Biometrika (2008), **94**, 2, *pp*. 1–7

Advance Access publication on 31 July 2008
Printed in Great Britain

On assessing vector valued parameters

BY A. C. DAVISON
Institute of Mathematics, Ecole Polytechnique Fédérale de Lausanne, Station 8, 1015 Lausanne, Switzerland
Anthony.Davison@epfl.ch

D. A. S. FRASER AND N. REID
Department of Statistics, University of Toronto, Toronto, Canada M5S 3G3
dfraser@utstat.toronto.edu reid@utstat.toronto.edu

N. SARTORI
Departimendo di Statistica, University of Venice, 30121 Venice, Italy
sartori@unive.it

1. INTRODUCTION

We consider the assessment of a vector valued parameter for a regular statistical model with data. The approach involves the Taylor expansion of the log-model, the separation of terms by asymptotic accuracy $O(1)$, $O(n^{-1/2})$, and $O(n^{-1})$, and the subsequent recombination of terms in a coordinate-free form. Our particular interest centers on vector valued parameters in the presence of nuisance parameters, with special concern for discrete data.

The model $f(y;\theta)$ with observed data y^0 leads immediately to the observed likelihood $L(\theta) = cf(y^0;\theta)$ which can often be examined directly. In turn the log-likelihood $\ell(\theta) = \log L(\theta)$ gives the maximum likelihood value $\hat{\theta}$ and observed information $\hat{\jmath}_{\theta\theta} = -(\partial^2/\partial\theta\partial\theta')\ell(\theta)|_{\hat{\theta}}$ recording curvature, a second-derivative value or matrix at the maximum $\ell(\hat{\theta})$.

The Taylor expansion of the model as demonstrated below gives an approximate Normal distribution $(\theta; \hat{\jmath}_{\theta\theta}^{-1})$ for the maximizing value $\hat{\theta}$. The departure of $\hat{\theta}$ from some hypothesis $H = \{\psi(\theta) = \psi\}$ can be put in a conventional coordinate-free form as twice the drop in the likelihood from the overall maximum $\ell(\hat{\theta})$ to the constrained maximum $\ell(\hat{\theta}_\psi)$,

$$r_\psi^2 = 2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_\psi)\},$$

where the constrained $\hat{\theta}_\psi$ gives the maximum subject to $\psi(\theta) = \psi$. The first-order reference distribution for r_ψ^2 is the chi-square distribution with degrees of freedom given as the number of constrained parameter coordinates in $\psi(\theta) = \psi$.

Our focus is on the higher-order separation of information concerning component parameters of interest, free of related nuisance parameters, and using directed measures of departure.

First consider the case where the variable y and the parameter θ have the same dimension and suppose the model has asymptotic properties so that $\ell(\theta; y) = \log f(y;\theta)$ has $O(n)$ dependence on some antecedent sample size. This can arise if the model is conditional or marginal or a mix of such from some larger background model with increasing data size n. We consider the Taylor expansion of the log-model $\ell(\theta; y)$ in terms of both parameter θ and variable y about the observed $(\hat{\theta}^0, y^0)$ using centered and scaled coordinates.

We begin by examining the scalar y and scalar θ case. Let a new θ designate departure of the initial θ from $\hat{\theta}^0$ as standardized by observed information: $(\theta - \hat{\theta}^0)\hat{\jmath}_{\theta\theta}^{1/2}$. And then let a new y designate departure of the initial y from y^0 scaled so that the new cross Hessian $\partial^2\ell(\theta;y)/\partial\theta\partial y = 1$ at the expansion point $(\hat{\theta}^0, y^0)$. This gives the following form to the model neglecting terms of order $O(n^{-1/2})$,

$$f_1(y;\theta) = \frac{1}{(2\pi)^{1/2}} \exp\left\{-\frac{\theta^2}{2} + y\theta - \frac{y^2}{2}\right\}; \tag{1}$$

the coefficient of the y^2 term follows from the norming property. The observed p-value function $p(\theta) = F_1(y^0;\theta)$ can then be written $p(\theta) = \Phi\{r(\theta)\}$ where

$$r(\theta) = \operatorname{sgn}(\hat{\theta}^0 - \theta)[2\{\ell(\hat{\theta}^0) - \ell(\theta)\}]^{1/2} \tag{2}$$

is the signed likelihood root.

