it is often the case that $T = (\hat{\theta}, C)$ and the log-likelihood may be written as $\ell(\theta; \hat{\theta}, C)$.

Differences Between Conditional and Unconditional (Frequentist) Inference
Central Concepts

Unconditional

- orthodox frequentist view
- focus is on how the procedure will perform in repeated trials
- statements about unknown parameters are made pre-data

Conditional

- Fisherian view
- statements about unknown parameters are connected to the data actually observed
- Motivations: relevance, sufficiency

- Bayesian: procedures based on the observed data only; do not account for other experiments that *might* have been conducted but were not
- i.e. no need to specify ancillaries in Bayesian setting

Conditional Inference

In key contexts, appropriate frequentist inference is conditional, conditioning on observed data value c for some statistic C.

- Inference on natural parameters in full multi-parameter exponential family (achieve relevance),

$$f(x;\theta) = b(\theta)h(x)\exp\left\{\sum_{i=1}^{k}\eta_i(\theta)\tau_i(x)\right\}, \;\; x \in \mathcal{X}, \;\; \theta \in \Theta$$

  when $\dim(\theta) = k$ (Examples: exponential, Beta, normal)
- Inference on natural parameters in curved multi-parameter exponential family (MLE not sufficient), $\dim(\theta) = d < k$ (Examples: $N(\mu, \mu^2)$, Fisher's hyperbola, exponential regression)
- Group transformation models (maximal invariant is ancillary statistic). (Example: location-scale)

Conditional PMPs

Appropriate frequentist inference to match is the conditional one.

The requirement should be 'conditional probability matching':

$$Pr_{\theta}\{\psi \leq \psi^{(1-\alpha)}(\pi, Y) \mid C(Y) = c\} = 1 - \alpha + O(n^{-r/2}).$$

Routes to Identify Conditional PMPs

1. Conditional version of shrinkage argument
2. Mean- and variance-adjusted signed root
3. ‘Saddlepoint’ method
4. p^* (DKY, 2016)

DKY 2k12

Theorem: Suppose the prior is such that $E_{Y|C}(\mu_B) = \mathring{\mu}_F + O(n^{-3/2})$. If the prior is also such that $\sigma_B^2 = \mathring{\sigma}_F^2 + O(n^{-3/2})$, then the Bayesian quantile is second-order conditional probability matching.

How we got there: Derive expansions for conditional frequentist mean, mean-adjusted conditional frequentist variance, posterior mean and mean-adjusted posterior variance of signed root statistic. Meditate, drink plenty of water and do a lot of algebra.

Idea of Proof: Given Y and unimodal log-likelihood, $R(\psi)$ is monotonic decreasing in ψ. Repeatedly apply delta method of Hall (1992) and manipulate higher-order relationships discovered in the expansions.

The Ugly Version

In notation of paper, let $\lambda_r = E\{\ell_r(\theta)\}$, $\lambda_{jr} = E\{\ell_{jr}(\theta)\} = I_{jr}$, etc. and let $\eta = (-\lambda^{11})^{-1/2}$; then these conditions correspond (respectively) to

$$\frac{\partial \log \pi(\theta)}{\partial \theta^r} \mathring{\lambda}^{r1} = \mathring{\lambda}_{rs/t} \mathring{\lambda}^{r1} \mathring{\lambda}^{st} + \frac{1}{2} \mathring{\eta}^2 \mathring{\lambda}_{rs/t} \mathring{\lambda}^{r1} \mathring{\lambda}^{s1} \mathring{\lambda}^{t1}$$

and

$$\eta \frac{\partial \log \pi(\theta)}{\partial \theta^r} \lambda^{r1} = -\sum_r \frac{\partial(\eta \lambda^{r1})}{\partial \theta^r},$$

$$\sum_r \frac{\partial \{\pi(\nu^{rs} + \frac{1}{3}\tau^{rs})\tau^{tu}\lambda_{stu}\}}{\partial \theta^r} + \sum_{r,s} \frac{\partial^2(\pi \tau^{rs})}{\partial \theta^r \partial \theta^s} = 0.$$

Relationship Between Conditional/Unconditional PMPs

Result 1:

If the prior is second-order probability matching in unconditional frequentist sense, it is second-order conditional probability matching provided a simple further condition ('conditional version of (D1)$'$) is satisfied.

This extra condition satisfied provided

$$E_{Y|C}(\mu_B) = \mathring{\mu}_F + O(n^{-3/2}),$$

$\mathring{\mu}_F$ is conditional frequentist mean of $R(\psi)$.

Relationship Between Conditional/Unconditional PMPs

Result 2: (to be submitted)

Think of statistical model as smooth manifold whose points are probability measures defined on common probability space. Fisher information is a natural Riemannian metric.

Consider a one-parameter model. Efron's (1975) curvature, $\gamma^2(\theta)$ is defined as

$$\gamma^2(\theta) = \frac{i_{02} - i_{20}^2 - \frac{i_{11}^2}{i_{20}}}{i_{20}^2}$$

where all quantities are expectations of log-likelihood derivatives, e.g. $i_{20} = E\left(\frac{\partial \ell}{\partial \theta}\frac{\partial \ell}{\partial \theta}\right)$ is the Fisher information component at the true value for a single observation.

It can be shown that Jeffreys' prior (a function of this information component), which is unconditional probability matching to 1st order, will also be conditional probability matching to 1st order if Efron curvature is constant.

Saddlepoint Method

Barndorff-Nielsen (1986) showed (by saddlepoint approximation) that

$$R^* = R(\psi) + R^{-1}(\psi)\log\left(U_F(\psi)/R(\psi)\right)$$

has a distribution which may be approximated by standard normal with error of order $O(n^{-3/2})$, conditionally on an ancillary statistic C

- the *conditional* frequentist adjustment term is:

$$U_F(\psi) = \frac{\left|\ell_{;\hat{\theta}}\left(\hat{\psi},\hat{\lambda}\right) - \ell_{;\hat{\theta}}\left(\psi,\hat{\lambda}_\psi\right) \quad \ell_{\lambda;\hat{\theta}}\left(\psi,\hat{\lambda}_\psi\right)\right|}{\left\{\left|\ell_{\lambda\lambda}(\psi,\hat{\lambda}_\psi)\right|\left|\ell_{\theta\theta}(\hat{\psi},\hat{\lambda})\right|\right\}^{1/2}}$$

- adjustment term parametrization invariant, does not depend on nuisance parameter
- relative error; holds in small deviation regions
- MLE is $\sqrt{n}$-consistent so in this small-deviation region, absolute error would be $O(n^{-1})$
- $\ell_{\psi\psi}, \ell_{\lambda\lambda}, \ell_{\lambda;\hat{\lambda}}$ and $\ell_{\lambda,\hat{\psi}}$ are second-order partial derivatives of full log-likelihood,
- but $\ell_{\lambda;\hat{\lambda}}$ and $\ell_{\lambda;\hat{\psi}}$ involve partial derivatives of the average log-likelihood with respect to the MLE holding the ancillary statistic fixed–these are sample space derivatives which require explicit specification of ancillary statistic

Bayesian adjustment version

DiCiccio & Martin (1993) showed the corresponding Bayesian version is

$$U_B(\psi) = \ell_\psi(\psi, \hat{\lambda}_\psi) \frac{\left| -\ell_{\lambda\lambda}(\psi, \hat{\lambda}_\psi) \right|^{1/2} \pi(\hat{\theta})}{\left| -\ell_{\theta\theta}(\hat{\theta}) \right|^{1/2} \pi(\psi, \hat{\lambda}_\psi)}$$

Theorem: If a prior is such that $U_F = U_B + O_p(n^{-3/2})$, then the Bayesian posterior credible sets match the conditional frequentist coverage probabilities to order $O(n^{-3/2})$.

Idea of Proof: Derive posterior and frequentist tail-area approximations. For fixed ψ, the event $\psi \leq \psi^{(1-\alpha)}(\pi, Y)$ is equivalent to

$$\Phi(R^*) + O(n^{-3/2}) \geq \alpha$$

Combine these expansions to express the first probability in terms of the second.

