Statistics in the Twenty-First Century: Special Volume
In Honour of Distinguished Professor Dr. Mir Masoom Ali
On the Occasion of his 75th Birthday Anniversary
PJSOR, Vol. 8, No. 3, pages 433-440, July 2012

Vector Exponential Models and Second Order Inference

D.A.S. Fraser, Uyen Hoang, Kexin Ji, Xufei Li, Li Li, Wei Lin and Jie Su
Department of Statistics, University of Toronto
Toronto, Ontario, Canada M5S 3G3
dfraser@utstat.toronto.edu

Summary

For an exponential model with scalar parameter, WelchP:1963 examined the role of Bayesian analysis in statistical inference, more specifically the use of the Jeffreys:1946 prior. They determined that Bayesian intervals and thus in effect Bayesian quantiles had second order confidence accuracy. We use a Taylor series expansion of the log-model to develop a second order version of the vector exponential model; this is developed as a contribution to theory in statistics at a time when algorithms are prominent, and it provides a basis for generalizing the Welch-Peers approach to the vector parameter context.

Some Keywords: Asymptotic model; Bayes as approximate confidence; Exponential model; Jeffreys prior; Likelihood analysis; Root-information prior; Second order expansion.

1. Introduction

Exponential models are widely used in contemporary statistics, offering rich model flexibility with relatively easy analysis that is largely immune to high data dimension: In basic theory they support the uniformly most powerful unbiased and similar tests (Lehmann & Romano,2005, Section 4.4); In model building they provide the structure to go beyond Normal theory to generlized linear models (Nelder & Wedderburn,1972; McCullagh & Nelder,1989); With graphical models they provide key structure (Lauritzen,1996); In machine learning they are a primary ingredient (Wainwright & Jordan,2008); In approximation theory they offer a basis for saddlepoint methodology (Daniels,1954); In current inference theory they underpin higher order methods (Barndorff-Nielsen,1986); And for determining the connections between Bayesian and frequentist inference they give access to the needed location models (Welch & Peers,1963).

A scalar exponential model $f(y;\theta) = \exp\{s(y)\varphi(\theta) + k(\theta)\}h(y)$ can be viewed as an exponential tilt of a basic density or relative density $h(y)$. By suitably recentering the variable and the parameter relative to observed data y^0 and corresponding $\hat{\theta}^0$ we can work with the more transparent form of model $g(s;\varphi) = \exp\{s\varphi + \ell(\varphi)\}g(s)$ where $\ell(\varphi)$ is the reexpressed observed log-likelihood with observed data value $s = 0$. Asymptotic

theory examines the effects of increasing data size n and produces remarkably accurate approximations for the density and the distribution function at the observed data point.

These approximations use the log-likelihood function $\ell(\varphi;s) = \log g(s;\varphi)$ which is typically directly available from the original model as $\log f(y;\theta)$ provided allowance is made for parameter change from θ to φ. The approximations also typically use two intermediate measures of departure, the signed likelihood root r and the maximum likelihood departure q in the canonical φ scaling:

$$r = \mathrm{sign}(\hat{\varphi}-\varphi)\{2[\ell(\hat{\varphi})-\ell(\varphi)]\}^{1/2}, \quad q = \hat{j}_{\varphi\varphi}^{1/2}(\hat{\varphi}-\varphi), \tag{1}$$

where $\hat{\varphi} = \hat{\varphi}(s)$ is the value that maximizes the likelihood $\ell(\varphi;s)$ and $\hat{j}_{\varphi\varphi} = \hat{j}_{\varphi\varphi}(s) = (\partial^2/\partial\varphi^2)\ell(\varphi;s)|_{\hat{\varphi}}$ is the curvature of likelihood at that maximum. For a model with data, the value of r and q are part of the usual output from many statistical packages, and come directly from the observed likelihood function: r comes from vertical change in likelihood and q comes from horizontal change, maximum likelihood value less parameter value of interest.