We can however further expand the log model to include cubic terms and then neglect terms of order $O(n^{-1})$. And to get a familiar form for the model we reexpress the variable and reexpress the parameter in accord with exponential model form thus eliminating the $\theta^2 y/2n^{1/2}$ and $\theta y^2/2n^{1/2}$ terms; this gives

$$f_2(y;\theta) = f_1(y;\theta)\exp\left\{-\alpha_3\frac{\theta^3}{6n^{1/2}} + \alpha_3\frac{y^3}{6n^{1/2}} - \alpha_3\frac{y}{2n^{1/2}}\right\}$$

where α_3 is a mathematical parameter partly describing observed likelihood and the coefficients of y and y^3 are determined by the norming constant using $E\{y^3 - 3y - \theta^3;\theta\} = 0$ for the Normal $(\theta;1)$ distribution indicated by $f_1(y;\theta)$; for some related detail see Cakmak *et al* (1998) and Andrews *et al* (2005).

Now suppose we expand the model to include quartic terms and then neglect terms of order $O(n^{-3/2})$; and suppose we also reexpress the parameter and reexpress the variable towards exponential model form; this reexpression forces the coefficients of $\theta^2 y/2n^{1/2}$, $\theta^3 y/6n$, $\theta y^2/2n^{1/2}$ and $\theta y^3/6n$ to be zero but cannot do so for the quadratic-quadratic term $\gamma\theta^2 y^2/4n$ where γ is a coefficient that reflects non-exponentiality in the model form. If $\gamma = 0$, then the model is exponential to order $O(n^{-3/2})$ and can be written

$$f_3(y;\theta) = f_2(y;\theta)\exp\left\{-\alpha_4\frac{\theta^4}{24n} + (\alpha_4 - 3\alpha_3^2)\frac{y^4}{24n} - (\alpha_4 - 2\alpha_3^2)\frac{y^2}{4n} + (3\alpha_4 - 5\alpha_3^2)\frac{1}{24n}\right\};$$

where α_4 is a mathematical parameter also describing an aspect of observed likelihood. The coefficients and the norming constant are available from Normal $(\theta, 1)$ integrals; see Andrews et al (2005).

As an exponential model, the observed p-value $p(\theta) = F_3(y^0;\theta)$ is available from Lugannani & Rice (1980), Barndorff-Nielson (1986), Daniels (1987) or by Normal$(\theta, 1)$ integrals in Andrews et al (2005). It can be written

$$p(\theta) = \Phi\left(r - r^{-1}\log\frac{r}{q}\right) \tag{3}$$

where r is the signed likelihood root (SLR) and q is the standardized maximum likelihood departure on a scale provided by the canonical parameter of the exponential model,

$$\varphi(\theta) = \frac{\partial}{\partial y}\log f(y;\theta)|_{y^0}; \tag{4}$$

thus

$$q = \{\varphi(\hat{\theta}^0) - \varphi(\theta)\} j_{\varphi\varphi}^{1/2}. \tag{5}$$

The full $O(n^{-3/2})$ expansion of the log-model as $\log f_3(y;\theta) = \sum a_{ij}\theta^i y^j / i!j!$ can be summarized by recording the matrix array $A = \{a_{ij}\}$ of coefficients to order $O(n^{-3/2})$ including contributions from the term $\gamma\theta^2 y^2/4n$ omitted above. The norming property gives all terms by using $E\{y^2\theta^2 - y^4 + 5y^2 - 2; \theta\} = 0$ for the Normal $(\theta; 1)$ distribution:

$$A = \begin{pmatrix} a + \frac{3\alpha_4 - 5\alpha_3^2 - 12\gamma}{24n} & \frac{-\alpha_3}{2n^{1/2}} & -\{1 + \frac{\alpha_4 - 2\alpha_3^2 - 5\gamma}{2n}\} & \frac{\alpha_3}{n^{1/2}} & \frac{\alpha_4 - 3\alpha_3^2 - 6\gamma}{24n} \\ 0 & 1 & 0 & 0 & - \\ -1 & 0 & \gamma/n & - & - \\ -\frac{\alpha_3}{n^{1/2}} & 0 & - & - & - \\ -\frac{\alpha_4}{n} & - & - & - & - \end{pmatrix}$$

where $a = -\frac{1}{2}\log(2\pi)$ and omitted terms are $O(n^{-3/2})$.