Example: Location-Scale Model

Suppose i.i.d. sample $Y = \{Y_1, \ldots, Y_n\}$ from location-scale family

$$\sigma^{-1} f\left(\frac{y-\mu}{\sigma}\right)$$

with $f(\cdot)$ known.

- ancillary statistic is configuration statistic $C = \{C_1, \ldots, C_n\}$ with

$$C_i = \frac{Y_i - \hat{\mu}}{\hat{\sigma}}$$

- take $\psi = \sigma$ as interest parameter and $\lambda = \mu$ as nuisance

continued

Log-likelihood of the form

$$\ell(\theta;\hat{\theta},c) = -n\log\psi - \sum_{i=1}^{n} g\left\{\frac{\hat{\psi}}{\psi}\left(C_i + \frac{\hat{\lambda}-\lambda}{\hat{\psi}}\right)\right\}$$

where $g(\cdot) = -\log f(\cdot)$

- let $\tilde{\delta}_i$ be the term in parentheses evaluated at constrained MLE
-

$$U_F = \frac{\left[-\frac{n}{\hat{\psi}} + \sum \frac{C_i}{\psi} g'(\tilde{\delta}_i)\right]\left[\sum \frac{1}{\psi^2} g''(\tilde{\delta}_i)\right]}{\left(\sum \frac{1}{\psi^2} g''(\tilde{\delta}_i)\right)^{1/2}},$$

$$U_B = \left[-\frac{n}{\psi} + \sum \frac{\hat{\psi}}{\psi^2} C_i g'(\tilde{\delta}_i)\right]\left(\sum \frac{1}{\psi^2} g''(\tilde{\delta}_i)\right)^{1/2} \frac{\pi(\hat{\theta})}{\pi(\psi,\hat{\lambda}_\psi)}$$

continued

Setting the adjustment terms equal, we have that the condition for 2nd order conditional matching is

$$\frac{\pi(\hat{\theta})}{\pi(\psi, \hat{\lambda}_\psi)} \propto \frac{\psi}{\hat{\psi}}$$

In original parameterization,

$$\pi(\theta) \propto \frac{1}{\sigma}$$

- this is (i) right-invariant Haar prior, (ii) Bernardo's reference prior
- by contrast, Jeffreys' prior is $\pi(\theta) \propto \sigma^{-2}$

Example: Cauchy location-scale

Let $Y_1, \dots, Y_n$ be IID from a Cauchy location-scale model, $\theta = (\mu, \sigma)$, common density

$$f(y; \mu, \sigma) = \frac{1}{\sigma} f_0\Big(\frac{y-\mu}{\sigma}\Big),\ f_0(z) = \frac{1}{\pi(1+z^2)},$$

interest parameter is scale parameter, σ.

MC evaluation of conditional coverage

There is an exact formula ('the p^* formula') for the conditional density of $(\hat{\mu}, \hat{\sigma})$, given $C = c$.

Given specified $\theta_0 = (\mu_0, \sigma_0)$, the exact conditional coverage of marginal posterior quantiles constructed by Monte Carlo.

(Objective) Bayes inference

Consider three priors

- $I: \ \pi(\mu, \sigma) \propto \sigma^{-1}$, Bernardo's reference prior (also right-invariant Haar prior and independence Jeffreys' prior)
- $II: \ \pi(\mu, \sigma) \propto \sigma^{-2}$, Jeffreys' prior (left-invariant Haar prior)
- $III: \ \pi(\mu, \sigma) \propto \mu^2 \sigma^{-1}$.

Prior I is exact probability matching, both unconditionally and conditionally.

Prior III is second-order probability matching, from an unconditional perspective: we are not able to assert second-order conditional probability matching.

Conditional coverage

Consider conditional inference, based on two datasets of size $n = 10$, generated from model with $\theta_0 = (0, 1)$.

Dataset A has conditional value c of configuration statistic C with $||c||$ around 85th percentile of the sampling distribution of $||C||$ under the Cauchy model.

Dataset B has $||c||$ around 75th percentile.

Algorithm

- Sample from p^* using random walk Metropolis-Hastings, yielding sequence of $\hat{\theta}^*$ (say 2000 samples after burn-in and taking every 20th point)

- For each $\hat{\theta}^*$, construct $Y = C\hat{\sigma}^* + \hat{\mu}^*$, yielding 2000 data sets Y^* of size 10

- For each Y^*, construct posterior for chosen prior. Use MH to sample from posteriors.

- Calculate $1 - \alpha$ quantiles for posteriors, ask question: *Is this* $1 - \alpha$ *posterior quantile for the interest parameter bigger than its 'true/pilot' value?*

- The number of 'yes' outcomes divided by the number of posteriors (2000) gives the frequency that the Bayesian quantile covers the true parameter value.

	A			B		
(1-α)(%)	I	II	III	I	II	III
1	1.02	0.18	5.34	1.04	0.19	4.06
5	5.09	1.33	18.96	4.98	1.44	11.48
10	10.20	3.10	30.73	9.75	3.43	18.63
90	90.05	75.08	95.71	90.08	78.11	95.22
95	95.01	84.96	97.92	95.03	87.31	97.82
99	99.00	95.69	99.59	99.03	96.50	99.66

Remarks

Of particular interest are results for Prior III, which is second-order probability matching, unconditionally.

Conditional coverage results way off nominal desired levels, for both conditioning values c, but perhaps no worse than the corresponding unconditional figures.

Formal relationship between conditional and unconditional probability matching priors in ancillary statistic context still to be understood: our sufficient condition for second-order conditional probability matching is not satisfied.

Bootstrap Refinements

Consider $\{X_i\}$, $i = 1, \ldots, n$ from CDF F_0.

- T_n is a statistic and let $G_n(\tau, F_0) = P(T_n \leq \tau)$ denote the exact, finite-sample CDF of T_n.
- We know $G_n(\tau, F_0)$ cannot be calculated analytically unless T_n is pivotal.
- Objective: obtain an approximation to $G_n(\tau, F_0)$ when T_n is <u>not</u> pivotal.

We will need to make some assumptions about the form of the function $T_n(X_1, \ldots, X_n) \Rightarrow$ assume T_n is a smooth function of sample moments of X or sample moments of functions of X.

- Specifically,

$$T_n = \sqrt{n}[H(\bar{Z}_1, \ldots, \bar{Z}_J) - H(\mu_{Z_1}, \ldots, \mu_{Z_J})],$$

where the H is scalar-valued, smooth in the sense defined below, and $\bar{Z}_j = n^{-1} \sum_{i=1}^{n} Z_j(X_i)$ for each $j = 1, \ldots, J$ and some non stochastic function Z_j, and $\mu_{Z_j} = E(Z_j)$.

After centering and normalization, many estimators and test statistics fit into this framework, or can be approximated by such functions with an asymptotically negligible approximation error.

- e.g. the OLS estimator in linear regression
- the t-statistic for testing hypotheses about coefficients are exact functions of sample moments
- MLEs for nonlinear models can be approximated with asymptotically negligible error by smooth functions of sample moments if the log-likelihood function or moment conditions have sufficiently many derivatives with respect to the unknown parameters.
- nonparametric density estimators, or quantile estimators such as the least absolute deviations estimator (LAD) of the slope coefficients of a median-regression model do not have sufficiently smooth objective functions, so they do not satisfy the smoothness conditions required here, and there are alternative bootstrap methods for those settings.

Returning to the problem of approximating $G_n(\tau, F_0)$, consider first-order asymptotic theory.

- Write $H(\bar{Z}_1, \ldots, \bar{Z}_J) = H(\bar{Z})$ where $\bar{Z} = (\bar{Z}_1, \ldots, \bar{Z}_J)'$. Define $\mu_Z = E(\bar{Z})$, $\partial H(z) = \partial H(z)/\partial z$ and $\Omega = E[(\bar{Z} - \mu_Z)(\bar{Z} - \mu_Z)']$ whenever these quantities exist.

