

# Does the “Sensitive Direction” point to the Likelihood Ancillary?

Ronald W. Butler

Colorado State University and University of Sydney

September 2, 2003

## 1 Introduction

Can we now put to rest the unthinking and unqualified use of “global repeated sampling properties” as a means for probability computation and inference? Professor Fraser has forcefully and eloquently stated the case against the use of this principle when the model structure would suggest otherwise. In §7 paragraph 3 he concedes that “On the surface this (principle) seems hard to argue against...”. However, after a careful reading of this paper, one must conclude from the multitude of examples and discussion that the unconditional and blanket use of this principle is seriously flawed. There are many modeling situations which would qualify for its use, particularly nonparametric modeling settings, however the models presented here clearly do not.

My comments are divided into two parts. First some consideration of what these ideas about ancillarity and conditional inference might mean for predictive inference. This is followed by the bulk of the discussion which presents a numerical example for a curved exponential family. No exact ancillaries are known for this example but it will be shown that (i) the likelihood ancillary is particularly appropriate, (ii) the approximate p value suggested in (3.14) agrees with that in Barndorff-Nielsen (1990) expression (1.2), (iii) the “sensitive direction” points tangent to the manifold created by holding the likelihood ancillary fixed at the data. The findings of the example pose further questions.

## 2 Predictive Inference

The dual problem to parametric inference is predictive inference for unobserved  $z$ . Criterion II in §5 needs slight modification if  $z$  is to be inferred from observed  $y^0$  using a parametric model. Criterion III seem particularly relevant to this setting: ignore the reason why  $z$  has not been observed, whether it be in the future, the past, or perhaps because it is a random effect in a model and therefore can never be observed.

The problem is dual because, rather than conditioning on ancillary statistics to make the inference more relevant to the model and data at hand, the reference set might now be a fixed value for the sufficient statistic. In the simplest setting considered in Butler (1986), suppose there

is a sufficient statistic  $s(y)$  of fixed dimension agreeing with that of the model parameter  $\theta$ . If a generic value  $z$  is adjoined with  $y$  in  $(y, z)$ -space, then the evidence for this  $z$  value should be with respect to the reference set  $s(y^0, z)$ . In entertaining the value of  $z$  as the unobserved value, then  $s(y^0, z)$  conveys all relevant information about  $\theta$  and hence about the state of model used to make the predictive inference. In the same way that conditioning on ancillary  $a$  is used to convey “key observable characteristics of the underlying error” for a location model, conditioning on  $s(y^0, z)$  provides complete information about the current state of understanding for  $\theta$  within the parametric model setting. Fixing this understanding allows the model structure to make the prediction and also have it relevant to the current level of understanding about the model. Professor Fraser’s arguments in Appendix A were also particularly interesting and relevant as they relate to determining the ancillary values in  $(y, z)$ -space for prediction. The approach in Butler (1986) worked instead with orthonormal coordinates that are locally orthogonal to  $s(y^0, z)$ .

Barnard (1986) has also suggested a pivotal approach to prediction which is the dual procedure to the parametric inference in §4.2. Working with the location/scale model, his approach also uses the marginalization step to remove dependence on all parameters to determine the marginal distribution of an ancillary  $a(y, z)$ . This ancillary is now transformed into predictive pivot  $p(y, z)$  and predictive ancillary  $q(y)$ , with the latter quantity offering evidence for model criticism derived from data  $y$ . The conditional distribution of  $p(y, z)$  given  $q(y)$  evaluated at the data  $y = y^0$  now provides the predictive extrapolation.

Based on the discussion above, it seems likely that the inferential structure proposed by Professor Fraser can neatly accommodate the dual problem of prediction. Other predictive approaches attempting to extend higher order asymptotic methods beyond the restriction of sufficiency have included Butler (1989), Vidoni (1995), and Barndorff-Nielsen and Cox (1996). The first paper suggests that conditioning on the proper reference set, e.g. the maximum likelihood estimator  $\hat{\theta}(y^0, z)$ , provides a more generally applicable principle than the restriction due to predictive sufficiency. For an overview of this issue and others, see Bjørnstad (1996).