The highly accurate approximations for the density and distribution functions are

$$f(s;\varphi) = e^{k/n}\phi(r)\hat{j}_{\varphi\varphi}^{-1/2}, \quad F(s;\varphi) = \Phi\{r - r^{-1}\log(r/q)\} \tag{2}$$

where $\phi(z)$ and $\Phi(z)$ are the standard Normal density and distribution functions and accuracy is third order $O(n^{-3/2})$ with k constant to that order. These approximations from Daniels (1954) and Barndorff-Nielsen(1991) provide exceptional access to statistical inference, both theoretical and practical; for some recent discussion see Fraser (2011).

In particular, the approximations lead to a simple proof of the Welch & Peers (1963) theorem, and provide the stronger statement that a scalar exponential model is a location model to second order, which thus justifies the Jeffreys (1946) choice of root information as a second order prior for the scalar exponential model; this is described from a Taylor expansion view in § 2 and § 3. We then pursue the Welch-Peers direction and derive the second order form for a vector exponential model in § 4. Some discussion in § 5 focusses on the implications for general statistical inference but does not pursue the details; some immediate consequences will developed in Fraser et al. (2012). Thus our present material can be viewed as a contribution to statistical theory and to the interpretation of statistical theory, particularly the profound Welch-Peers result.

2. Welch-Peers and log-model expansions

Statistics has two rather different methodologies for statistical inference: the frequentist which is based on frequency properties in the statistical model, and the Bayesian which augments this with a prior density purporting to describe origins for the particular parameter value. The literature records support for one or the other of these, or discusses

conflicts between or within the approaches, or proposes alternative approaches such as (Fisher,1956). A profound link among these emerged with Welch & Peers (1963) who showed that the two approaches lead to the same result to second order in the presence of a scalar exponential model, provided the prior used is the Jeffreys root information prior $\pi(\varphi)d\varphi = j_{\varphi\varphi}^{1/2}(\varphi)d\varphi$; for some recent discussion see Fraser (2011).

A more transparent route to this Welch & Peers (1963) result is available using a Taylor expansion of the log-model in terms of appropriately standardized variable and parameter; for some background see Fraser & Reid (1993) and Cakmak et al. (1998). These expansions are usually examined to third order, but for the Welch-Peers result to be discussed here a second order expansion suffices, and is flexible for extension to the vector parameter context.

For the present scalar exponential model $f(y;\theta) = \exp\{s(y)\varphi(\theta) + k(\theta)\}h(y)$ we first consider the log-model using a centered and scaled version of φ, a centered and scaled version of s, plus expansion to the second order; this leads to the second order expansion of log-likelihood as

$$\ell(\varphi;s) = s\varphi + \ell(\varphi) = -\varphi^2/2 - \gamma\varphi^3/6n^{1/2} + s\varphi$$

where $\gamma/n^{1/2} = j_{\varphi\varphi\varphi} = -\ell_{\varphi\varphi\varphi}(\hat{\varphi})$ is the negative third derivative of log-likelihood at the maximum. The standardized statistical model then has the second order form:

$$\begin{aligned} g(s;\varphi) &= (2\pi)^{-1/2}\exp(-\varphi^2/2 + \varphi s)\exp(-\gamma\varphi^3/6n^{1/2})h(s) && (3)\\ &= \phi(s-\varphi)\exp\{-\gamma\varphi^3/6n^{1/2} + \gamma(s^3-3s)/6n^{1/2}\} && (4)\\ &= \phi(s-\varphi)\exp\{-\gamma\varphi^3/6n^{1/2} + \gamma s^3/6n^{1/2}\}(1-\gamma s/2n^{1/2}) && (5)\end{aligned}$$

where $h(s)$ has been determined to the second order by expanding the second exponential in (3), then using $E(s^3-3s;\varphi) = \varphi^3$ for the pure Normal $(\varphi;1)$, and then returning one term to the exponent.