Assume that

(i) $T_n = \sqrt{n}[H(\bar{Z}) - H(\mu_Z)]$, where $H(z)$ is six times continuously differentiable with respect to any mixture of components of z in a neighborhood of μ_Z.

(ii) $\partial H(\mu_Z) \neq 0$.

(iii) The expected value of the product of any 16 components of Z exists

Comments about these assumptions:

- the proof that the bootstrap provides asymptotic refinements is based on an Edgeworth expansion of a sufficiently high-order Taylor-series approximation to T_n.
- the above assumptions ensure that H has derivatives and Z has moments of sufficiently high order to obtain the Taylor series and Edgeworth expansions that are used to obtain a bootstrap approximation to the distribution of T_n that has error of size $O(n^{-2})$
- these assumptions may not be the weakest condition needed to obtain this result.
- this assumes the existence of more derivatives of H and moments of Z than needed to obtain less accurate approximations!

For example, asymptotic normality of T_n can be proved if H has only one continuous derivative and Z has only two moments.

- Hall (1992) states the regularity conditions needed to obtain various levels of asymptotic and bootstrap approximations.

Now, under these assumptions, a Taylor series approximation gives

$$\sqrt{n}[H(\bar{Z}) - H(\mu_Z)] = \partial H(\mu_Z)' \sqrt{n}(\bar{Z} - \mu_Z) + o_p(1). \tag{30}$$

Applying the Lindeberg-Levy CLT to the RHS shows that $\sqrt{n}[H(\bar{Z}) - H(\mu_Z)] \to_d N(0, V)$ where $V = \partial H(\mu_Z)' \Omega \partial H(\mu_Z)$. Thus the asymptotic CDF of T_n is

$$G_\infty(\tau, F_0) = \Phi(\tau / V^{1/2}),$$

where Φ is the standard normal CDF. This is just the standard result of the delta method. Moreover, it follows from the Berry-Essen theorem that

$$\sup_\tau |G_n(\tau, F_0) - G_\infty(\tau, F_0)| = O(n^{-1/2}).$$

Thus, under the above assumptions of the smooth function model, first-order asymptotic approximations to the exact finite-sample distribution of T_n make an error of size $O(n^{-1/2})$.

Many statistics have asymptotic chi-square distributions.

- Such statistics often satisfy the assumptions of the smooth function model but with

$$\partial H(\mu_Z) = 0$$

  and

$$\partial^2 H(z)/\partial z \partial z' \Big|_{z=\mu_Z} \neq 0.$$

- Versions of the result described here for asymptotically normal statistics *are also available* for asymptotic chi-square statistics.
- First-order asymptotic approximations to the finite-sample distributions of asymptotic chi-square statistics typically make errors of size $O(n^{-1})$.

Now consider the bootstrap.

- The bootstrap approximation to the CDF of T_n is $G_n(\cdot, F_n)$.
- *Under the smooth function model assumptions*, it follows that the bootstrap is consistent.
- It is possible to prove the stronger result that $\sup_\tau |G_n(\tau, F_n) - G_\infty(\tau, F_0)| \to 0$ almost surely.
- This result ensures that the bootstrap provides a good approximation to the asymptotic distribution of T_n if n is sufficiently large.
- However, it says nothing about the accuracy of $G_n(\cdot, F_n)$ as an approximation to the exact finite-sample distribution function $G_n(\cdot, F_0)$.

To investigate this question, we need to develop higher-order asymptotic approximations to $G_n(\cdot, F_0)$ and $G_n(\cdot, F_n)$.

Theorem (Hall 1992)

Let the smooth function model assumptions hold. Also assume that

$$\limsup_{\|t\|\to\infty} |E[\exp(it'Z)]| < 1, \tag{31}$$

where $i = \sqrt{-1}$. *Then*

$$G_n(\tau, F_0) = G_\infty(\tau, F_0) + \frac{1}{\sqrt{n}} g_1(\tau, F_n) + \frac{1}{n} g_2(\tau, F_0) + \frac{1}{n^{3/2}} g_3(\tau, F_0) + O(n^{-2}) \tag{32}$$

uniformly over τ *and*

$$G_n(\tau, F_n) = G_\infty(\tau, F_n) + \frac{1}{\sqrt{n}} g_1(\tau, F_n) + \frac{1}{n} g_2(\tau, F_n) + \frac{1}{n^{3/2}} g_3(\tau, F_n) + O(n^{-2}) \tag{33}$$

uniformly over τ *almost surely. Moreover,* g_1 *and* g_3 *are even, differentiable functions of their first arguments,* g_2 *is an odd, differentiable function of its first argument, and* G_∞, g_1, g_2 *and* g_3 *are continuous functions of their second arguments relative to the supremum norm on the space of distribution functions.*

If T_n is asymptotically pivotal, then G_∞ is the standard normal distribution function.

- Otherwise, $G_\infty(\cdot, F_0)$ is the $N(0, V)$ distribution function and $G_\infty(\cdot, F_n)$ is the $N(0, V_n)$ distribution function where V_n is the quantity obtained from V by replacing population expectations and moments with expectations and moments relative to F_n.
- Condition (31) is Cramer's condition. It is satisfied if the random vector Z has a probability density with respect to Lebesgue measure.

  More generally, it is satisfied if the distribution of Z has a non-degenerate absolutely continuous component in the sense of the Lebesgue decomposition.

Now we can evaluate the accuracy of the bootstrap estimator $G_n(\tau, F_n)$ as an approximation to the exact, finite-sample CDF $G_n(\tau, F_0)$. It follows from (32) and (33) that

$$G_n(\tau, F_n) - G_n(\tau, F_0) = [G_\infty(\tau, F_n) - G_\infty(\tau, F_0)] + \frac{1}{\sqrt{n}}[g_1(\tau, F_n) - g_1(\tau, F_0)] + \frac{1}{n}[g_2(\tau, F_n) - g_2(\tau, F_0)] + O(n^{-3/2}) \tag{34}$$

almost surely uniformly over τ.

- The leading term on the RHS is of size $O(n^{-1/2})$ almost surely uniformly over τ since $F_n - F_0 = O(n^{-1/2})$ almost surely uniformly over the support of F_0.
- Thus, the bootstrap makes an error of size $O(n^{-1/2})$ almost surely, which is the same as the size of the error made by first-order asymptotic approximations.
- In terms of the rate of convergence to zero of the approximation error, the bootstrap has the same accuracy as first-order asymptotic approximations.

In this sense, nothing is lost in terms of accuracy by using the bootstrap instead of the first-order approximations, but nothing is gained either.

Now suppose that T_n is *asymptotically pivotal.* Then the asymptotic distribution of T_n is independent of F_0, and $G_\infty(\tau, F_n) = G_\infty(\tau, F_0)$ for all τ. (32) and (33) now yield

$$\begin{aligned} G_n(\tau, F_n) - G_n(\tau, F_0) = {} & \frac{1}{\sqrt{n}}[g_1(\tau, F_n) - g_1(\tau, F_0)] \\ & + \frac{1}{n}[g_2(\tau, F_n) - g_2(\tau, F_0)] + O(n^{-3/2}) \end{aligned} \tag{35}$$

almost surely.

Consider the leading term on the RHS. It follows from continuity of g_1 w.r.t. its second argument that this term has size $O(n^{-1})$ almost surely uniformly over τ.

Now the bootstrap makes an error of size $O(n^{-1})$, which is smaller as $n \to \infty$ than the error made by first-order asymptotic approximations.

Thus the bootstrap is more accurate than first-order asymptotic theory for estimating the distribution of a smooth asymptotically pivotal statistic.

Bootstrap References

[1] Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika* **74** 457–468.

[2] Beran, R. (1997). Diagnosing bootstrap success. *Annals of the Institute of Statistical Mathematics* **49** 1–24.

[3] Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics* **9** 1196–1217.

[4] Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press.

[5] Davison, A. C., Hinkley, D. V., and Young, G.A. (2003). Recent developments in bootstrap methodology. *Statistical Science* **18**(2), 141–157.

[6] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics* **7** 1–26.

