hoa: An R Package Bundle for Higher Order Likelihood Inference

by Alessandra R. Brazzale

Rnews, **5/1** May 2005, pp. 20–27

Introduction

The likelihood function represents the basic ingredient of many commonly used statistical methods for estimation, testing and the calculation of confidence intervals. In practice, much application of likelihood inference relies on first order asymptotic results such as the central limit theorem. The approximations can, however, be rather poor if the sample size is small or, generally, when the average information available per parameter is limited. Thanks to the great progress made over the past twenty-five years or so in the theory of likelihood inference, very accurate approximations to the distribution of statistics such as the likelihood ratio have been developed. These not only provide modifications to well-established approaches, which result in more accurate inferences, but also give insight on when to rely upon first order methods. We refer to these developments as *higher order asymptotics*.

One intriguing feature of the theory of higher order likelihood asymptotics is that relatively simple and familiar quantities play an essential role. The higher order approximations discussed in this paper are for the significance function, which we will use to set confidence limits or to calculate the p-value associated with a particular hypothesis of interest. We start with a concise overview of the approximations used in the remainder of the paper. Our first example is an elementary one-parameter model where one can perform the calculations easily, chosen to illustrate the potential accuracy of the procedures. Two more elaborate examples, an application of binary logistic regression and a nonlinear growth curve model, follow. All examples are carried out using the R code of the `hoa` package bundle.

Basic ideas

Assume we observed n realizations $y_1, \ldots, y_n$ of independently distributed random variables $Y_1, \ldots, Y_n$ whose density function $f(y_i; \theta)$ depends on an

unknown parameter θ. Let $\ell(\theta) = \sum_{i=1}^{n} \log f(y_i; \theta)$ denote the corresponding log likelihood and $\hat{\theta} = \text{argmax}_\theta \ell(\theta)$ the maximum likelihood estimator. In almost all applications the parameter θ is not scalar but a vector of length d. Furthermore, we may re-express it as $\theta = (\psi, \lambda)$, where ψ is the d_0-dimensional *parameter of interest*, about which we wish to make inference, and λ is a so-called *nuisance parameter*, which is only included to make the model more realistic.

Confidence intervals and p-values can be computed using the *significance function* $p(\psi; \hat{\psi}) = \Pr(\hat{\Psi} \leq \hat{\psi}; \psi)$ which records the probability left of the observed "data point" $\hat{\psi}$ for varying values of the unknown parameter ψ (Fraser, 1991). Exact elimination of λ, however, is possible only in few special cases (Severini, 2000, Sections 8.2 and 8.3). A commonly used approach is to base inference about ψ on the *profile log likelihood* $\ell_{\text{p}}(\psi) = \ell(\psi, \hat{\lambda}_\psi)$, which we obtain from the log likelihood function by replacing the nuisance parameter with its constrained estimate $\hat{\lambda}_\psi$ obtained by maximising $\ell(\theta) = \ell(\psi, \lambda)$ with respect to λ for fixed ψ. Let $j_{\text{p}}(\psi) = -\partial^2 \ell_{\text{p}}(\psi)/\partial\psi\partial\psi^\top$ denote the observed information from the profile log likelihood. Likelihood inference for scalar ψ is typically based on the

- Wald statistic, $\quad w(\psi) = j_{\text{p}}(\hat{\psi})^{1/2}(\hat{\psi} - \psi)$;
- likelihood root,

$$r(\psi) = \text{sign}(\hat{\psi} - \psi)\left[2\{\ell_{\text{p}}(\hat{\psi}) - \ell_{\text{p}}(\psi)\}\right]^{1/2};$$

  or

- score statistic, $\quad s(\psi) = j_{\text{p}}(\hat{\psi})^{-1/2} d\ell_{\text{p}}(\psi)/d\psi$.

Under suitable regularity conditions on $f(y; \theta)$, all of these have asymptotic standard normal distribution up to the first order. Using any of the above statistics we can approximate the significance function by $\Phi\{w(\psi)\}$, $\Phi\{r(\psi)\}$ or $\Phi\{s(\psi)\}$. When $d_0 > 1$, we may use the quadratic forms of the Wald, likelihood root and score statistics whose finite sample distribution is $\chi^2_{d_0}$ with d_0 degrees of freedom up to the second order. We refer the reader to Chapters 3 and 4 of Severini (2000) for a review of first order likelihood inference.