The observed p-value $p(\theta) = F_3(y^0; \theta)$ for the general case that includes γ is immediately available from the same expression (3) above, as a consequence of the extraordinary fact that the p-value does not depend on γ to the third order; this can be verified by direct computer integration (Andrews et al, 2005) or direct integration by parts although it is implicit in many of the references cited above.

If the general density $f_4(y;\theta)$ acquires an appended factor $a(y;\theta)$ that is constant to first order and comes say from the elimination of nuisance parameters, then the observed distribution function is again given to third order by (3) but with the following modification of (5),

$$q = \{\varphi(\hat{\theta}^0) - \varphi(\theta)\} j_{\varphi\varphi}^{1/2} \frac{a(y^0; \hat{\theta}^0)}{a(y^0; \theta)};$$

for details, see Cheah et al (1995).

The third order p-value (3) uses only $\ell(\theta)$ and $\varphi(\theta)$, the observed likelihood and observed likelihood gradient. One of the possible forms for the initial model is the exponential which can be written

$$f(y;\theta)dy = c\exp\{\ell(\theta) + s'(\varphi(\theta) - \hat{\varphi}^0(\theta))\}h(s)ds \tag{6}$$

and then rewritten (Daniels, 1954) in the saddlepoint form as

$$f(y;\theta)dy = \frac{e^{k/n}}{(2\pi)^{p/2}} \exp\left\{-\frac{r^2(s;\theta)}{2}\right\} |\hat{\jmath}_{\varphi\varphi}(s)|^{1/2} ds \tag{7}$$

with observed value say $s^0 = 0$, where s designates a score variable and k is constant to third order. We have written (7) for the vector y and θ case of dimension p although the preceding discussion focused on the scalar y and θ case. As (6) and (7) involve only likelihood and first derivative likelihood at the data, the exponential approximation is reasonably called the tangent exponential model at the observed data.

In Section 2 for an interest parameter $\psi(\theta)$ we record the various levels of accuracy that are available for the observed likelihood $\ell(\psi)$ and also for the observed likelihood gradient $\varphi(\psi)$, depending on the available model information. In Section 3 we show how the related density function can be determined on a score line that joins the expected score say s_0 under the hypothesis $\psi(\theta) = \psi$ with the observed score s^0, and then how a directed p-value for testing $\psi(\theta) = \psi$ can be determined. Examples including discrete data models are examined in Section 4; Section 5 presents some discussion.

2. ACCURACY OF LIKELIHOOD AND LIKELIHOOD GRADIENT.

(i) Accuracy of p-values. In Section 1 we have indicated how p-values for a scalar parameter can be obtained from $\ell(\theta)$ and $\varphi(\theta)$ and thus from the nominal exponential model $f(s;\theta)ds = \exp\{\ell(\theta) + s'\varphi(\theta)\}h(s)ds$, with data $s^0 = 0$. The accuracy of the p-value is the lesser of the accuracies of $\ell(\theta)$ and $\varphi(\theta)$ and functions such as $\hat{\varphi}(s)$ and $\hat{\jmath}(s)$ are defined as functions of the score variable s of the nominal exponential model..

(ii) Accuracy concerning the full parameter θ. The observed likelihood $\ell(\theta) = \ell(\theta; y^0)$ is obtained by substituting y^0 in the log-model $\log f(y;\theta)$ and would then typically be available with full accuracy. For discrete models however we will use a continuous approximation to the model; the usual third order accuracy for the continuous model makes available just $O(n^{-1})$ accuracy for the discrete model; for some discussion see Davison et al (2006).