### 3 Gamma Exponential Example

An example is given that is similar to that considered by Pederson (1981). The example is used to consider (and partially answer) the following questions and speculations:

1. What sort of ancillary, affine or likelihood, should be used for inference about  $\theta$  and in what format?

2. Which ancillary is “more ancillary”?
3. What are the relationships between these ancillaries and the “sensitive” or ancillary directions suggested in the paper? Are there any deeper connections between the results of this paper and the suggestions of Barndorff-Nielsen (1990)?

### 3.1 Model and Ancillaries

A (2, 1) curved exponential family may be defined by supposing that  $y_1 \sim \text{Exponential}(\theta)$  independently of  $y_2 \sim \text{Exponential}(e^\theta)$ . To keep numerical computation simple, suppose the data are  $y_1^0 = 1$  and  $y_2^0 = 2$ . The MLE is

$$\hat{\theta}^0 = \text{LambertW}(1/2) \simeq 0.3517$$

and solves an equation which, when rearranged, allows  $y_2$  to be expressed in terms of  $\hat{\theta}$  and  $y_1$  as

$$y_2 = e^{-\hat{\theta}}(1/\hat{\theta} + 1 - y_1). \quad (1)$$

Two ancillaries are considered. The first is an affine ancillary  $a$  as discussed in Efron and Hinkley (1978) and Barndorff-Nielsen (1980) and sometimes named after the former authors. If vector  $y = (y_1, y_2)'$  has mean  $\mu_\theta$  and covariance  $\Sigma_\theta$  then the affine ancillary is computed as the MLE of the Studentized vector or

$$a^2 = (y - \mu_{\hat{\theta}})' \Sigma_{\hat{\theta}}^{-1} (y - \mu_{\hat{\theta}}) = (\hat{\theta}y_1 - 1)^2 + (e^{\hat{\theta}}y_2 - 1)^2 \quad (2)$$

and  $a^0 \simeq 1.954$ . In order to compute the  $p^*$  density for conditionality resolution  $\hat{\theta}|a$ , the transformation  $(y_1, y_2) \rightarrow (\hat{\theta}, a)$  needs to be inverted from (2) which leads to

$$y_1 = 1/\hat{\theta} - |a|/\sqrt{1 + \hat{\theta}^2} \quad (3)$$

with  $y_2$  given in (1).

The second ancillary is a likelihood ancillary. It is defined through the process of completing the (2, 1) curved exponential family so it is (2, 2) with the addition of another parameter  $\chi > 0$ . This is most simply done by assuming that  $y_2 \sim \text{Exponential}(\chi e^\theta)$  with the value  $\chi = 1$  creating the curved exponential family. The likelihood ancillary is now based on the likelihood ratio test that  $\chi = 1$ . If  $l_a(\theta, \chi)$  denotes the log-likelihood under the alternative, then the ancillary  $a_\chi$  assumes the value

$$\frac{1}{2}a_\chi^2 = l_a(\tilde{\theta}, \tilde{\chi}) - l_a(\hat{\theta}, 1) = -\ln \hat{\theta} + 1/\hat{\theta} - 1 - (1 - \hat{\theta})y_1 - \ln y_1 - \ln(1/\hat{\theta} + 1 - y_1), \quad (4)$$

where  $(\hat{\theta}, \tilde{\chi})$  denotes the MLE under the alternative. In (4), any dependence on  $y_2$  has already been replaced with  $y_1$  using (1). Let the sign of  $a_\chi^0$  be  $\text{sgn}(\tilde{\chi} - 1) = \text{sgn}(0.3517 - 1) = -1$  so that  $a_\chi^0 \simeq -1.546$ .