Although it is common to treat $\ell_{\text{p}}(\psi)$ as if it were an ordinary log likelihood, first order approximations can give poor results, particularly if the dimension of λ is high and the sample size small. An important variant of the likelihood root is the *modified likelihood root*

$$r^*(\psi) = r(\psi) + \frac{1}{r(\psi)} \log\left\{q(\psi)/r(\psi)\right\}, \tag{1}$$

where $q(\psi)$ is a suitable correction term. Expression (1) is a higher order pivot whose finite sample distribution is standard normal up to the third

order. As it was the case for its first order counterpart r, the significance function is approximated by $\Phi\{r^*(\psi)\}$, and there is a version of r^* for multidimensional ψ (Skovgaard, 2001, Section 3.1). More details about the computation of the $q(\psi)$ correction term are given in the Appendix.

It is sometimes useful to decompose the modified likelihood root as

$$r^*(\psi) = r(\psi) + r_{\text{inf}}(\psi) + r_{\text{np}}(\psi),$$

where r_{inf} is the *information adjustment* and r_{np} is the *nuisance parameter adjustment*. The first term accounts for non normality of r, while the second compensates r for the presence of the nuisance parameter λ. Pierce and Peters (1992, Section 3) discuss the behaviour of these two terms in the multiparameter exponential family context. They find that while r_{np} is often appreciable, the information adjustment r_{inf} has typically a minor effect, provided the ψ-specific information $j_{\text{p}}(\hat{\psi})$ is not too small relative to the dimension of λ.

A simple example

Suppose that a sample $y_1, \ldots, y_n$ is available from the Cauchy density

$$f(y;\theta) = \frac{1}{\pi\{1+(y-\theta)^2\}}. \tag{2}$$

The maximum likelihood estimate $\hat{\theta}$ of the unknown location parameter θ is the value which maximises the log likelihood function

$$\ell(\theta;y) = -\sum_{i=1}^{n} \log\{1+(y_i-\theta)^2\}.$$

For $n = 1$, we obtain the exact distribution of $\hat{\theta} = y$ from (2) as $F(\hat{\theta};\theta) = F(y;\theta) = \pi^{-1}\arctan(y-\theta)$.

Assume that $y = 1.32$ was observed. In Figure 1 we compare the exact significance function $p(\theta;y) = \Pr(Y \leq y;\theta)$ (bold line) to the two first order approximations obtained from the Wald statistic

$$w(\theta) = \sqrt{2}(y-\theta),$$

(dotted line), and from the likelihood root

$$r(\theta) = \text{sign}(\hat{\theta}-\theta)\left[2\log\{1+(y-\theta)^2\}\right]^{1/2},$$

(dashed line). We also show the third order approximation $\Phi\{r^*(\theta)\}$ (solid line). Since this is a location model and there is no nuisance parameter, the statistic $q(\theta)$ in (1) is the score statistic

$$s(\theta) = \sqrt{2}(y-\theta)/\{1+(y-\theta)^2\}$$

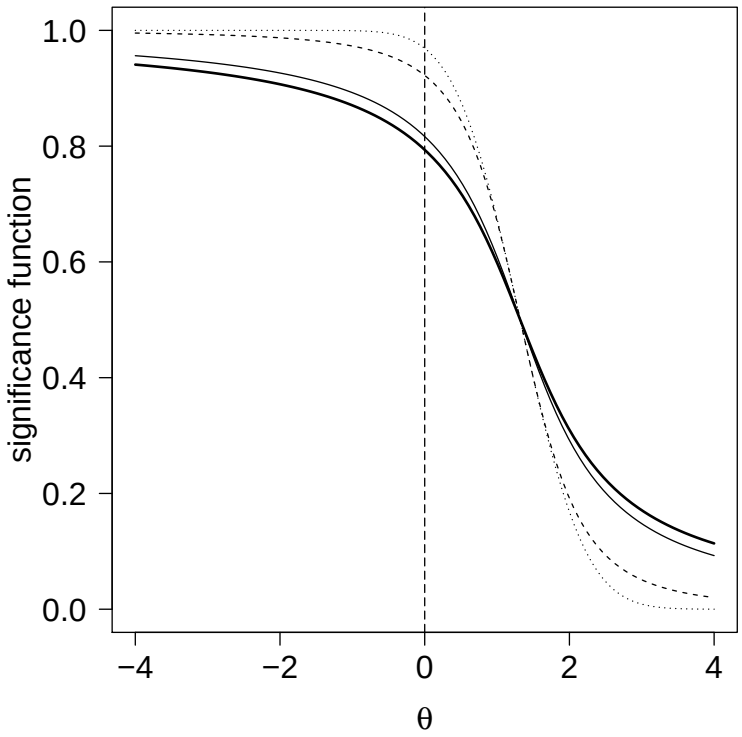

Figure 1: Significance functions for the location parameter of a Cauchy distribution when $y = 1.32$: exact (bold), Wald pivot (dotted), r (dashed) and r^* (solid). The vertical dashed line corresponds to the null hypothesis $\theta = 0$.