The observed likelihood gradient $\varphi_3(\theta) = (d/dV)\ell(\theta; y)|_{y^0}$ is the gradient of likelihood in critically chosen directions V, known to be tangent to a second order ancillary (Fraser et al, 1999; Fraser & Reid, 1995; Fraser et al, 2009). The individual directions recorded in V provide exact first derivative ancillary directions for one-dimensional departures from the observed maximum likelihood value (Fraser et al, 2009), but can often be difficult to calculate or determine. The resulting $\varphi_3(\theta)$ however does lead to third order accuracy for inference concerning a scalar parameter.

Several alternatives are available that lead to a $\varphi_2(\theta)$ having second order accuracy. Barndorff-Nielsen (1986) works within a finite dimensional conditioned model and then further conditions iteratively on likelihood ratio values within an embedding exponential model; this gives direction $\tilde{V}_{BN}$. Fraser & Reid (1998, 2001) and Davison et al (2006) replace the given variable y by the score variable $s = \ell_{;\theta}(\hat{\theta}^0; y)$ and examine directions $\tilde{V}_S = (d/d\theta)E\{s;\theta\}|_{\hat{\theta}^0}$.

Skovgaard (1996) estimates directions for conditioning and derives p-values for scalar parameters. Reid & Fraser (2009) note that the resulting Skovgaard directions depend on just the maximum likelihood value and thus in effect can be obtained from the marginal model $f(\hat{\theta};\theta)$ for $\hat{\theta}$. In this latter case a second order accurate version of $\varphi(\theta)$ can be obtained as $\varphi_2(\theta) = (\partial/\partial\hat{\theta}) \log f(\hat{\theta};\theta)|_{\hat{\theta}=\hat{\theta}^0}$; and a variation on this is given as

$$\tilde{\varphi}_2(\theta) = \frac{\partial}{\partial\theta_0} E\{\log f(y;\theta);\theta_0\}|_{\theta_0=\hat{\theta}^0}. \tag{8}$$

The use of $\tilde{\varphi}_2(\theta)$ reproduces Skovgarrd's second order p-values; in effect the calculation of $\varphi(\theta)$ replaces an observed log-likelihood by an observed average likelihood,

$$E\{\log f(y;\theta);\theta_0\}_{\theta_0=\hat{\theta}^0},$$

which is called an information by Kent (1982).

(iii) Accuracy concerning a component $\psi(\theta)$. Now consider inference for a component parameter of dimension d. A first-order accurate likelihood is available immediately as the profile

$$\ell_1(\psi) = \sup_{\lambda} \ell(\psi,\lambda) = \ell(\psi,\hat{\lambda}_\psi) = \ell(\psi).$$

where $\hat{\lambda}_\psi$ is the constrained maximum given a complementary nuisance parameter λ.

A second-order accurate version of marginal likelihood is available (Fraser, 2003) using a nuisance parameter correction based on a second order accurate likelihood gradient $\varphi(\theta)$:

$$\ell_2(\psi) = \ell(\psi,\hat{\lambda}_\psi) + \frac{1}{2}\log|\hat{\jmath}_{(\lambda\lambda)}(\hat{\theta}_\psi)|$$

where $\hat{j}_{(\lambda\lambda)}(\hat{\theta}_\psi)$ is the estimated nuisance information $\hat{j}_{\lambda\lambda}(\hat{\theta}_\psi)$ but rescaled to the $\varphi(\theta)$ parameterization giving

$$|\hat{j}_{(\lambda\lambda)}(\hat{\theta}_\psi)| = |\hat{j}_{\lambda\lambda}(\hat{\theta}_\psi)| \cdot |\varphi_\lambda(\hat{\theta}_\psi)|^{-2}$$

where $|X| = |X'X|^{1/2}$ uses the $p \times d$ array of partial derivatives

$$X = \frac{\partial\varphi(\psi,\lambda)}{\partial\lambda}\Big|_{\lambda=\hat{\lambda}_\psi}.$$