The likelihood $\ell(\varphi;s) = -\varphi^2/2 - \gamma\varphi^3/6n^{1/2} + \varphi s$ has score function $\ell_\varphi(\varphi;s) = s - \varphi - \gamma\varphi^2/2n^{1/2}$ which gives the maximum likelihood value $\hat{\varphi}(s) = s - \gamma s^2/2n^{1/2}$. The information function $j_{\varphi\varphi}(\varphi;s) = j_{\varphi\varphi}(\varphi) = -\ell_{\varphi\varphi}(\varphi)$ is the negative Hessian of likelihood or the negative derivative of score ℓ_φ; thus $j_{\varphi\varphi} = 1 + \gamma\varphi/n^{1/2}$. This then gives the Welch-Peers prior

$$\pi(\varphi) = j_{\varphi\varphi}^{1/2}(\varphi) = (1+\gamma\varphi/n^{1/2})^{1/2} = 1 + \gamma\varphi/2n^{1/2} \qquad (6)$$

to second order: see Reid & Fraser(2010) and Fraser et al.(2011).

The essence of Welch & Peers (1963) analysis comes from a location relationship between a reexpressed parameter say β, a reexpressed variable say $\hat{\beta}$, and an associated model $f(\hat{\beta}-\beta)d\hat{\beta}$; we demonstrate this essence of Welch & Peers (1963) in § 3. But first we record a fundamental consequence that comes from a location model. With such

a model say $f(y-\theta)$ with scalar variable, scalar parameter and data y^0 we have the p-value

$$p(\theta; y^0) = \int_{-\infty}^{y^0} f(y-\theta)dy$$

which records the statistical position of the data y^0 in the population labelled θ; it can be viewed as the primitive or mother-of-all p-values. And with the addition of a flat prior for θ as motivated by location invariance we have the posterior distribution $f(y^0-\theta)$ for θ and then the Bayesian survivor value

$$s(\theta; y^0) = \int_{\theta}^{\infty} f(y^0-\theta)d\theta;$$

these present the conditional probability consequences of introducing or formally assuming probability properties for the mathematical prior $\pi(\theta)$ when no such properties are part of the given. Introductory calculus then shows that the two integrals are equal. This gives a p-value or confidence calibration for the Bayes approach, and provides a primary or fundamental connection between Bayesian and frequentist methodologies. For some recent discussion see Fraser et al.(2010) and Fraser & Reid (2011), and for some background on the Bayes-frequentist connection see Fraser (2011).

And then for the scalar exponential model considered here using the location results from the next section we have the WelchP:1963 result that the frequentist p-value and the Bayes survivor value $s(\beta;\hat{\beta})$ are equal to the second order, where

$$p(\beta;\hat{\beta}) = \int f(\hat{\beta}-\beta)d\hat{\beta}, \quad s(\beta;\hat{\beta}) = \int_{\beta} L(\beta;\hat{\beta})cd\beta = \int_{\beta} f(\hat{\beta}-\beta)d\beta \qquad (7)$$

with the constant prior $c=1$ for β: the two integrals are numerically equal so that $p(\beta;\hat{\beta}) = s(\beta;\hat{\beta})$. In § 3 we determine the second order location parameter β and the related model $f(\hat{\beta}-\beta)d\hat{\beta}$.

3. Location parameterization and the reexpressed model

We use the information function to rescale on the parameter space and thereby derive a new parameterization. For the standardized model (2) the information $j_{\varphi\varphi}(\varphi;s) = -\ell_{\varphi\varphi}(\varphi)$ arises as a second derivative; the root information can then be viewed as a rate for the given parameter and accordingly we can rescale locally to obtain an increment for a new parameterization and then integrate to obtain the new β in terms of the old φ:

$$d\beta = j_{\varphi\varphi}^{1/2}(\varphi)d\varphi = (1+\gamma\varphi/2n^{1/2})d\varphi, \qquad (8)$$

$$\beta = \int j_{\varphi\varphi}^{1/2}(\varphi)d\varphi = \int(1+\gamma\varphi/2n^{1/2})d\varphi = \varphi + \gamma\varphi^2/4n^{1/2} \qquad (9)$$

where the lower limit of integration for convenience is at $\hat{\varphi}^0$ which is 0 in the standardized notation The reverse transformation expressing the old φ in terms of the