(see formula (6) in the Appendix). The code used to generate Figure 1 is given below.

```
> ## likelihood pivots
> wald.stat <- function(theta, y) {
+     sqrt(2) * (y - theta) }
> lik.root <- function(theta, y) {
+     sign(y - theta) * sqrt( 2 * log(1 + (y - theta)^2) ) }
> score.stat <- function(theta, y) {
+     ( sqrt(2) * (y - theta) )/( 1 + (y - theta)^2 ) }
> rstar <- function(theta, y) {
+     lik.root(theta, y) + 1/lik.root(theta, y) *
+         log( score.stat(theta, y)/lik.root(theta, y) ) }
> ## significance functions : Figure 1
> theta.seq <- seq(-4, 4, length = 100)
> par( las = 1, mai = c(0.9, 0.9, 0.2, 0.2) )
> plot( theta.seq, pcauchy( q = 1.32 - theta.seq ), type = "l", lwd = 2,
+       ylim = c(0,1), xlab = expression(theta),
+       ylab = "significance function", cex.lab = 1.5, cex.axis = 1.5 )
> lines( theta.seq, pnorm( wald.stat(theta.seq, 1.32) ), lty = "dotted" )
> lines( theta.seq, pnorm( lik.root(theta.seq, 1.32) ), lty = "dashed" )
> lines( theta.seq, pnorm( rstar(theta.seq, 1.32) ), lty = "solid" )
> abline( v = 0, lty = "longdash" )
```

The vertical dashed line corresponds to the null hypothesis that $\theta = 0$. The exact p-value is

```
> ## exact p-value
> round( 2 * ( min( tp <- pt(1.32, df = 1), 1 - tp ) ), digits = 3 )

[1] 0.413
```

while the first and third order approximations yield

```
> ## Wald pivot p-value
> round( 2 * ( min( tp <- pnorm( wald.stat(0, 1.32) ), 1 - tp ) ),
+        digits = 3 )

[1] 0.062

> ## likelihood root p-value
> round( 2 * ( min( tp <- pnorm( lik.root(0, 1.32) ), 1 - tp ) ),
+        digits = 3 )

[1] 0.155
```

```
> ## modified likelihood root p-value
> round( 2 * ( min( tp <- pnorm( rstar(0, 1.32) ), 1 - tp ) ),
+        digits = 3 )

[1] 0.367
```

respectively for the Wald, likelihood root and modified likelihood root pivot. The r^* statistic is strikingly accurate, while the first order approximations are very poor. This is surprising if we consider that the score function is not monotonic in y and that only one observation is available.

Suppose now that we observed a sample of size $n = 15$ from the Student t distribution with 3 degrees of freedom. It is no longer possible to derive the exact distribution of the maximum likelihood estimator $\hat{\theta}$, but we may use the code provided in the `marg` package of the `hoa` package bundle to compute the p-values for testing the significance of the location parameter.

```
> ## simulated data
> library(marg)
> set.seed(321)
> y <- rt(n = 15, df = 3)
> y.rsm <- rsm(y ~ 1, family = student(3))
> y.cond <- cond(y.rsm, offset = 1)
> summary(y.cond, test = 0)

 Formula:  y ~ 1
 Family:  student
 Offset:  Intercept

          Estimate   Std. Error
uncond.     -0.4208       0.3907
cond.       -0.4065       0.4313

Test statistics
---------------
 hypothesis : Intercept = 0
                           statistic  tail prob.
Wald pivot                   -1.0770      0.1408
Wald pivot (cond.)           -0.9426      0.1729
Likelihood root              -1.0250      0.1528
Modified likelihood root     -0.9277      0.1768

"q" correction term: -0.9277

Diagnostics:
-----------
```

```
   INF     NP
0.2057 0.3291

 Approximation based on 20 points
```

The previous set of instructions yields the p-values 0.282 (Wald), 0.306 (r) and 0.354 (r^*). The difference between first order and higher order approximations is slightly smaller than it was the case before. For this particular model a sample size of $n = 15$ still does not provide enough information on the scalar parameter θ to wipe out completely the effect of higher order corrections.

Higher order asymptotics in R

`hoa` is an R package bundle which implements higher order inference for three widely used model classes: logistic regression, linear non normal models and nonlinear regression with possibly non homogeneous variance. The corresponding code is organised in three packages, namely `cond`, `marg` and `nlreg`. We already saw a (very elementary) application of the `marg` code. The two examples which follow will give a glimpse of the use of the routines in `cond` and `nlreg`. Attention is restricted to the calculation of p-values and confidence intervals, although several routines for accurate point estimation and model checking are also available. The `hoa` bundle includes a fourth package, called `sampling`, which we will not discuss here. It implements a Metropolis-Hastings sampler which can be used to simulate from the conditional distribution of the higher order statistics considered in `marg`.