[7] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

[8] Hall, P. (1986). On the bootstrap and confidence intervals. *Annals of Statistics* **14** 1431–1452.

[9] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

[10] Horowitz, J. (2001). The bootstrap. In *Handbook of Econometrics Volume 5*, Chapter 52. Elsevier.

[11] Mammen, E. (1992). *When Does Bootstrap Work?* Lecture Notes in Statistics **77**. Springer, New York.

[12] Putter, H. (1994). *Consistency of Resampling Methods*. Ph.D. dissertation, Leiden University.

[13] Young, G.A. (1994). Bootstrap: More than a stab in the dark? *Statistical Science* **9**(3), 382–415.

Is this relevant for Bayesians?

Recall the Bernstein von Mises theorem:

Theorem (Following van der Vaart's book)

Let the experiment $(P_\theta : \theta \in \Theta)$ be differentiable in quadratic mean at θ_0 with nonsingular Fisher information matrix I_{θ_0}, and suppose that $\forall\ \varepsilon > 0$, $\exists$ a sequence of tests s.t.

$$P^n_{\theta_0}\phi_n \to 0, \qquad \sup_{\|\theta-\theta_0\|\geq\varepsilon} P^n_\theta(1-\phi_n) \to 0.$$

Furthermore, let the prior measure be absolutely continuous in a neighborhood of θ_0 with a continuous positive density at θ_0. Then the corresponding posterior distributions satisfy

$$\left\| P_{\sqrt{n}(\overline{\Theta}_n-\theta_0)\ |X_1,\ldots,X_n} - \mathcal{N}(\Delta_{n,\theta_0}, I^{-1}_{\theta_0}) \right\| \overset{P^n_{\theta_0}}{\to} 0.$$

Expansions/refinements in finite dimensions:

- Johnson (1970s)
- Sinha/Joshi/Ghosh (1980s)
- Woodroofe & Weng

Approximate Bayesian inference:

- Bickel & Ghosh (1990)
- DiCiccio & Martin (1993)
- Sweeting (every couple years)
- Ventura et al.

Some Recent Bayesian Work

METRON
DOI 10.1007/s40300-014-0041-4

Approximate Bayesian computation with modified log-likelihood ratios

Laura Ventura · Nancy Reid

Received: 30 December 2013 / Accepted: 11 April 2014

Abstract The aim of this contribution is to discuss approximate Bayesian computation based on the asymptotic theory of modified likelihood roots and log-likelihood ratios. Results on third-order approximations for univariate posterior distributions, also in the presence of nuisance parameters, are reviewed and the computation of asymptotic credible sets for a vector parameter of interest is illustrated. All these approximations are available at little additional computational cost over simple first-order approximations. Some illustrative examples are discussed, with particular attention to the use of matching priors.

Keywords Bayesian simulation · Credible set · Higher-order asymptotics · Laplace approximation · Marginal posterior distribution · Matching priors · Modified likelihood root · Nuisance parameter · Pereira–Stern measure of evidence · Precise null hypothesis · Tail area probability

1 Introduction

Asymptotic arguments are widely used in Bayesian inference, and in recent years there have been considerable developments of so-called higher-order asymptotics. The aim of this contribution is to discuss recent advances in approximate Bayesian computation based on the theory of higher-order asymptotics. The theory provides very accurate approximations to posterior distributions, and to various summary quantities of interest, including tail areas and credible regions. The approximations are based on modifications of the usual log-likelihood ratio statistic. It is argued that analytic approximations still have an important role to play in Bayesian statistics.

L. Ventura (✉)
Department of Statistical Sciences, University of Padova, Padua, Italy
e-mail: ventura@stat.unipd.it

N. Reid
Department of Statistics, University of Toronto, Toronto, ON, Canada
e-mail: reid@utstat.utoronto.ca

Bayesian Analysis (2014) **9**, Number 1, pp. 129–146

Marginal Posterior Simulation via Higher-order Tail Area Approximations

Erlis Ruli *, Nicola Sartori * and Laura Ventura *

Abstract. A new method for posterior simulation is proposed, based on the combination of higher-order asymptotic results with the inverse transform sampler. This method can be used to approximate marginal posterior distributions, and related quantities, for a scalar parameter of interest, even in the presence of nuisance parameters. Compared to standard Markov chain Monte Carlo methods, its main advantages are that it gives independent samples at a negligible computational cost, and it allows prior sensitivity analyses under the same Monte Carlo variation. The method is illustrated by a genetic linkage model, a normal regression with censored data and a logistic regression model.

Keywords: Asymptotic expansion, Bayesian computation, Inverse transform sampling, Marginal posterior distribution, MCMC, Modified likelihood root, Nuisance parameter, Sensitivity analysis.

1 Introduction

Consider a parametric statistical model with density $f(y;\boldsymbol{\theta})$, with $\boldsymbol{\theta} = (\psi, \boldsymbol{\lambda})$, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$ $(d > 1)$, where ψ is a scalar parameter of interest and $\boldsymbol{\lambda}$ is a $(d-1)$-dimensional nuisance parameter. Let $\ell(\boldsymbol{\theta}) = \ell(\psi, \boldsymbol{\lambda}) = \ell(\psi, \boldsymbol{\lambda}; \mathbf{y})$ denote the log-likelihood function based on data $\mathbf{y} = (y_1, \ldots, y_n)$, $\pi(\boldsymbol{\theta}) = \pi(\psi, \boldsymbol{\lambda})$ be a prior distribution of $(\psi, \boldsymbol{\lambda})$ and let $\pi(\boldsymbol{\theta}|\mathbf{y}) = \pi(\psi, \boldsymbol{\lambda}|\mathbf{y}) \propto \pi(\psi, \boldsymbol{\lambda}) \exp\{\ell(\psi, \boldsymbol{\lambda})\}$ be the posterior distribution of $(\psi, \boldsymbol{\lambda})$. Bayesian inference on ψ, in the presence of the nuisance parameter $\boldsymbol{\lambda}$, is based on the marginal posterior distribution

$$\pi(\psi|\mathbf{y}) = \int \pi(\psi, \boldsymbol{\lambda}|\mathbf{y})\, d\boldsymbol{\lambda}, \tag{1}$$

which is typically approximated numerically, by means of Monte Carlo integration methods. In order to approximate (1), a variety of Markov chain Monte Carlo (MCMC) schemes have been proposed in the literature (see, e.g., Robert and Casella 2004). However, MCMC methods in practice may need to be specifically tailored to the particular model (e.g. choice of proposal, convergence checks, etc.) and they may have poor tail behavior, especially when d is large.

Parallel with these simulation-based procedures has been the development of analytical higher-order approximations for parametric inference in small samples (see, e.g., Brazzale and Davison 2008, and references therein). Using higher-order asymptotics it is possible to avoid the difficulties related to MCMC methods and obtain accurate

*Department of Statistics, University of Padova, Padova, Italy ruli@stat.unipd.it

 DOI:10.1214/13-BA851

Biometrika (2016), **103**, 2, *pp.* 337–349
Printed in Great Britain
doi: 10.1093/biomet/asw018

Exponential tilting in Bayesian asymptotics

BY S. A. KHARROUBI
Department of Nutrition and Food Sciences, Faculty of Agricultural and Food Sciences, American University of Beirut, P.O. Box 11–0236, Riad El Solh 1107–2020 Beirut, Lebanon
sk157@aub.edu.lb

AND T. J. SWEETING
Department of Statistical Science, University College London, London WC1E 6BT, U.K.
t.sweeting@ucl.ac.uk

SUMMARY

We use exponential tilting to obtain versions of asymptotic formulae for Bayesian computation that do not involve conditional maxima of the likelihood function, yielding a more stable computational procedure and significantly reducing computational time. In particular we present an alternative version of the Laplace approximation for a marginal posterior density. Implementation of the asymptotic formulae and a modified signed root based importance sampler are illustrated with an example.

Some key words: Approximate Bayesian inference; Exponential tilting; Higher-order asymptotic theory; Importance sampling; Laplace approximation; Signed root loglikelihood ratio.