new β is then $\varphi = \beta - \gamma\beta^2 / 4n^{1/2}$. And somewhat similarly we obtain the connections between the old and the new variables; using (9) we obtain

$$\hat{\beta} = \hat{\varphi} + \gamma\hat{\varphi}^2 / 4n^{1/2} = s - \gamma s^2 / 2n^{1/2} + \gamma(s - \gamma s^2 / 2n^{1/2})^2 / 4n^{1/2} = s - \gamma s^2 / 4n^{1/2}, \quad (10)$$

$$d\hat{\beta} = (1 - \gamma s / 2n^{1/2})ds; \quad (11)$$

and for the reverse transformation we have $s = \hat{\beta} + \gamma\hat{\beta}^2 / 4n^{1/2}$.

We now reexpress the standardized model in terms of the new parameter β and the related maximum likelihood variable $\hat{\beta}$. For this we make the change of variable $s = \hat{\beta} + \gamma\hat{\beta}^2 / 4n^{1/2}$ and parameter $\varphi = \beta - \gamma\beta^2 / 4n^{1/2}$ in (4), and obtain

$$g(s;\varphi)ds = \phi(s - \varphi)\exp\{-\gamma\varphi^3 / 6n^{1/2} + \gamma s^3 / 6n^{1/2}\}(1 - \gamma s / 2n^{1/2})ds \quad (12)$$

$$= (2\pi)^{-1/2}\exp\{-s^2 / 2 + s\varphi - \varphi^2 / 2 - \gamma\varphi^3 / 6n^{1/2} + \gamma s^3 / 6n^{1/2}\}d\hat{\beta} \quad (13)$$

$$= (2\pi)^{-1/2}\exp\{-(\hat{\beta} - \beta)^2 / 2 - \gamma(\hat{\beta} - \beta)^3 / 12n^{1/2}\}d\hat{\beta} \quad (14)$$

$$= (2\pi)^{-1/2}\exp\{-z^2 / 2 - \gamma z^3 / 12n^{1/2}\}dz, \quad (15)$$

where the first equality brings one term down from the exponent to form the factor before ds, the second equality come from moving the Normal density to the exponent, the third collects the quadratic and cubic terms to order $O(n^{-1})$, and the fourth expresses the model in terms of the centered pivot $z = \hat{\beta} - \beta$. We thus see that the model for $\hat{\beta}$ is location with constant observed information $\hat{j}_{\beta\beta} = 1$.

A simple example with immediate verification. Consider y with an exponential life model $f(y;\varphi) = \varphi\exp(-y\varphi)$ and positive y and φ. The canonical variable is $s = -y$ and the likelihood becomes $\ell(\varphi;s) = s\varphi + \log\varphi$ giving $\ell_\varphi = s + \varphi^{-1}$ and $j_{\varphi\varphi} = \varphi^{-2}$. Then from (9) we obtain

$$\overset{\varphi}{\beta} = \int\varphi^{-1}d\varphi = \log\varphi,$$

and then from $\hat{\varphi} = y^{-1}$ obtain $\hat{\beta} = -\log y$. Accordingly consider the distribution of $\hat{\beta} - \beta = -\log y - \log\varphi = -\log y\varphi = -\log z$ where $z = y\varphi$. Of course z has the standard exponential distribution $g(z) = \exp(-z)$ on $(0,\infty)$ and thus $w = \hat{\beta} - \beta = -\log z$ has the extreme value distribution $\exp(-e^{-w} - w)$ on $(-\infty,+\infty)$, free of the parameter; thus $\hat{\beta}$ has a location distribution exactly, which is centered at β.