The `hoa` package bundle is be available on CRAN. More examples of applications, and generally of the use of likelihood asymptotics, are given in Brazzale et al. (to appear).

Example 1: Binary data

Collet (1998) gives a set of binary data on the presence of a sore throat in a sample of 35 patients undergoing surgery during which either of two devices was used to secure the airway.

```
> ## `airway' data
> library(cond)
> head( airway, n = 3 )

  response age sex lubricant duration type
1        0  48   1         0       45    0
2        0  48   1         0       15    0
3        1  39   0         1       40    0
```

In addition to the variable of interest, device `type` (1=tracheal tube or 0=laryngeal mask), there are four further explanatory variables: the `age` of the patient in years, an indicator variable for `sex` (1=male, 0=female), an indicator variable for `lubricant` use (1=yes, 0=no) and the `duration` of the surgery in minutes. The observations form the data frame `airway` which is part of the `hoa` bundle.

A natural starting point for the analysis is a logistic regression model with success probability of the form

$$\Pr(Y = 1; \beta) = \frac{\exp(x^\top \beta)}{1 + \exp(x^\top \beta)},$$

where x represents the explanatory variables associated with the binary `response` Y (1=sore throat and 0=no sore throat). The following set of instructions fits this model to the data with all five explanatory variables included.

```
> ## binomial model fit
> airway.glm <- glm( formula(airway), family = binomial, data = airway )
> summary( airway.glm )

Call:
glm(formula = formula(airway), family = binomial, data = airway)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1888  -0.5633   0.3029   0.7444   1.5954

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.75035    2.08914  -1.316   0.1880
age          0.02245    0.03763   0.597   0.5507
sex1         0.32076    1.01901   0.315   0.7529
lubricant1   0.08448    0.97365   0.087   0.9309
duration     0.07183    0.02956   2.430   0.0151 *
type1       -1.62968    0.94737  -1.720   0.0854 .
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 46.180  on 34  degrees of freedom
Residual deviance: 29.477  on 29  degrees of freedom
AIC: 41.477

Number of Fisher Scoring iterations: 5
```

The coefficient of device `type` is only marginally significant.

As in the previous example we may wonder whether the sample size is large enough to allow us to rely upon first order inference. For the airway data we have $n = 35$ and $p = 5$, so we might expect higher order corrections to the usual approximations to have little effect. We can check this using the routines in the `cond` package.

```
> ## higher order inference
> airway.cond <- cond( airway.glm, offset = type1 )
> summary( airway.cond )    #    produces 95% confidence intervals

 Formula:  response ~ age + sex + lubricant + duration + type
 Family:  binomial
 Offset:  type1

          Estimate   Std. Error
uncond.      -1.630       0.9474
cond.        -1.394       0.8466

Confidence intervals
--------------------
 level = 95 %
                                        lower two-sided  upper
Wald pivot                             -3.486           0.2271
Wald pivot (cond. MLE)                 -3.053           0.2656
Likelihood root                        -3.682           0.1542
Modified likelihood root               -3.130           0.2558
Modified likelihood root (cont. corr.) -3.592           0.5649

Diagnostics:
-----------
    INF      NP
0.05855 0.51426

 Approximation based on 20 points

> plot(airway.cond, which = 4)    # Figure 2
```

As our model is a canonical exponential family, the correction term $q(\psi)$ in (1) involves the Wald statistic $w(\psi)$ plus parts of the observed information matrix (see formula (5) in the Appendix). The 95% confidence intervals obtained from the Wald pivot and from the likelihood root are respectively $(-3.486, 0.227)$ and $(-3.682, 0.154)$. The third order statistic r^* yields a 95% confidence interval of $(-3.130, 0.256)$. First and third order results are rather different, especially with respect to the lower bound. Figure 2 plots the

profiles of the first and third order pivots $w(\psi)$ (dashed line), $r(\psi)$ (solid line) and $r^*(\psi)$ (bold line). The correction term $q(\psi)$ is particularly significant for values of ψ belonging to the lower half of the confidence interval. The nuisance parameter correction is $r_{\text{np}} = 0.51$, while $r_{\text{inf}} = 0.059$ is about ten times smaller.