A third-order accurate likelihood can be obtained by using a third order $\varphi(\theta)$ and making the result sample space invariant; this is achieved by using a third order version of $\varphi(\theta)$ that has observed information equal to the identity,

$$\bar{\varphi}(\theta) = \hat{j}_{\varphi\varphi}^{1/2}\varphi(\theta), \tag{9}$$

where $\hat{j}_{\varphi\varphi}^{1/2}$ is a right square root of $\hat{j}_{\varphi\varphi} = (\hat{j}_{\varphi\varphi}^{1/2})'\hat{j}_{\varphi\varphi}^{1/2}$. This gives

$$\ell_3^*(\psi) = \ell(\psi, \hat{\lambda}_\psi) + \frac{1}{2}\log|\hat{j}_{(\bar{\lambda}\bar{\lambda})}(\hat{\theta}_\psi)|$$

where

$$|\hat{j}_{(\bar{\lambda}\bar{\lambda})}(\hat{\theta}_\psi)| = |\hat{j}_{\lambda\lambda}(\hat{\theta}_\psi)| \cdot |\varphi'_\lambda(\hat{\theta}_\psi)\hat{j}_{\varphi\varphi}\varphi_\lambda(\hat{\theta}_\psi)|^{-1}.$$

The likelihood gradient parameterization to use with the component $\psi(\theta)$ is available as rotated coordinates of $\bar{\varphi}(\theta)$ perpendicular to the $\lambda = \hat{\lambda}_\psi$ contour at $\bar{\varphi}(\hat{\theta}_\psi)$. The likelihood is modified by the transformation since it is calculated typically in different subspaces for different ψ values Fraser (2003).

3. ASSESSING A VECTOR INTEREST PARAMETER

Consider a d-dimensional interest parameter $\psi(\theta)$, and suppose that a corresponding observed log-likelihood $\ell(\psi)$ is chosen and an observed log-likelihood gradient $\varphi(\psi)$ is chosen, each with specified accuracies as described in the preceding section. To be of particular interest here the accuracies would need to be second order or higher. But both of third order is possible in some situations, and just one of third order can have advantages in removing certain biases.

We assess the parameter $\psi(\theta) = \psi_0$ in the context of the exponential model based on $\{\ell(\psi), \varphi(\psi)\}$ using a nominal canonical variable s of dimension d and observed value $s^0 = 0$; from (6) and (7) we then have

$$f(s;\psi)ds = c\exp\{\ell(\psi) + s'\varphi(\psi)\}h(s)ds \tag{10}$$

$$= \frac{e^{k/n}}{(2\pi)^{d/2}}\exp\left\{-\frac{r^2(\psi;s)}{2}\right\}|\hat{j}_{\varphi\varphi}(s)|^{1/2}ds, \tag{11}$$

with subsequent third-order accuracy relative to the model (10)using the observed value $s^0 = 0$. The specified value $\psi = \psi_0$ has a corresponding mean score value $s_0 = E\{s;\psi_0\} = -\ell_\varphi(\psi_0)$. The line $\mathcal{L}^+(s^0 - s_0)$ radiating from s_0 to and beyond the observed s^0 makes available a vector space departure of data s^0 from expectation s_0, and defines a conditional p-value using the related conditional distribution. This gives the directional test of ψ_0 following Fraser & Massam (1985), Skovgaard (1988), Cheah et al (1994) and Fraser & Reid (2006):

$$p(\psi_0) = \frac{\int_0^1 t^{d-1} h\{s_0 + t(s^0 - s_0)\} dt}{\int_0^\infty t^{d-1} h\{s_0 + t(s^0 - s_0)\} dt} \tag{12}$$

where $h(s)$ is the density (10) and (11) with $\psi = \psi_0$. The preceding references focus on approximations of the type (3), but here we will work with direct numerical integration of the scalar variable integrals in (12); this bypasses various instabilities that arise with the approximation formulas coming from the Jacobian factor t^{d-1} when the dimension d is greater than one.

The density function $h(s)$ in the integrals for the p-value (12) is available from (11) with third-order relative accuracy: computationally, all that is needed is the maximum likelihood value $\hat{\varphi}(s)$ and the observed information value $\hat{\jmath}_{\varphi\varphi}(s)$ as calculated from the log-model $\ell(\psi, s) = \ell(\psi) + s'\varphi(\psi)$ using the available $\{\ell(\psi), \varphi(\psi)\}$; and the values $\hat{\varphi}(s)$ and $\hat{\jmath}_{\varphi\varphi}(s)$ are needed only for points s on the line $s^0 + \mathcal{L}(s_0 - s^0)$. If we take $a = (a_1, \ldots, a_d)'$ to be orthogonal to $s_0 - s^0$ and for convenience, of unit length, then the scalar parameter $X(\psi) = a'\varphi(\psi)$ provides the basic tilting and the corresponding profile likelihood for the computation.