4. Vector exponential model and the second order reexpression

Consider a full p-dimensional continuous exponential model with canonical parameter $\varphi(\theta)$ and canonical variable $s(y)$, as reexpressed in the standardized form used in the preceding section:

$$f(s;\varphi) = \exp\{\varphi' s + \ell(\varphi)\}h(s), \tag{16}$$

where $\ell(\varphi)$ is the observed likelihood at a data point y^0 of interest with $s^0 = s(y^0) = 0$ by recentering the canonical variable and $\hat{\varphi}^0 = 0$ by recentering the canonical parameter. With moderate regularity we then further standardize to obtain observed information $\hat{j}^0_{\varphi\varphi} = I$, the identity matrix. For this it is often convenient to order the coordinates of φ such as $\varphi = (\psi, \lambda)^T$ in the $p = 2$ case so as to maintain the integrity of an interest parameter say ψ; we can then take $j^{1/2}$ to be the right positive lower triangular square root of the observed information matrix, $\hat{j}^0_{\varphi\varphi} = (j^{1/2})'(j^{1/2})$, and define a new parameterization as $j^{1/2}\varphi$ in terms of the old parameterization. The new parameterization is then centered at the observed maximum but with an identity observed information and has new first parameter coordinate monotone increasing in the old first parameter. With asymptotic properties the model can then be expressed in the pattern (3),

$$g(s;\varphi) \quad = \quad \phi(s-\varphi)\exp\{-\Sigma_{ijk}\gamma_{ijk}\varphi_i\varphi_j\varphi_k / 6n^{1/2}\}h(s), \tag{17}$$

to the second order, where $\gamma_{ijk} / n^{1/2} = j_{\varphi_i\varphi_j\varphi_k}$ is the third derivative of negative log-likelihood with respect $\varphi_i, \varphi_j, \varphi_k$ but calculated in the standardized coordinates.

We now derive $h(s)$ to second order following the pattern from (3) to (4) and (5) in § 2. There a term say $-\gamma\varphi_1^3 / 6n^{1/2}$ in the exponent is brought down as $1 - \gamma\varphi_1^3 / 6n^{1/2}$ which then requires $h(s) = 1 + \gamma(s_1^3 - 3s_i) / 6n^{1/2}$ based on $E(s^3 - 3s) = \varphi^3$ for the Normal$(\varphi;1)$. Now a term $-\gamma\varphi_1^2\varphi_2 / 2n^{1/2}$ comes down as $1 - \gamma\varphi_1^2\varphi_2 / 2n^{1/2}$ which then requires $h(s) = 1 + \gamma(s_1^2 - 1)s_2 / 2n^{1/2}$ based on $E(s_1^2 - 1)s_2 = \varphi_1^2\varphi_2$ for the Normal$(\varphi_1, \varphi_2; I)$. Also a term $-\gamma\varphi_1\varphi_2\varphi_3 / n^{1/2}$ comes down as $1 - \gamma\varphi_1\varphi_2\varphi_3 / n^{1/2}$ which then requires $h(s) = 1 + \gamma s_1 s_2 s_3 / n^{1/2}$ based on $E(s_1 s_2 s_3) = \varphi_1\varphi_2\varphi_3$ for the Normal$(\varphi_1, \varphi_2, \varphi_3; I)$. We can then combine these steps and obtain

$$\begin{aligned} g(s; \quad \varphi) = \phi(s-\varphi)\exp\{-\Sigma_i\gamma_{iii}(\varphi_i^3 - s_i^3)6n^{1/2} - \Sigma_{i\neq j}\gamma_{iij}(\varphi_i^2\varphi_j - s_i^2 s_j) / 2n^{1/2}\}\times \\ \exp\{-\Sigma_{i<j<k}\gamma_{ijk}(\varphi_i\varphi_j\varphi_k - s_i s_j s_k) / n^{1/2}\}(1 - \Sigma_i\gamma_{iii}s_i / 2n^{1/2} - \Sigma_{i\neq j}\gamma_{iij}s_j / 2n^{1/2}). \end{aligned}$$

Example: Second order two-dimension exponential model. The two dimensional model in standardized notation has the form

$$\begin{aligned} g(s_1, s_2; \quad \varphi_1, \varphi_2) = \phi(s_1 - \varphi_1)\phi(s_2 - \varphi_2)\exp\{-\gamma_3(\varphi_1^3 - s_1^3) / 6n^{1/2}\}\times \\ \exp\{-\gamma_2(\varphi_1^2\varphi_2 - s_1^2 s_2) / 2n^{1/2} - \gamma_1(\varphi_1\varphi_2^2 - s_1 s_2^2) / 2n^{1/2} - \gamma_0(\varphi_2^3 - s_2^3) / 6n^{1/2}\}\times \\ (1 - \gamma_3 s_1 / 2n^{1/2} - \gamma_2 s_2 / 2n^{1/2} - \gamma_1 s_1 / 2n^{1/2} - \gamma_0 s_2 / 2n^{1/2}). \end{aligned}$$