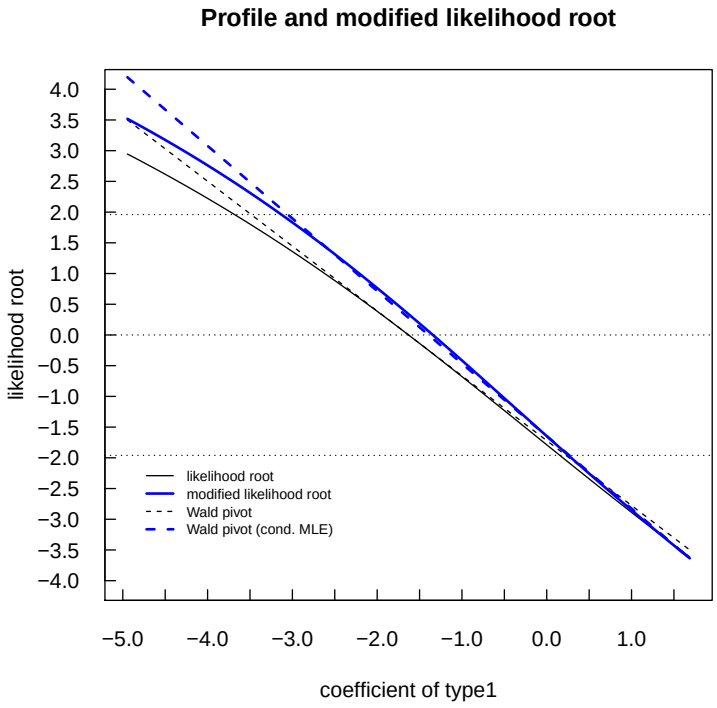

Figure 2: `airway` data analysis: profile plots of the pivots $w(\psi)$ (dashed line), $r(\psi)$ (solid line) and $r^*(\psi)$ (bold line), where ψ is the coefficient of the covariate device `type`.

Example 2: Nonlinear regression

A simple illustration of nonlinear regression is Example 7.7 of Davison and Hinkley (1997), which refers to the `calcium` data of package `boot`. This data set records the calcium uptake (in nmoles/mg) of cells y as a function of time x (in minutes), after being suspended in a solution of radioactive calcium.

```
> ## `calcium' data
> library(boot)
> head( calcium, n = 3 )
```

```
  time      cal
1 0.45  0.34170
2 0.45 -0.00438
3 0.45  0.82531
```

The variables `cal` and `time` represent respectively the calcium uptake and suspension time. There are 27 observations in all. The model is

$$y_i = \beta_0\{1 - \exp(-\beta_1 x_i)\} + \sigma_i \varepsilon_i, \tag{3}$$

where β_0 and β_1 are unknown regression coefficients and the error term $\varepsilon_i \sim N(0,1)$ is standard normal. We complete the definition of model (3) by allowing the response variance $\sigma_i^2 = \sigma^2(1 + x_i)^\gamma$ to depend nonlinearly on the time covariate through the two variance parameters γ and σ^2.

Model (3) is fitted by maximum likelihood using the `nlreg` routine of package `nlreg`.

```
> library(nlreg)
> ## maximum likelihood fit
> calcium.nl <- nlreg( cal ~ b0 * (1 - exp(-b1 * time)),
+                      weights = ~ (1 + time)^g, data = calcium,
+                      start = c(b0 = 4, b1 = 0.1, g = 0) )
> summary( calcium.nl )    # yields estimates and standard errors

differentiating mean function -- may take a while
differentiating variance function -- may take a while
Call:
nlreg(formula = cal ~ b0 * (1 - exp(-b1 * time)), weights = ~(1 +
    time)^g, data = calcium, start = c(b0 = 4, b1 = 0.1, g = 0))

Regression coefficients:
   Estimate Std. Error z value Pr(>|z|)
b0  4.31698    0.32274   13.38  < 2e-16 ***
b1  0.20746    0.03589    5.78 7.47e-09 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Variance parameters:
     Estimate Std. Error
g      0.5364     0.3196
logs  -2.3426     0.6338

No interest parameter

Total number of observations: 27
```

```
Total number of parameters: 4
-2*Log Likelihood 39.31

Algorithm converged in 3 iterations
```

