Parameterizations for Natural Exponential Families With Quadratic Variance Functions

Elizabeth H. SLATE*

Parameterizations for natural exponential families (NEF's) with quadratic variance functions (QVF's) are compared according to the nearness to normality of the likelihood and posterior distribution. Nonnormality of the likelihood (posterior) is measured using two criteria. The first is the magnitude of a standardized third derivative of the log-likelihood (logposterior density); the second is a comparison of the probability of particular tail regions under the normalized likelihood (posterior distribution) and under the corresponding normal approximation. A relationship is given that links these two criteria. Sample sizes are recommended for adequate normality in the likelihood for various parameterizations of the NEF-QVF models, and these results are extended to Bayesian models with a conjugate prior.

KEY WORDS: Conjugate prior; Natural exponential family; Nonnormality; Skewness reduction transformation; Third derivative; Variance Stabilizing parameterization.

1. INTRODUCTION

It is well known that the parameterization of a statistical model affects the accuracy of numerical calculations and the validity of inferences based on asymptotic approximations. Computation of maximum likelihood estimates (MLE's) or posterior modes, for example, requires maximization of the likelihood or posterior, which is typically performed using a Newton-like method. Such optimization methods will be more efficient if the likelihood or posterior distribution is approximately normal. In addition, the calculation of posterior densities, means, and probabilities requires some form of numerical integration. Popular methods include Gauss–Hermite quadrature (Naylor and Smith 1982) and Laplace's approximation (Tierney and Kadane 1986), which are also based on an assumption that the posterior is approximately normal.

Normality is highly desirable for inference as well. Exact posterior probability regions may be quite expensive to compute (see Ross 1978, for instance, for likelihood regions for nonlinear models) and complicated (and hence difficult to interpret), whereas regions based on limiting normality are easily obtained. If large amounts of data may be collected, then an adequate normal approximation may be assured for nearly any parameterization. In most cases, however, only a relatively small amount of data is available, and it may be possible to improve the validity of inferences based on asymptotic normality by transforming to a more well-behaved parameterization.

This article investigates the behavior of some key parameterizations for a model that supposes that the observations are a random sample from a natural exponential family distribution with quadratic variance function (NEF-QVF), as studied by Morris (1982, 1983). The parameterizations are evaluated according to the nearness to normality of the likelihood function and also according to the accuracy of the modal normal approximation to the posterior distribution under a conjugate prior.

Two nonnormality diagnostics are used as criteria for selecting a good parameterization. The first is a standardized third derivative of the log-likelihood function, considered by Sprott (1973, 1975, 1980). From the Bayesian viewpoint, the corresponding measure is based on the third derivative of the logposterior, which has also been considered by Albert (1989), Hills (1989), and, through a more general multivariate measure, by Kass (1989) and Kass and Slate (1992, 1993). The second parameterization criterion is a comparison of the probability of particular tail regions under the posterior distribution (or normalized likelihood) and its normal approximation. It is shown in Section 2 that the tail probability criterion can be expressed in terms of the standardized third derivative, and this relationship is used to translate a definition of acceptable normality in terms of the tail probabilities into an upper bound on the magnitude of the third-derivative measure. The parameterizations are ranked by the size of the sample required for adequate normality of the likelihood and posterior.

Although one can often qualitatively evaluate the accuracy of a normal approximation for one-parameter models by directly graphing the likelihood or posterior, it is essential to quantify the nonnormality and to calibrate these measures to improve understanding of parameterization behavior. The knowledge of the parameterizations and nonnormality diagnostics gained through this study can be used to improve the understanding of more complex multiparameter models. In addition, the class of NEF-QVF distributions is of interest because it contains many of the distributions more commonly used in statistical modeling.

The six NEF-QVF distributions are the normal, gamma, Poisson, negative binomial, binomial, and generalized hyperbolic secant distributions. For each of these distributions, a number of parameterizations are considered, with particular emphasis on the mean-value, variance-stabilizing, and vanishing third-derivative parameterizations. The NEF-QVF models and parameterizations are described in Section 3. The results discussed in Section 4 indicate that the vanishing

* Elizabeth H. Slate is Assistant Professor, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853. This research was partially supported by the National Science Foundation Grants DMS-9005858 and DMS-9309990. Much of the research comes from a portion of the author's Ph.D. dissertation research conducted under the supervision of Robert E. Kass at Carnegie Mellon University. The author thanks Robert E. Kass for many helpful discussions during the course of this research and comments that improved this presentation. The comments of two referees also greatly improved the presentation.

Journal of the American Statistical Association
December 1994, Vol. 89, No. 428, Theory and Methods

third-derivative parameterization performs optimally, or nearly so, within the class of parameterizations considered for each family. The extension of the likelihood results to the Bayesian model with a conjugate prior is given in Section 5, and further discussion is given in Section 6.