1. INTRODUCTION

Accurate asymptotic approximations based on signed root loglikelihood ratios for Bayesian inference have been obtained by a number of authors; see for example DiCiccio et al. (1990), DiCiccio & Martin (1991), DiCiccio & Field (1991) and Sweeting (1995). Sweeting (1996) and Sweeting & Kharroubi (2003) derive asymptotic formulae for posterior expectations and predictive distributions that are correct to the same asymptotic order as the corresponding Tierney & Kadane (1986) expressions. In Kharroubi & Sweeting (2010) asymptotic formulae based on signed roots form the basis for an importance sampling scheme for Bayesian computation.

These formulae all require the repeated computation of conditional maxima of the likelihood function, which can cause computational difficulties especially in higher dimensions and when many signed root inversions are required, such as in the importance sampling scheme of Kharroubi & Sweeting (2010). Here we use exponential tilting to develop alternative asymptotic posterior approximations that do not require conditional maximization. We further indicate how these new approximations can be incorporated into the signed root based importance sampling scheme of Kharroubi & Sweeting (2010). The theory is illustrated by a nonlinear repeated measures model for bacterial clearance.

Exponential tilting has been extensively used in statistics in a variety of contexts. It is a powerful tool for the development of accurate saddlepoint approximations in frequentist inference; see for example Goutis & Casella (1999). Fraser et al. (1999) implement a one-dimensional exponential tilt based on a hypothetical value of a scalar parameter of interest for the purpose

An example of a recent breakthrough

Biometrika (2016), *pp.* 1–15
Printed in Great Britain
doi: 10.1093/biomet/asw022

Accurate directional inference for vector parameters

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

Department of Statistical Sciences, University of Toronto, Toronto M5S 3G3, Canada
dfraser@utstat.toronto.edu reid@utstat.toronto.edu

AND N. SARTORI

Dipartimento di Scienze Statistiche, Università degli Studi di Padova, 35121 Padova, Italy
sartori@stat.unipd.it

SUMMARY

We consider statistical inference for a vector-valued parameter of interest in a regular asymptotic model with a finite-dimensional nuisance parameter. We use highly accurate likelihood theory to derive a directional test, in which the p-value is obtained by one-dimensional numerical integration. This extends the results of Davison et al. (2014) for linear exponential families to nonlinear parameters of interest and to more general models. Examples and simulations provide comparisons with the likelihood ratio test and adjusted versions of the likelihood ratio test. The directional approach gives extremely accurate inference, even in high-dimensional settings where the likelihood ratio versions can fail catastrophically.

Some key words: Behrens–Fisher problem; Box–Cox model; Higher-order asymptotics; Likelihood ratio test; Marginal independence; Tangent exponential model.

1. INTRODUCTION

In this paper we develop tests for vector parameters of interest in parametric models, and thus generalize the directional approach of Davison et al. (2014). In that paper attention was restricted to exponential-family models and to parameters of interest that are linear in the canonical parameter of the model. Here we extend the approach to nonlinear parameters in general continuous models; this requires a substantially different reduction process.

The usual likelihood-based method of inference for vector parameters is to refer the loglikelihood ratio statistic to a χ^2 distribution with degrees of freedom equal to the number of parameters of interest. The accuracy of the χ^2 approximation can be improved by the large-deviation method proposed in Skovgaard (2001), or by Bartlett correction (Bartlett, 1937), where the latter rescales the likelihood ratio statistic by its expected value under the hypothesis for the parameter of interest. The likelihood ratio statistic and these two adjusted versions are omnibus test statistics, in the sense that values more extreme than the observed value can correspond to alternatives in any part of the parameter space.

The directional method proposed in Fraser & Massam (1985) and further developed by Skovgaard (1988) uses a test statistic that is conditioned on the direction of departure of the observed data from the data expected under the hypothesis. In Davison et al. (2014) the directional method was applied to inference in contingency tables, logistic regression, equality of normal variances, and Gaussian graphical models. Simulations there showed the method to be

Stuff that I have worked on

Statistica Sinica **25** (2015), 1355-1376
doi:http://dx.doi.org/10.5705/ss.2013.055

STABILITY AND UNIQUENESS OF p-VALUES FOR LIKELIHOOD-BASED INFERENCE

Thomas J. DiCiccio, Todd A. Kuffner, G. Alastair Young and Russell Zaretzki

Cornell University, Washington University in St. Louis, Imperial College London and University of Tennessee, Knoxville

Abstract: Likelihood-based methods of statistical inference provide a useful general methodology that is appealing, as a straightforward asymptotic theory can be applied for their implementation. It is important to assess the relationships between different likelihood-based inferential procedures in terms of accuracy and adherence to key principles of statistical inference, in particular those relating to conditioning on relevant ancillary statistics. An analysis is given of the stability properties of a general class of likelihood-based statistics, including those derived from forms of adjusted profile likelihood, and comparisons are made between inferences derived from different statistics. In particular, we derive a set of sufficient conditions for agreement to $O_p(n^{-1})$, in terms of the sample size n, of inferences, specifically p-values, derived from different asymptotically standard normal pivots. Our analysis includes inference problems concerning a scalar or vector interest parameter, in the presence of a nuisance parameter.

Key words and phrases: Adjusted profile likelihood, ancillary statistic, likelihood, modified signed root likelihood ratio statistic, nuisance parameter, pivot, stability.

1. Introduction

A highly useful statistical methodology for inference on a scalar or vector interest parameter in the presence of a nuisance parameter is furnished by procedures based on the likelihood function, including tests and confidence sets based on the likelihood ratio statistic. Though no explicit optimality criteria are invoked, a quite general asymptotic theory allows straightforward implementation of such methodology in a wide range of settings. However, accuracy and what may be termed inferential correctness are (Young (2009)) key desiderata of any parametric inference. When constructing, say, a confidence set for a parameter of interest in the presence of nuisance parameters, we desire high levels of coverage accuracy from the confidence set. Further, it is important that procedures are inferentially correct, meaning that they respect key principles of inference, in particular those relating to appropriate conditioning on ancillary information when this is relevant. The crucial issue here is the stability of the statistic used for inference, the extent to which the unconditional distribution of the statistic

Journal of Statistical Planning and Inference 165 (2015) 1–12

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

ELSEVIER

Quantifying nuisance parameter effects via decompositions of asymptotic refinements for likelihood-based statistics

CrossMark

T.J. DiCiccio[a], T.A. Kuffner[b,*], G.A. Young[c]

[a] *Cornell University, Ithaca, NY 14853, USA*
[b] *Washington University in St. Louis, St. Louis, MO 63130, USA*
[c] *Imperial College London, London SW7 2AZ, UK*

A R T I C L E I N F O

Article history:
Received 11 December 2014
Received in revised form 23 February 2015
Accepted 12 March 2015
Available online 25 March 2015

Keywords:
Adjusted signed root likelihood ratio
Ancillary statistic
Bartlett correction
Cornish–Fisher
Decomposition
Exponential family
Nuisance parameter
Profile likelihood

A B S T R A C T

Accurate inference on a scalar interest parameter in the presence of a nuisance parameter may be obtained using an adjusted version of the signed root likelihood ratio statistic, in particular Barndorff-Nielsen's R^* statistic. The adjustment made by this statistic may be decomposed into a sum of two terms, interpreted as correcting respectively for the possible effect of nuisance parameters and the deviation from standard normality of the signed root likelihood ratio statistic itself. We show that the adjustment terms are determined to second-order in the sample size by their means. Explicit expressions are obtained for the leading terms in asymptotic expansions of these means. These are easily calculated, allowing a simple way of quantifying and interpreting the respective effects of the two adjustments, in particular of the effect of a high dimensional nuisance parameter. Illustrations are given for a number of examples, which provide theoretical insight to the effect of nuisance parameters on parametric inference. The analysis provides a decomposition of the mean of the signed root statistic involving two terms: the first has the property of taking the same value whether there are no nuisance parameters or whether there is an orthogonal nuisance parameter, while the second is zero when there are no nuisance parameters. Similar decompositions are discussed for the Bartlett correction factor of the likelihood ratio statistic, and for other asymptotically standard normal pivots.