This yields $\hat{\beta}_0 = 4.317$ (s.e. 0.323), $\hat{\beta}_1 = 0.207$ (s.e. 0.036), $\hat{\gamma} = 0.536$ (s.e. 0.320), and $\log \hat{\sigma}^2 = -2.343$ (s.e. 0.634). Note that the baseline variance σ^2 is fitted on the logarithmic scale. This does not affect inference based on the r and r^* statistics, which are parametrisation invariant, and ensures positive values for σ^2 when the Wald statistic is used. The `profile` method of the `nlreg` package can be used to set various first and higher order 95% confidence intervals for the variance parameter γ.

```
> ## pivot profiling for \gamma
> calcium.prof <- profile( calcium.nl, offset = g )

differentiating mean function -- may take a while
differentiating variance function -- may take a while

> summary( calcium.prof )

 Two-sided confidence intervals for g
                   lower upper
r* - Fr (0.95) -0.14270 1.191
r* - Sk (0.95) -0.14250 1.190
r (0.95)       -0.12431 1.154
Wald (0.95)    -0.08992 1.163

  Estimate Std. Error
g   0.5364     0.3196

14 points calculated exactly
50 points used in spline interpolation

INF (Sk): 0.05799
INF (Fr): 0.0699
NP (Sk): 0.1523
NP (Fr): 0.1413
```

A difficulty we had not to face in the previous two examples is that it is no longer possible to calculate the correction term in (1) exactly. The `profile` function implements two slightly different versions of the higher order pivot r^* which we obtain by using the two approximations of $q(\psi)$ discussed in the Appendix. The four statistics agree in letting us question the heterogeneity of the response variance.

Davison and Hinkley (1997, p. 356) consider not only inference on the nonlinear mean function, but also on other aspects of the model such as the "proportion of maximum", $\pi = 1 - \exp(-\beta_1 x)$. For $x = 15$ minutes they give the estimate $\hat{\pi} = 0.956$ and the associated 95% bootstrap confidence interval $(0.83, 0.98)$. We may obtain the corresponding first and higher order likelihood analogues by reparametrizing the mean response curve into (π, β_0) and re-running the whole analysis. This time we assume that the response variance is homogeneous.

```
> ## inference on proportion of maximum
> calcium.nl <- nlreg( cal ~ b0 * (1 - exp(- log(1 + exp(psi)) * time / 15)),
+                      data = calcium, start = c(b0 =4.3, psi =2) )
> calcium.prof <- profile( calcium.nl, offset = psi )

differentiating mean function -- may take a while
differentiating variance function -- may take a while

> calcium.sum <- summary( calcium.prof )
> exp(calcium.sum$CI) / (1 + exp(calcium.sum$CI))    # 95% confidence intervals for

                    lower     upper
r* - Fr (0.95) 0.8748270 0.9897534
r* - Sk (0.95) 0.8715957 0.9892020
r (0.95)       0.8777616 0.9882782
Wald (0.95)    0.8728598 0.9857717
```

Because of the constraint that π must lie in the interval $(0, 1)$, we actually fit the model for $\psi = \log\{\pi/(1-\pi)\}$ and back-transform to the original scale by $\pi = \exp(\psi)/\{1 + \exp(\psi)\}$. This yields the intervals $(0.87, 0.99)$ and $(0.88, 0.99)$ for respectively the Wald and likelihood root statistics and $(0.87, 0.99)$ for both versions of r^*, which is in agreement with the bootstrap simulation.

The `profile` method of the `nlreg` package provides also all elements needed to display graphically a fitted nonlinear model.

```
> ## profile and contour plots : Figure 3
> calcium.prof <- profile( calcium.nl )

long calculation --- may take a while

differentiating mean function -- may take a while
differentiating variance function -- may take a while

> par( las = 1, mai = c(0.5, 0.5, 0.2, 0.2) )
> contour( calcium.prof, alpha = 0.05, cl1 = "black", cl2 = "black",
+          lwd2 = 2 )
```

```
Higher order method used: Skovgaard's r*
```

The result is Figure 3. The `contour` method of the `nlreg` package represents, in fact, an enhanced version of the original algorithm by Bates and Watts (1988, Chapter 6), to which we refer the reader for the interpretation of these plots. The dashed, solid and bold lines represent respectively the Wald pivot, the likelihood root r and Skovgaard's (1996) approximation to the r^* statistic (see the Appendix). The bivariate contour plots in the lower triangle are plotted on the original scale, whereas the ones in the upper triangle are on the r scale. Figure 3 highlights different aspects of the model fit. First, the maximum likelihood estimate of $\log \sigma^2$ is biased downwards, which we can tell from the fact the corresponding r^* profile is shifted to the right of r. Otherwise, there does not seem to be a huge difference between first and higher order methods as the corresponding profiles and contours are not too different. The finite sample estimates of β_0 and ψ are strongly correlated, while they are almost independent of $\log \hat{\sigma}^2$. The contours of $r(\psi)$ and $r^*(\psi)$ are close to elliptic which indicates that the log likelihood function is not too far from being quadratic. A further indication for a small curvature effect due to parametrisation is that the contours on the original and on the r scale look similar.

Acknowledgments

I am in debt with Ruggero Bellio, Anthony Davison and Nancy Reid for the many helpful discussions on and around higher order likelihood asymptotics. I would like to thank two anonymous referees whose suggestions and comments helped improving a previous version of this paper.