### 3.2 Which Ancillary and in What Format?

The format to be used for inference is the  $p^*$  density. It uses the likelihood shape to approximate the conditional density of  $\hat{\theta}|a; \theta$  as the normalized ( $d\hat{\theta}$ ) version of

$$p^\dagger(\hat{\theta}|a; \theta) = \sqrt{j_{\hat{\theta}}/(2\pi)} \exp\{l(\theta; \hat{\theta}, a) - l(\hat{\theta}; \hat{\theta}, a)\}.$$

In the case of the conditioning on the observed affine ancillary  $a^0$ , plots of  $p^*$  (dashed),  $p^\dagger$  (dotted) and the true density  $f(\hat{\theta}|a^0; \theta)$  (solid) are shown in Figures 1-4 below and are obtained through the inverse transformation  $\hat{\theta}|a^0 \rightarrow (y_1, y_2)$  given in (3). As  $\theta$  moves from  $\theta = 4$  (top left), 2, 1, to 1/2 (bottom right) the accuracy of  $p^*$  and  $p^\dagger$  diminish markedly.

Compare this with the use of  $p^*$  and  $p^\dagger$  but conditioning instead on the observed likelihood ancillary  $a_\chi^0$ . Figures 5-8 show the same quantities as their counterparts in Figures 1-4 as concerns the assessment of accuracy of  $p^*$  and  $p^\dagger$  for there respective true densities. However, the true conditional densities are different in the two sets of plots since Figures 1-4 fix  $a = a^0$  while Figures 5-8 fix  $a_\chi = a_\chi^0$ . Fixing  $a_\chi^0$  rather than affine  $a$  is a considerably more difficult computation since the inverse transformation  $\hat{\theta}|a_\chi^0 \rightarrow (y_1, y_2)$  requires selecting the correct  $y_1$  roots in (4) over a fine grid of  $\hat{\theta}$ -values. The true joint density of  $(\hat{\theta}, a_\chi)$  has also been computed the same way but with the additional complication of a Jacobian determination based on implicit differentiation.

### 3.3 Which Ancillary is “More Ancillary”?

The normalization constants ( $d\hat{\theta}$ ) of the joint densities  $f(\hat{\theta}, a_\chi^0; \theta)$  and  $f(\hat{\theta}, a^0; \theta)$  provide the marginal densities  $f(a_\chi; \theta)$  and  $f(a; \theta)$  which should not show extraordinary dependence on  $\theta$  if  $a$  and  $a_\chi$  are “good ancillaries”. Figures 9-10 plot  $f(a_\chi^0; \theta)$  (solid) and  $f(a^0; \theta)$  (dashed) versus  $\theta$ . These plots show the marginal evidence about  $\theta$  contained in each of the observed ancillaries. The observed likelihood ancillary is clearly “more ancillary” as revealed by the comparison in the right plot. All numerical computations for the likelihood ancillary here and in the previous subsection have used the grid  $\hat{\theta} \in \{.02(.04)9.98, 10\frac{1}{16}(\frac{1}{16})12, 12\frac{1}{8}(\frac{1}{8})16\}$ . The superior performance of the likelihood ancillary has been previously suggested in the asymptotics of Barndorff-Nielsen and Wood (1998). This superior performance can now be confirmed using a sample size of  $n = 1$  for this data set and model.

### 3.4 “Sensitive” Directions, P value Computations, and $r^*$ Connections

For this example, the ancillary direction is computed as

$$v' = -(y_1/\hat{\theta}, y_2)$$

which leads to the data dependent parameterization

$$\varphi(\theta) = \theta y_1/\hat{\theta} + e^\theta y_2.$$

Computation of the standardized maximum likelihood departure value leads to

$$q(\theta) = \text{sgn}(\hat{\theta} - \theta) \left| y_1(1 - \theta/\hat{\theta}) + y_2(e^{\hat{\theta}} - e^\theta) \right| \sqrt{j_{\hat{\theta}}} \left| y_1/\theta + e^{\hat{\theta}} y_2 \right|^{-1} \quad (5)$$

where

$$j_{\hat{\theta}} = 1/\hat{\theta}^2 + 1/\hat{\theta} + 1 - y_1.$$

At this junction, quite remarkably, it can be shown for any data  $(y_1^0, y_2^0)$ , that  $q(\theta)$  is analytically the same as the value for the standardized maximum likelihood departure  $u$  suggested in Barndorff-Nielsen (1990) as (1.4) and computed as in (5.5). We return to the implications of this equivalence below but first pause to tabulate some p values in Table 1.