1. Introduction

We are concerned with inference on a scalar interest parameter in the presence of a, possibly high dimensional, nuisance parameter, based on a data sample of size n, and with identification of procedures which yield repeated sampling accuracy. In this setting, inference accurate to third order, that is with repeated sampling error of order $O(n^{-3/2})$, may be obtained using an adjusted version of the signed root likelihood ratio statistic, in particular through use of Barndorff-Nielsen's R^* statistic (Barndorff-Nielsen, 1986).

The R^* statistic is particularly useful in two contexts. In full, multi-parameter exponential family models inference based on standard normal approximation to the sampling distribution of the R^* statistic approximates to third order the optimal, conditional, but generally intractable, inference, which is based on conditioning on the sufficient statistic for the nuisance parameter. In more general models which admit an ancillary statistic, taken to mean an approximately distribution free statistic which together with the maximum likelihood estimator constitutes a minimal sufficient statistic for the full

* Corresponding author.
E-mail addresses: tjd9@cornell.edu (T.J. DiCiccio), kuffner@math.wustl.edu (T.A. Kuffner), alastair.young@imperial.ac.uk (G.A. Young).

http://dx.doi.org/10.1016/j.jspi.2015.03.006

Visual Bibliography

Gentle Introductions Part I (2003, 2009)

The Annals of Statistics
2003, Vol. 31, No. 6, 1695–1731

THE 2000 WALD MEMORIAL LECTURES

ASYMPTOTICS AND THE THEORY OF INFERENCE

BY N. REID

University of Toronto

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-some 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 comparing 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?

Received May 2002; revised March 2003.
AMS 2000 subject classifications. Primary 62-02; secondary 62E20, 62F05.
Key words and phrases. Ancillarity, approximation, Bayesian inference, conditioning, Laplace approximation, likelihood, matching priors, p^*, p-values, r^*, saddlepoint approximation, tail area, tangent exponential model.

Australian & New Zealand Journal of Statistics

Aust. N. Z. J. Stat. **51**(2), 2009, 115–126 doi: 10.1111/j.1467-842X.2009.00548.x

ROUTES TO HIGHER-ORDER ACCURACY IN PARAMETRIC INFERENCE

G. ALASTAIR YOUNG[1]

Imperial College London

Summary

Developments in the theory of frequentist parametric inference in recent decades have been driven largely by the desire to achieve higher-order accuracy, in particular distributional approximations that improve on first-order asymptotic theory by one or two orders of magnitude. At the same time, much methodology is specifically designed to respect key principles of parametric inference, in particular conditionality principles. Two main routes to higher-order accuracy have emerged: analytic methods based on 'small-sample asymptotics', and simulation, or 'bootstrap', approaches. It is argued here that, of these, the simulation methodology provides a simple and effective approach, which nevertheless retains finer inferential components of theory. The paper seeks to track likely developments of parametric inference, in an era dominated by the emergence of methodological problems involving complex dependences and/or high-dimensional parameters that typically exceed available data sample sizes.

Key words: analytic methods; ancillary statistic; bootstrap; conditionality; full exponential family; likelihood; likelihood ratio statistic; nuisance parameter; objective Bayes; signed root likelihood ratio statistic; simulation.

1. Introduction

The primary purpose of this paper is to review key aspects of frequentist parametric inference methodology, as it has developed over the last 25 years or so. Developments have been driven largely by the desire to achieve higher-order accuracy, in particular distributional approximations that improve on first-order asymptotic theory by one or two orders of magnitude.

Two main routes to achieving this higher-order accuracy have emerged: analytic methods based on the techniques of small-sample asymptotics, such as saddlepoint approximation, and simulation, or bootstrap, approaches. Objective Bayes methods provide a further route to higher-order frequentist accuracy in many circumstances.

A personal evaluation of these different routes is provided, with the aim of justifying the assertion that parametric bootstrap procedures, with appropriate handling of nuisance parameters, provide satisfactory, simple approaches to inference, yielding the desired higher-order accuracy while retaining finer inferential components of statistical theory, in particular those associated with conditional inference. Parametric bootstrap methods might, therefore, be viewed as the inferential approach of choice in many settings.

[1]Department of Mathematics, Imperial College London, London SW7 2AZ, UK.
e-mail: alastair.young@imperial.ac.uk
Acknowledgments. This paper is based on a keynote address given at the Australian Statistical Conference, Melbourne, July 2008. The author is grateful to the organizers of the conference, in particular Michael Martin and Steven Roberts, for the invitation to speak. Many of the ideas expressed in the paper have emerged from collaborative work with Tom DiCiccio.

Gentle Introductions Part II (2008, 2016)

Statistical Science
2008, Vol. 23, No. 4, 465–484
DOI: 10.1214/08-STS273

Accurate Parametric Inference for Small Samples

Alessandra R. Brazzale and Anthony C. Davison

Abstract. We outline how modern likelihood theory, which provides essentially exact inferences in a variety of parametric statistical problems, may routinely be applied in practice. Although the likelihood procedures are based on analytical asymptotic approximations, the focus of this paper is not on theory but on implementation and applications. Numerical illustrations are given for logistic regression, nonlinear models, and linear non-normal models, and we describe a sampling approach for the third of these classes. In the case of logistic regression, we argue that approximations are often more appropriate than 'exact' procedures, even when these exist.

Key words and phrases: Conditional inference, heteroscedasticity, logistic regression, Lugannani–Rice formula, Markov chain Monte Carlo, nonlinear model, R, regression-scale model, saddlepoint approximation, spline, statistical computing.

1. INTRODUCTION

Monte Carlo inference has developed remarkably over the last 30 years. Bootstrap procedures (Efron, 1979) are used for a wide range of problems (Efron and Tibshirani, 1993, Davison and Hinkley, 1997). Markov chain Monte Carlo simulation has transformed Bayesian modelling (Robert and Casella, 2004). The combination of iterative simulation with importance sampling and improved algorithms for full enumeration of discrete sample spaces has had a strong impact on the analysis of contingency tables (Forster, McDonald and Smith, 1996, Smith, Forster and McDonald, 1996, Diaconis and Sturmfels, 1998, Mehta, Patel and Senchaudhuri, 2000). More recently there has been a rise in Bayesian nonparametric modelling (Denison et al., 2002), which parallels the use of the bootstrap for nonparametric frequentist inference. All these techniques use simulation to avoid tailoring analytical work to specific problems.

Parallel with these developments has been the development of analytical approximations for parametric inference in small samples, initiated by Fisher (1934) but largely overlooked until new developments were stimulated by Efron and Hinkley (1978) and Barndorff-Nielsen and Cox (1979). A flood of subsequent work is summarized in the books of Barndorff-Nielsen and Cox (1994), Pace and Salvan (1997), and Severini (2000). The efforts of many researchers, particularly O. E. Barndorff-Nielsen, (1983, 1986) and D. A. S. Fraser (e.g., Fraser, 1990; Fraser, Reid and Wu, 1999) and their co-workers, have led to an elegant theory of near-exact inference based on small samples from parametric models. Its theoretical basis is saddlepoint and related approximation (Daniels, 1954, 1987), and further developments have been well described by Reid (1988, 1995, 2003). These methods are highly accurate in many situations, but are nevertheless under-used compared to the simulation procedures mentioned above. One reason for this may be their arcane basis in the conditionality principle, ancillary statistics and marginalization, and another may be the forbidding technical details, but the main reason is undoubtedly the lack of suitable software. Unlike the bootstrap libraries available in general-purpose

Alessandra R. Brazzale is Associate Professor of Statistics, Department of Social, Cognitive and Quantitative Sciences (DSSCQ) University of Modena and Reggio Emilia, Viale Allegri 9, 42100 Reggio Emilia, Italy (e-mail: alessandra.brazzale@unimore.it) and Institute of Biomedical Engineering, Italian National Research Council. Anthony C. Davison is Professor of Statistics, Institute of Mathematics (IMA-FSB-EPFL), Ecole Polytechnique Fédérale de Lausanne, Station 8, 1015 Lausanne, Switzerland (e-mail: anthony.davison@epfl.ch).