5. Discussion

Exponential models can be used to construct larger models in applications, to deconstruct larger models to determine p-values for interest parameters, and also to get highly accurate approximations for those p-values. With this as background we have discussed the Welch & Peers (1963) demonstration that Bayes analysis can reproduce standard results to the second order, but just in the scalar-parameter exponential model context. We have then developed the second order version of the vector parameter exponential model, for general purposes and for seeking vector-parameter extensions of the Welch-Peers. We view the second order exponential model as a contribution to statistical theory; some aspects of the vector Welch-Peers have been mentioned in Fraser et al. (2005) and further aspects will be reported separately (Fraser et al.,2012).

Acknowledgement

The authors acknowledge support from the Natural Sciences and Engineering Research Council of Canada.

References

1. Barndorff-Nielsen, O. E. (1986). Inference on full or partial parameters based on the standardized, signed log likelihood ratio. *Biometrika*, 73, 307–322.
2. Barndorff-Nielsen, O. E. (1991). Modified signed log likelihood ratio. *Biometrika*,78, 557–563.
3. Cakmak, S., Fraser, D. A. S., McDunnough, P., Reid, N. & Yuan, X. (1998). Likelihood centered asymptotic model: exponential and location model versions. *J. Statist. Planning and Inference*, 66, 211–222.
4. Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Annals Math. Statist*. 46, 21–31.
5. Fisher, R. (1956). *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.
6. Fraser, A. M., Fraser, D. A. S. & Fraser, M. J. (2012). An adjusted Jeffreys for higher-order inference. *J.Royal Statist. Soc., B* .
7. Fraser, D., Rekkas, M. & Wong, A. (2005). Highly accurate likelihood analysis for the seemingly unrelated regression problem. *Journal of Econometrics*, 127, 17–33.
8. Fraser, D. A. S. (2011). Is Bayes posterior just quick and dirty confidence? with discussion. *Statistical Science*, 26, 299–316.
9. Fraser, D. A. S., Naderi, A., Ji, K. & Su, J. (2011). Exponential models: Approximations for probabilities.*Jour. Iranian Statist. Soc.* 10, to appear.
10. Fraser, D. A. S. & Reid, N. (1993). Third order asymptotic models: Likelihood functions leading to accurate approximations for distribution functions. *Statist. Sinica*,3, 67–82.

11. Fraser, D. A. S. & Reid, N. (2011). On default priors and approximate location models. *Brazilian Jour. Prob.Statist*. to appear, 1–16.
12. Fraser, D. A. S., Reid, N., Marras, E. & Yi, G. (2010). Default priors for Bayesian and frequentist inference. *J. Royal Statist. Soc.* 75, 631–654.
13. Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. A* 186,453–461.
14. Lauritzen, S. (1996). *Graphical Models*. Oxford: Clarendon.
15. Lehmann, E. L. & Romano, J. P. (2005). *Testing Statistical Hypotheses*. New York: Springer.
16. McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models*. Boca Raton: Chapman and Hall.
17. Nelder, J. & Wedderburn, R. (1972). Generalized linear models. *Jour. Royal Statist. Soc. A*,135, 370–384.
18. Reid, N. & Fraser, D. A. S. (2010). Mean likelihood and higher order approximations. *Biometrika,* 97, 159–170.
19. Wainwright, M. & Jordan, M. (2008). *Graphical Models, Exponential Families, and Variational Methods.* Boston: Now Publishers.
20. Welch, B. & Peers, H. (1963). On formulae for confidence points based on intervals of weighted likelihoods. *Jour. Royal Statist. Soc. B*,25, 318–329.