Appendix: $q(\psi)$ correction term

In this appendix we give the general expression of the correction term $q(\psi)$ in (1) and the explicit formulae for two special model classes, that is, linear exponential families and regression-scale models. We will furthermore discuss two ways of approximating $q(\psi)$ in case we cannot calculate it explicitly.

Basic expression

Let $\ell(\theta) = \ell(\psi, \lambda)$ be the log likelihood function, $\hat{\theta} = (\hat{\psi}, \hat{\lambda})$ the maximum likelihood estimator of the d-dimensional parameter $\theta = (\psi, \lambda)$, and $j(\theta) = -\partial^2 \ell(\theta)/\partial\theta\partial\theta^\top$ the $d \times d$ observed information matrix. Denote $\hat{\lambda}_\psi$ the constrained maximum likelihood estimator of the nuisance parameter λ given the value of the scalar parameter of interest ψ. Write $j_{\lambda\lambda}(\theta)$ the corner of $j(\theta) = j(\psi, \lambda)$ which corresponds to λ, and $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$.

```
Higher order method used: Skovgaard's r*
```

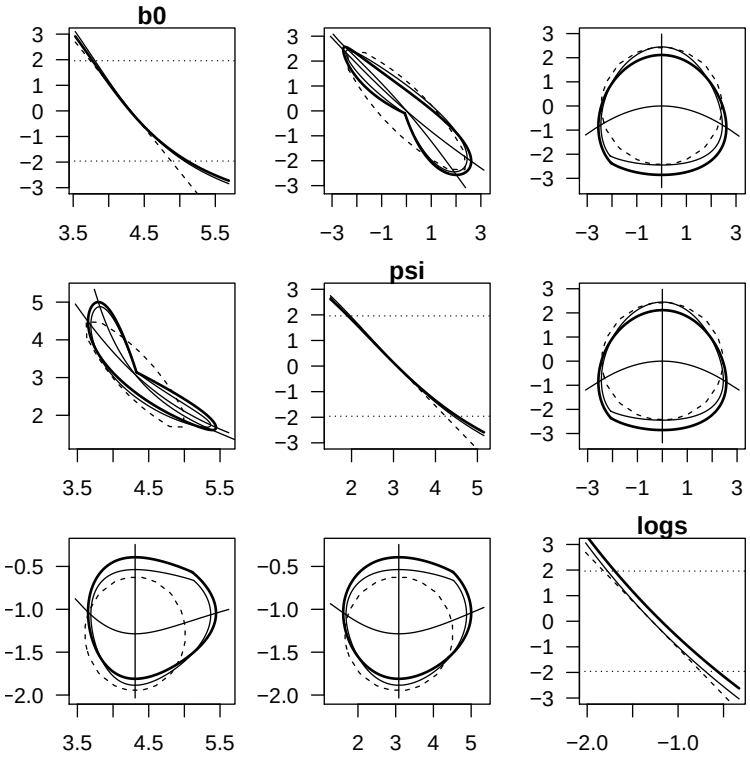

Figure 3: `calcium` uptake data analysis: profile plots and profile pair sketches for the parameters β_0, ψ and $\log \sigma^2$ using the Wald statistic (dashed), the likelihood root r (solid) and Skovgaard's (1996) approximation to r^* (bold).

The basic expression for $q(\psi)$ is

$$q(\psi) = \frac{|\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\hat{\theta}}(\hat{\theta}_\psi) \quad \ell_{\lambda^\top;\hat{\theta}}(\hat{\theta}_\psi)|}{\left\{|j_{\lambda\lambda}(\hat{\theta}_\psi)||j(\hat{\theta})|\right\}^{1/2}}, \tag{4}$$

where $|\cdot|$ indicates determinant (Severini, 2000, Section 7.4.1). The $d \times d$ matrix appearing in the numerator of $q(\psi)$ consists of a column vector formed using so-called *sample space derivatives*

$$\ell_{;\hat{\theta}}(\theta) = \frac{\partial \ell(\theta; \hat{\theta}|a)}{\partial \hat{\theta}},$$

and a $d \times (d-1)$ matrix of *mixed derivatives*

$$\ell_{\lambda^\top;\hat{\theta}} = \frac{\partial^2 \ell(\psi, \lambda; \hat{\theta}|a)}{\partial \lambda^\top \partial \hat{\theta}}.$$

The former are defined as the derivatives of the log likelihood function $\ell(\psi, \lambda; \hat{\theta}|a)$ with respect to the maximum likelihood estimator $\hat{\theta}$; mixed derivatives furthermore involve differentiation with respect to the whole parameter θ or parts of it (Severini, 2000, Section 6.2.1). Note that to do so, the data vector has to be re-expressed as $y = (\hat{\theta}, a)$, where a represents the observed value of an ancillary statistic upon which we condition.