Modern Likelihood-Frequentist Inference

Donald A. Pierce
Oregon Health and Sciences University
Portland, Oregon U.S.A
Ruggero Bellio
University of Udine
Udine, Italy

Summary

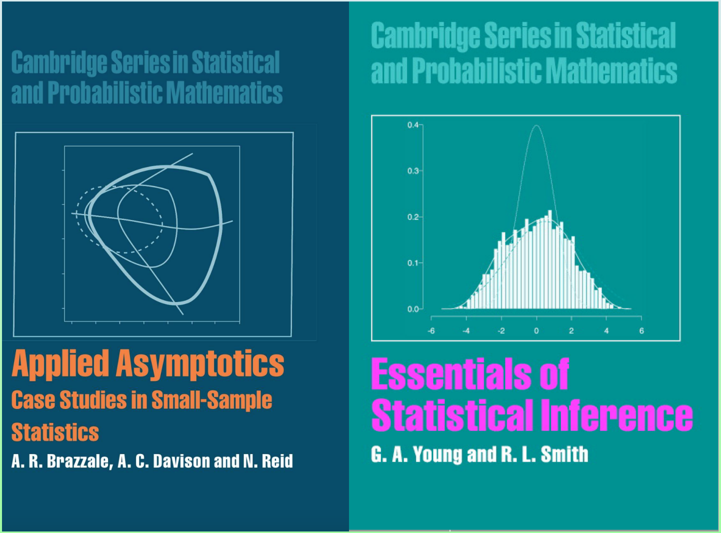
We offer an exposition of modern higher-order likelihood inference, and introduce software to implement this in a fairly general setting. The aim is to make more accessible an important development in statistical theory and practice. The software, implemented in an *R* package, requires only that the user provide code to compute the likelihood function and to specify the extra-likelihood aspects of the model. The exposition charts a narrow course through the developments, intending thereby to make these more widely accessible. It includes the likelihood ratio approximation to the distribution of the maximum likelihood estimator, and transformation of this yielding a second-order approximation to the distribution of the likelihood ratio test statistic. This follows developments of Barndorff-Nielsen and others. The software utilizes the approximation to required Jacobians as developed by Skovgaard, which is included in the exposition. Several examples of using the software are provided.

Some Key Words: ancillary statistic, conditional inference, likelihood asymptotics, modified likelihood ratio, modified profile likelihood, neo-Fisherian inference, nuisance parameter, saddlepoint approximation, p^* formula.

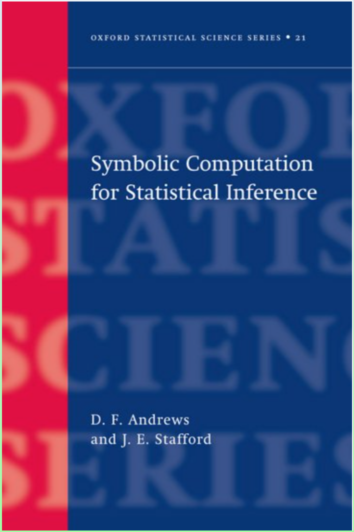
1. INTRODUCTION AND BASIC CONCEPTS

Special likelihood-based procedures, modifying usual inferential approximations for much higher accuracy, have emerged in recent years (Davison, 2003, Ch. 12; Brazzale and Davison, 2008; Lozada-Can and Davison (2010). The performance of these methods is uncanny, and it is not unreasonable to characterize them as often 'close to exact'. However, there is considerably more to modern likelihood

Accessible Monographs (2007, 2005)

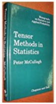


Symbolic Computation (2000)



Making asymptotic expansions more friendly!

Bookshelf Essentials Part I (1987)



Tensor Methods in Statistics (Chapman & Hall/CRC Monographs on Statistics & Applied Probability) (Hardcover)

by P. McCullagh

Be the first to review this item | Share

Access codes and supplements are not guaranteed with used items.

Refine by Clear all

Shipping

- ☐ Free shipping

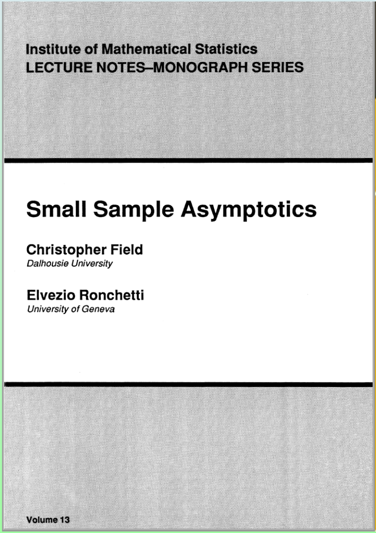
Condition

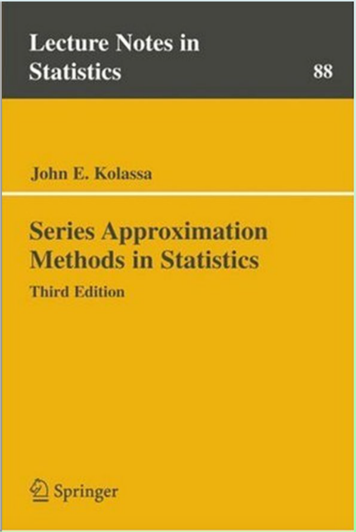
- ☑ Used
  - ☑ Like New
  - ☑ Very Good
  - ☑ Good
  - ☑ Acceptable

Price + Shipping	Condition (Learn more)	Delivery	Seller Information	Buying Options
$370.18 + $3.99 shipping + $0.00 estimated tax	**Used - Very Good** .	• **Arrives between** October 6-24. • Ships from GA, United States. • Shipping rates and return policy.	**Brooke Books** 99% positive over the past 12 months. (71,904 total ratings)	Add to cart or Sign in to turn on 1-Click ordering.
$763.49 + $3.99 shipping + $0.00 estimated tax	**Used - Acceptable** Cover appears used and pages may show heavy wear. May have notes,... » Read more	• **Arrives between** October 6-24. • Ships from NC, United States. • Shipping rates and return policy.	**Red Rhino** 96% positive over the past 12 months. (4,761 total ratings)	Add to cart or Sign in to turn on 1-Click ordering.
$818.60 + $3.99 shipping + $0.00 estimated tax	**Used - Like New** Like New Condition.	• **Arrives between** October 6-24. • Ships from FL, United States. • Shipping rates and return policy.	**Cloud-9-Books** 97% positive over the past 12 months. (17,938 total ratings)	Add to cart or Sign in to turn on 1-Click ordering.
$1,872.21 + $3.99 shipping + $0.00 estimated tax	**Used - Acceptable**	• **Arrives between** October 6-24. • Ships from TX, United States. • Shipping rates and return policy.	**Lucky's Fulfillment** 96% positive over the past 12 months. (7,277 total ratings)	Add to cart or Sign in to turn on 1-Click ordering.
$3,995.00 + $3.99 shipping + $0.00 estimated tax	**Used - Acceptable**	• **Arrives between** October 6-24. • Ships from WA, United States. • Shipping rates and return policy.	**1st_class_books** 98% positive over the past 12 months. (4,437 total ratings)	Add to cart or Sign in to turn on 1-Click ordering.