Method	$\theta = 1/2$	$3/4$	1	$3/2$	2
Exact (trapezoidal)	.189	.0689	.0194	.03489	.05120
(3.14) with $q$	.238	.0864	.0239	.03583	.05140
Skovgaard	.259	.0990	.0289	.03796	.05219
Normal	.325	.130	.0392	.02112	.05315

*Table 1.* P values  $p^0(\theta)$  for the various methods listed in the rows. “Exact” refers to trapezoidal summation for  $\Pr(\hat{\theta} < \hat{\theta}^0 | a_\chi^0; \theta)$ , (3.14) accounts for the sensitive direction as well as Barndorff-Nielsen’s (1990) value of  $u$ , “Skovgaard” (1996) computes  $u$  using the author’s approximate sample space derivatives, and Normal uses the normal approximation to  $r$  in (3.11).

Even for this  $n = 1$  setting, the sensitive direction approach and that using Skovgaard’s (1996) approximate sample space derivatives show remarkable accuracy particularly for large  $\theta$ . Taking the inference for  $\theta$  further, the exact confidence interval by inverting  $\Pr(\hat{\theta} < \hat{\theta}^0 | a_\chi^0; \theta)$  gives (.0276, .664) while (3.14) gives (.0446, .717) and Skovgaard’s method gives (.0478, .748).

The analytical equivalence of Fraser’s  $q(\theta)$  with  $u$  from Barndorff-Nielsen’s (1990) approach which explicitly conditions on  $a_\chi^0$ , suggests that the sensitive direction in which the directional derivative is taken in (3.13) to define  $\varphi(\theta)$ , is tangent to the manifold  $\{(y_1, y_2) : a_\chi(y_1, y_2) = a_\chi^0\}$ .

This is indeed the case. Implicit differentiation of (1) to determine  $\partial y_2/\partial y_1$  holding  $a_\chi^0$  fixed requires the determination of  $\partial \hat{\theta}/\partial y_1$  through (4). After long computations,

$$\partial y_2/\partial y_1 = \hat{\theta} y_2/y_1 = v_2/v_1, \quad (6)$$

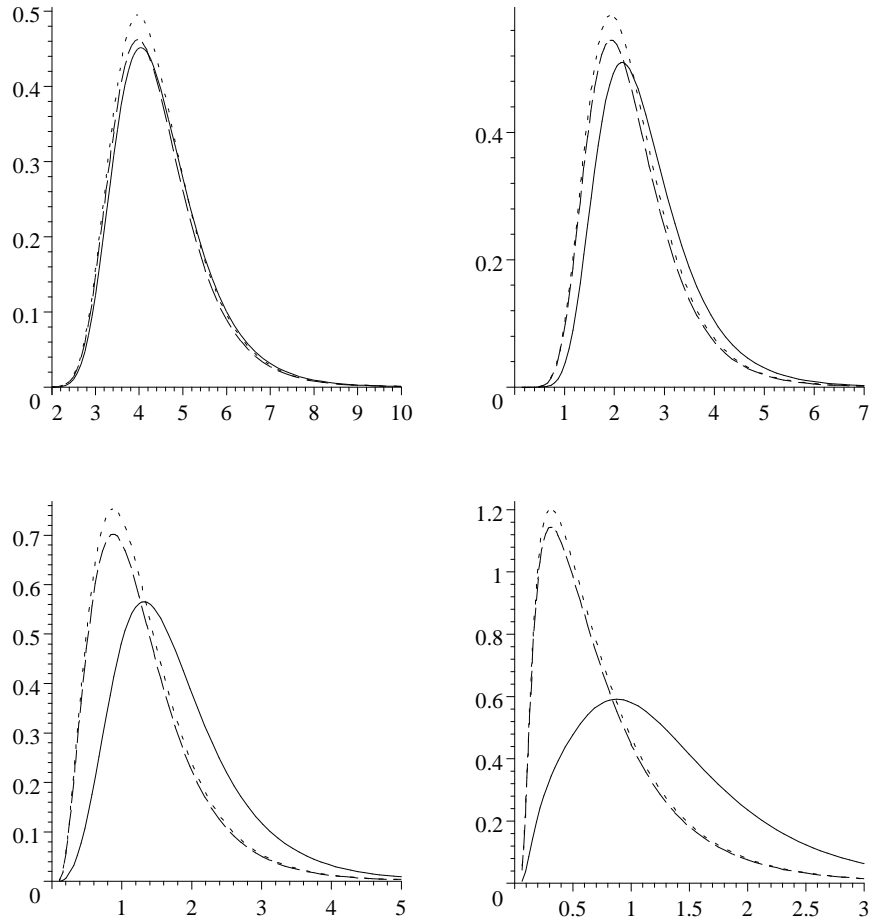
the direction of  $v$ . At the data this slope is 0.7035.