Approximations

Exact computation of the sample space derivatives involved in expression (4) requires that we are able to write the data vector y as a function of the maximum likelihood estimator $\hat{\theta}$ and of an ancillary statistic a. This is, with few exceptions, only feasible for linear exponential families and transformation models, in which cases the $q(\psi)$ term involves familiar likelihood quantities. If the reference model is a full rank exponential family with ψ and λ taken as canonical parameters, the correction term

$$q(\psi) = w(\psi) \left\{|j_{\lambda\lambda}(\hat{\theta})|/|j_{\lambda\lambda}(\hat{\theta}_\psi)|\right\}^{1/2} \tag{5}$$

depends upon the Wald statistic. In case of a regression-scale model, that is, of a linear regression model with non necessarily normal errors,

$$q(\psi) = s(\psi) \left\{|j_{\lambda\lambda}(\hat{\theta}_\psi)|/|j_{\lambda\lambda}(\hat{\theta})|\right\}^{1/2} \tag{6}$$

involves the score statistic. Here, ψ is linear in (β, σ) and the nuisance parameter λ is taken linear in β and $\xi = \log \sigma$, where β and σ represent respectively the regression coefficients and the scale parameter.

In general, the calculation of the sample space derivatives $\ell_{;\hat{\theta}}(\theta)$ and mixed derivatives $\ell_{\lambda^\top;\hat{\theta}}(\theta)$ may be difficult or impossible. To deal with this difficulty, several approximations were proposed. For a comprehensive review we refer the reader to Section 6.7 of Severini (2000). Here we will mention two of them. A first approximation, due to Fraser et al. (1999), is based upon the idea that in order to differentiate the likelihood function $\ell(\theta;\hat{\theta}|a)$ on the surface in the n-dimensional sample space defined by conditioning on a we need not know exactly the transformation from y to $(\hat{\theta}, a)$, but only the d vectors which are tangent to this surface (Severini, 2000, Section 6.7.2). Skovgaard (1996) on the other hand suggests to approximate the sample space and mixed derivatives by suitable covariances of the log likelihood and of the score vector (Severini, 2000, Section 6.7.3). While the first approximation maintains the third order accuracy of r^*, we lose one degree when following Skovgaard's (1996) approach. See Sections 7.5.3 and 7.5.4 of Severini (2000) for the details.

The `hoa` package

The expressions of $q(\psi)$ implemented in the `hoa` package bundle are: i) (5) and (6) for respectively the `cond` and `marg` packages (logistic and linear non normal regression), and ii) the two approximations discussed above for the `nlreg` package (nonlinear heteroscedastic regression). The formulae are given in Brazzale et al. (to appear). The `nlreg` package also implements Skovgaard's (2001, Section 3.1) multiparameter extension of the modified likelihood root. The implementation of the `cond` and `nlreg` packages is discussed in Brazzale (1999) and Bellio and Brazzale (2003).

References

D. M. Bates and D. G. Watts. *Nonlinear Regression Analysis and Its Applications*. Wiley, New York, 1988.

R. Bellio and A. R. Brazzale. Higher-order asymptotics unleashed: Software design for nonlinear heteroscedastic models. *Journal of Computational and Graphical Statistics*, 12:682–697, 2003.

A. R. Brazzale. Approximate conditional inference in logistic and loglinear models. *Journal of Computational and Graphical Statistics*, 8:653–661, 1999.

A. R. Brazzale, A. C. Davison, and N. Reid. *Applied Asymptotics*. Cambridge University Press, Cambridge, to appear.

D. Collet. Binary data. In P. Armitage and T. Colton, editors, *Encyclopedia of Biostatistics*. John Wiley & Sons, Chichester, 1998.

A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge, 1997.

D. A. S. Fraser. Statistical inference: Likelihood to significance. *Journal of the American Statistical Association*, 86:258–265, 1991.

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

D. A. Pierce and D. Peters. Practical use of higher-order asymptotics for multiparameter exponential families. *Journal of the Royal Statistical Society Series B*, 54:701–737, 1992.

T. A. Severini. *Likelihood Methods in Statistics*. Oxford University Press, Oxford, 2000.

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

I. M. Skovgaard. Likelihood asymptotics. *Scandinavian Journal of Statistics*, 28:3–32, 2001.

Alessandra R. Brazzale
Institute of Biomedical Engineering, Italian National Research Council
`alessandra.brazzale@isib.cnr.it`