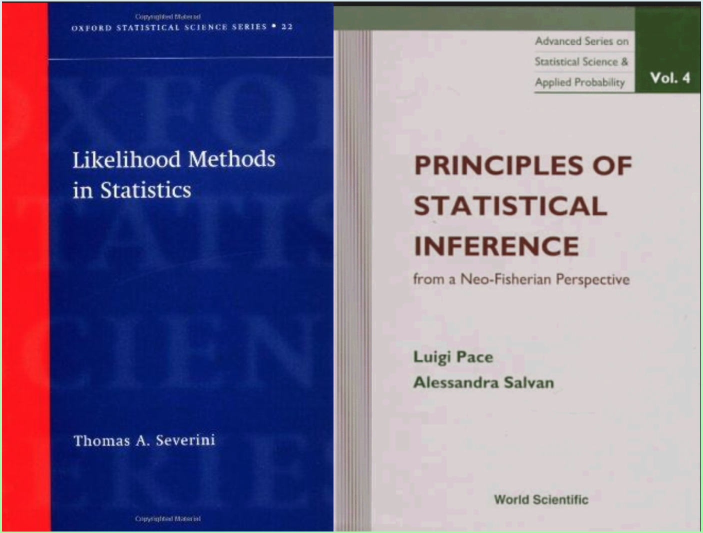
There are cheaper copies, but Peter McCullagh's 1987 'Tensor Methods in Statistics' is obviously very dear to some of us!

Bookshelf Essentials Part II (1990, 2006)

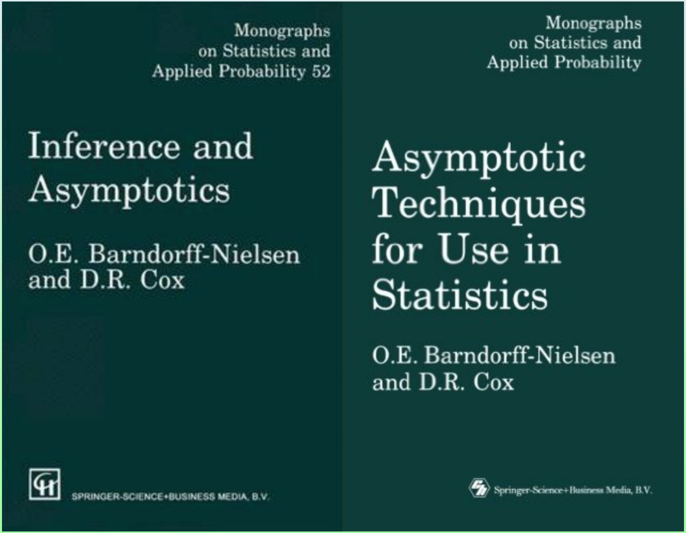




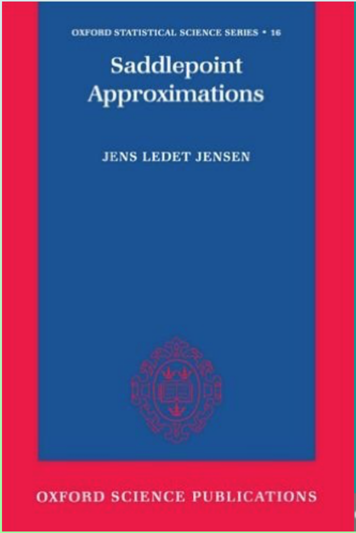
Bookshelf Essentials Part III (2000, 1997)



Bookshelf Essentials Part IV (1994, 1989)



Bookshelf Essentials Part V (1995, 2007)



`R` packages

`hoa` Alessandra Brazzale (2015)

- higher order likelihood inference for regression models; emphasis on saddlepoint methods

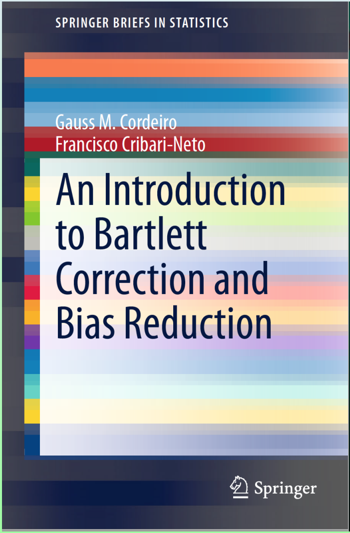
`likelihoodAsy` Ruggero Bellio and Don Pierce (2016)

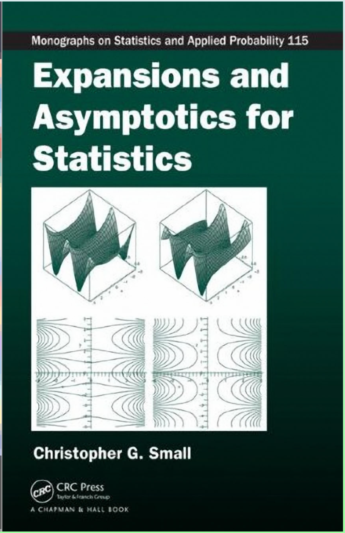
- implement inference using R and R^* for general parametric models; also compute modified profile likelihood

`nlreg` Alessandra Brazzale and Ruggero Bellio (2014)

- higher-order inference for nonlinear, possibly heteroscedastic models

Recent Monographs (2014, 2010)





Who and Where Part I

Australia
- Michael Martin, *Australian National University*
- John Robinson, *University of Sydney*
- Steven Stern, *Australian National University*

Brazil
- Gauss Cordeiro, *Federal University of Pernambuco*
- Francisco Cribari-Neto, *Federal University of Pernambuco*
- Silvia Ferrari, *University of Sao Paulo*
- Artur Lemonte, *Federal University of Pernambuco*

Canada
- Christopher Field, *Dalhousie University*
- Don Fraser, *University of Toronto*
- Nancy Reid, *University of Toronto*
- Chris Small, *University of Waterloo*
- Augustine Wong, *York University*

Denmark
- Ole Barndorff-Nielsen, *Aarhus University*
- Jens Jensen, *Aarhus University*
- Ib Skovgaard, *University of Copenhagen*

Greece
- Anna Nicolaou, *University of Macedonia*

Hong Kong
- Ngai Hang Chan, *Chinese University of Hong Kong*
- Stephen M.S. Lee, *University of Hong Kong*
- Phillip Yam, *Chinese University of Hong Kong*

Who and Where Part II

- India
  - Rahul Mukerjee, *Indian Institute of Management Calcutta*
- Italy
  - Ruggero Bellio, *University of Udine*
  - Alessandra Brazzale, *University of Padova*
  - Luigi Pace, *University of Udine*
  - Walter Racugno, *University of Cagliari*
  - Alessandra Salvan, *University of Padova*
  - Nicola Sartori, *University of Padova*
  - Laura Ventura, *University of Padova*
- Japan
  - Shinto Eguchi, *Institute of Statistical Mathematics*
  - Akimichi Takemura, *University of Tokyo / Shiga University*
- Switzerland
  - Anthony Davison, *EPFL*
  - Riccardo Gatto, *University of Bern*
  - Elvezio Ronchetti, *University of Geneva*
- Taiwan
  - Ruby Weng, *National Chengchi University*
- United Kingdom
  - David Cox, *University of Oxford*
  - Trevor Sweeting, *University College London*
  - Andrew Wood, *University of Nottingham*
  - G. Alastair Young, *Imperial College London*

Who and Where Part III

USA

- Rabi Bhattacharya, *University of Arizona*
- Peter Bickel, *UC Berkeley*
- Jim Booth, *Cornell University*
- Ronald Butler, *Southern Methodist University*
- Anirban DasGupta, *Purdue University*
- Gauri Datta, *University of Georgia*
- Tom DiCiccio, *Cornell University*
- Brad Efron, *Stanford University*
- J.K. Ghosh, *Purdue University*
- Malay Ghosh, *University of Florida*
- David Hinkley, *UC Santa Barbara*
- Robert Kass, *Carnegie Mellon University*
- Soumendra Lahiri, *NC State University*
- Nicole Lazar, *University of Georgia*
- Subhashis Ghosal, *NC State University*
- John Kolassa, *Rutgers University*
- Todd Kuffner, *Washington University in St. Louis*
- Peter McCullagh, *University of Chicago*
- Per Mykland, *University of Chicago*
- Don Pierce, *Oregon State University*
- Joseph Romano, *Stanford University*
- Tom Severini, *Northwestern University*
- Ana-Maria Staicu, *NC State University*
- Rob Strawderman, *University of Rochester*
- Suojin Wang, *Texas A&M University*
- Martin Wells, *Cornell University*
- Russell Zaretzki, *University of Tennessee Knoxville*