Is this example merely a coincidence or are there any greater generalities to these agreements? To be tangent to the likelihood ancillary curve, the curve must be a solution to the differential equation in (6) which is complicated by the dependence of  $\hat{\theta}$  on  $(y_1, y_2)$ . General differential equation theory (see Ross, 1974, theorem 1.1) only guarantees a local solution to (6) at the data but this is all that is required for a local ancillary. This seems to say that the sensitive direction approach has greater mathematical generality when a likelihood ancillary does not exist. If it does exist, when is the sensitive direction equal to or “close” to the direction of the likelihood ancillary curve? To what extent can the equivalence between Fraser’s  $q(\theta)$  and Barndorff-Nielsen’s  $u$  be asserted with or without nuisance parameters in curved exponential families or in other classes of models?

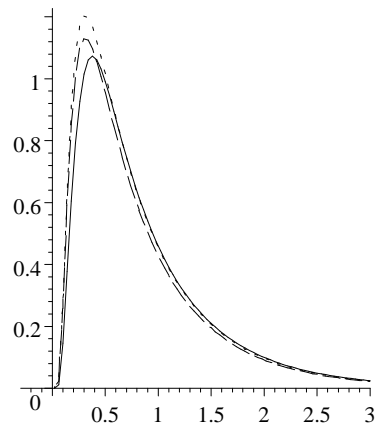
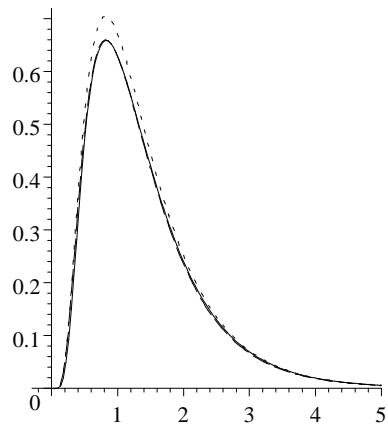
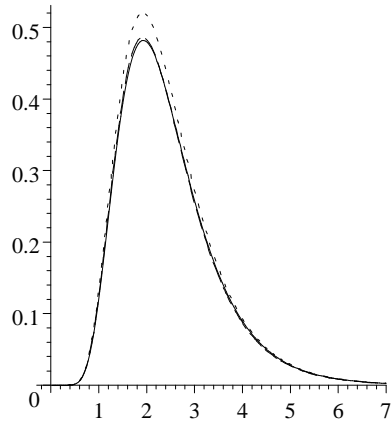
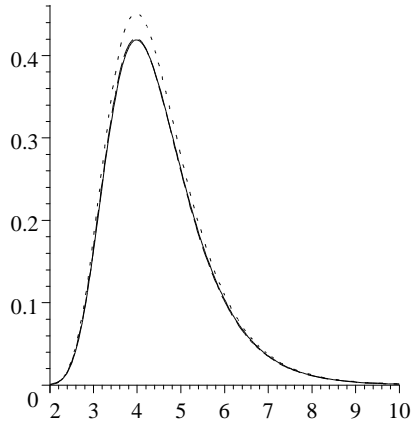
## References