The directional test can reveal important aspects of the departure of data from the value $\psi = \psi_0$ for a vector parameter. In particular the upper limit for t in the p-value (12) may be less than infinity, as forced by the density $h\{s_0 + t(s^0 - s_0)\}$. This can arise with continuous and with discrete models and would be important additional information to record alongside a likelihood ratio $r^2(\psi, y)$, or alongside the related chi-square p-value; see Examples .... In addition a likelihood ratio value may be associated with a longer tail or a shorter tail, important additional information that would only emerge as part of the p-value calculation (12); this would typically be of order $O(n^{-1/2})$ and inappropriately would be removed by averaging the conditional distribution of the likelihood ratio itself. Similar aspects arise with the Bartlett correction when varying types of departure are hidden by the key marginalization step.

4. EXAMPLES

Some possibilities

1. Simple continuous exponential model with say 2 coordinates and a finite boundary in one or two directions.
2. A continuous generalized exponential model with a non canonical link function.
3. A discrete example, possibly from Davison et al (2006), where the directional value could be compared with the usual p-value for some scalar parameter.
4. Contingency tables, testing say independence or symmetry.

REFERENCES

Andrews, D.F., Fraser, D.A.S. and Wong, A. (2005). Computation of distribution functions from likelihood information near observed data. *Journal Statist. Plann. Inference* **134**, 180-193.

Barndorff-Nielsen, O.E. (1986). Inference on full or partial parameters based on the standardised signed log likelihood ratio. *Biometrika* **73**, 307-22.

Cakmak, S., Fraser, D.A.S., McDunnough, P., Reid, N., and Yuan, X. (1998). Likelihood centered asymptotic model exponential and location model versions. *J. Statist. Planning and Inference* **66**, 211-222.

Cheah, P.K., Fraser, D.A.S., and Reid, N. (1994). Multiparameter testing in exponential models: third order approximations from likelihood. *Biometrika* **81**, 271-278.

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

Daniels, H. E. (1954). Saddlepoint approximation in statistics. *Ann. Math. Statist.*, **25**, 631-650.

Daniels, H. E. (1987). Tail probability approximations. *Int. Statist. Rev.*, **54**, 37-48.

Davison, A.C., Fraser, D.A.S. and Reid, N. (2006). Improved likelihood inference for discrete data. *J. Royal Statist. Soc.*, **B 68**, 495-508.

Fraser, D.A.S. (2003). Likelihood for component parameters. *Biometrika* **90**, 327-339.

Fraser, A.M., Fraser, D.A.S. and Staicu, A.-M. (2009). The second order ancillary: A differential view with continuity. *Bernoulli*, in revision.

Fraser, D. A. S. and Massam, H. (1985). Conical tests: Observed levels of significance and confidence regions. *Statistische Hefte* **26**, 1-17.

Fraser, D.A.S., and Reid, N. (1995). Ancillaries and third order significance. *Utilitas Mathematica*, **47**, 33-53.

Fraser, D.A.S., and Reid, N. (2001). Ancillary information for statistical inference. In S.E. Ahmed and N. Reid (Eds), *Empirical Bayes and Likelihood Inference*, 185-209. New York: Springer-Verlag.

Fraser, D. A. S. and Reid, N. (2006). Assessing a vector parameter. *Student*, **5**, 247-256.

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

Kent, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika*, **69**, 19-27.

Lugannani, R. and Rice, S. (1980), Saddlepoint approximation for the distribution of the sum of independent random variables, *Advances in Applied Probability* **12**, 475–490.

Reid, N, and Fraser, D.A.S. (2009) Mean likelihood and higher order inference. *Biometrika*; in review.

Skovgaard, Ib M. (1988). Saddlepoint expansions for directional test probabilities. *J. R. Statist. Soc.* B **50**, 269-280.

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

[*Received January* 2008. *Revised March* 2009]