- [1] Barnard, G. (1986). Discussion of paper by Butler (1986).
- [2] Barndorff-Nielsen, O. (1990). Approximate interval probabilities. *J. Roy. Statist. Soc. B* **52** 485-496.
- [3] Barndorff-Nielsen, O. (1991). Modified signed log likelihood ratio. *Biometrika* **78** 557-563.
- [4] Barndorff-Nielsen, O. and Cox, D.R. (1996) Prediction and asymptotics. *Bernoulli* **2** 319-340.
- [5] Barndorff-Nielsen, O. and Wood, A.T.A. (1996) On large deviations and choice of ancillary for  $p^*$  and  $r^*$ . *Bernoulli* **4** 35-63.
- [6] Bjørnstad, J.F. (1996) On the generalization of the likelihood function and the likelihood principle. *J. Amer. Statist. Assoc.* **91** 791-806.
- [7] Butler, R.W. (1986). Predictive likelihood inference with applications (with Discussion). *J. Roy. Statist. Soc. B* **48** 1-38.
- [8] Butler, R.W. (1989). Approximate predictive pivots and densities. *Biometrika* **76** 489-501.
- [9] Efron, B. and Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika* **65** 457-482.
- [10] Pederson, B. V. (1981). Comparison of the Efron-Hinkley ancillary and the likelihood ratio ancillary in a particular example. *Ann. Statist.*, **9** 1328-1333.
- [11] Ross, S.L. (1974). *Differential Equations*. 2nd Ed. Xerox College Publishing. Lexington, MA.
- [12] Skovgaard, I.M. (1996). An explicit large-deviation approximation to one-parameter tests. *Bernoulli* **2** 145-165.

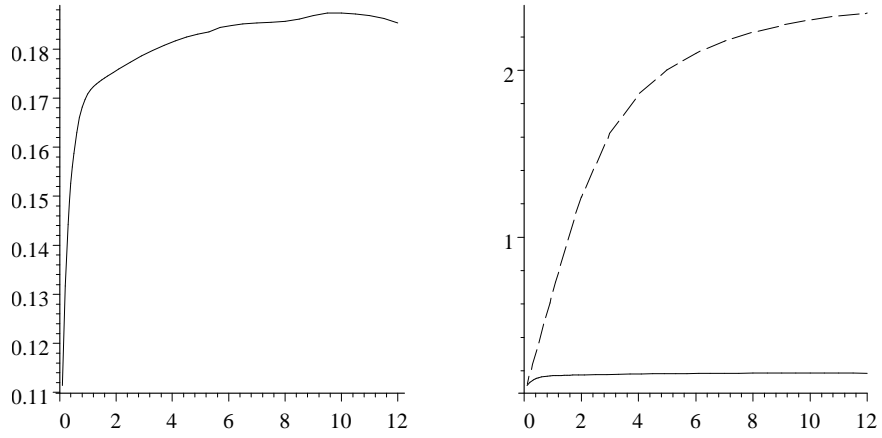
- [13] Vidoni, P. (1995). A simple predictive density based on the  $p^*$  formula. *Biometrika* **82** 855-863.



Figures 1-4. Densities for  $\hat{\theta}$  in the Gamma Exponential example when conditioning on the affine ancillary  $a^0 = 1.954$ . The plots, showing a range of accuracy from good to poor, depicting the exact density  $f(\hat{\theta}|a^0; \theta)$  (solid),  $p^\dagger(\hat{\theta}|a; \theta)$  (dotted), and  $p^*(\hat{\theta}|a; \theta)$  (dashed) for  $\theta = 4, 2, 1$ , and  $1/2$  respectively.



Figures 5-8. Densities for  $\hat{\theta}$  when conditioning on the likelihood ancillary  $a_{\chi}^0 = -3.140$ . In each plot,  $f(\hat{\theta}|a_{\chi}^0; \theta)$  (solid),  $p^{\dagger}(\hat{\theta}|a_{\chi}^0; \theta)$  (dotted) and  $p^*(\hat{\theta}|a_{\chi}^0; \theta)$  (dashed) are shown.



Figures 9-10. Marginal likelihood plots for  $f(a_\chi^0; \theta)$  (solid) and  $f(a; \theta)$  (dashed) versus  $\theta$  where  $a_\chi^0$  and  $a$  are the likelihood and affine ancillaries respectively.