Notation. Let the log-likelihood for the one-dimensional parameter θ be denoted by $l(\theta)$, the maximum likelihood estimator (MLE) be $\hat{\theta}$, the logposterior be $\tilde{l}(\theta)$ and the posterior mode be $\tilde{\theta}$. Also let $\hat{\sigma}^2 = -1/l''(\hat{\theta})$ and $\tilde{\sigma}^2 = -1/\tilde{l}''(\tilde{\theta})$. Then the usual normal approximation to the likelihood is the $N(\hat{\theta}, \hat{\sigma}^2)$ distribution, and the modal normal approximation to the posterior is the $N(\tilde{\theta}, \tilde{\sigma}^2)$ distribution.

2. THE THIRD DERIVATIVE AND TAIL PROBABILITY CRITERIA

The diagnostics used in this article to assess the degree of nonnormality in the posterior distribution are based on the third derivative of the logposterior and the coverage of tail probability regions. This section defines these two measures, from both the likelihood and Bayesian viewpoints, and then gives a theorem that illustrates a relationship between them. The result of the theorem is used to translate a specified level of acceptable normality in terms of the tail probability criterion into an upper bound on the magnitude of the third derivative measure.

2.1 The Third-Derivative Diagnostic

If the likelihood function is normal, then the log-likelihood is quadratic and the third derivative of the log-likelihood is identically zero. Hence any nonzero value for the third derivative of the log-likelihood indicates a departure from normality. Sprott (1973, 1975, 1980) argued that the magnitude of the standardized third derivative of the log-likelihood, given by $\hat{F}_\theta = l'''(\hat{\theta}) \cdot [-l''(\hat{\theta})]^{-3/2} = l'''(\hat{\theta})\hat{\sigma}^3$, can be used to measure nonnormality in the likelihood. The standardization is important for the nonnormality measure because it makes $\hat{F}_\theta$ invariant under affine transformations of the parameter, which preserve normality (and nonnormality).

Similar reasoning leads to a measure for nonnormality in the posterior distribution. Let $\tilde{F}_\theta = \tilde{l}'''(\tilde{\theta}) \cdot [-\tilde{l}''(\tilde{\theta})]^{-3/2} = \tilde{l}'''(\tilde{\theta})\tilde{\sigma}^3$. The measure $\tilde{F}_\theta$ is the first term beyond the modal normal approximation in the Taylor series expansion of the logposterior about the mode, standardized so that it is affine invariant. Note that $\tilde{F}_\theta$ has also been considered by Hills (1989) and Albert (1989) and has been extended to a multivariate measure by Kass (1989) and Kass and Slate (1992, 1993).

If a flat prior is chosen for θ, then $\tilde{F}_\theta = \hat{F}_\theta$. But if the model is then reparameterized and instead expressed in terms of $\phi = \phi(\theta)$, then typically $\tilde{F}_\phi \neq \hat{F}_\phi$ despite the equality for θ. Nevertheless, $\hat{F}_\theta - \tilde{F}_\theta = O(n^{-3/2})$ in general, so that one may study the behavior of the somewhat simpler diagnostic $\hat{F}_\theta$ as an approximation to $\tilde{F}_\theta$. When the parameterization is understood from the context, these standardized third-derivative diagnostics may be referred to simply as $\hat{F}$ and $\tilde{F}$.

2.2 The Tail Probability Criterion

The tail probability criterion compares the probability of tail regions under the posterior and its modal normal approximation. This criterion may be used to evaluate nonnormality in the likelihood by comparing the normalized likelihood to the MLE-based normal approximation. Of course, the normalized likelihood may be obtained as the posterior distribution under a flat prior. When various parameterizations are studied from the likelihood viewpoint, the correspondence requires using a flat prior for each parameterization, which is not an acceptable Bayesian practice in general but is equivalent to using altered data sets and conjugate priors.

In this article, only the lower and upper .05 tail regions of the posterior are considered. Letting F be the posterior cumulative distribution function for θ and letting Φ be the standard normal cumulative distribution function, the values $F(\Phi^{-1}(.05)\tilde{\sigma} + \tilde{\theta})$ and $F(\Phi^{-1}(.95)\tilde{\sigma} + \tilde{\theta})$ are compared to the nominal values of .05 and .95. Acceptable normality in the .05 tails of the posterior distribution for θ is defined in terms of the accuracy of the modal normal approximation according to the following guideline:

> The modal normal approximation to the distribution of the random variable θ, the $N(\tilde{\theta}, \tilde{\sigma}^2)$ distribution, is acceptable if $.025 \leq F(\Phi^{-1}(.05)\tilde{\sigma} + \tilde{\theta}) \leq .10$ and $.90 \leq F(\Phi^{-1}(.95)\tilde{\sigma} + \tilde{\theta}) \leq .975$.

Thus a posterior distribution is adequately normal if the probabilities of the tail regions that have coverage .05 under the modal normal approximation are between half and twice the nominal value.

2.3 The Tail Probability and $\tilde{F}$

Using an approximation given by DiCiccio, Field, and Fraser (1990), a tail probability under the posterior distribution may be expressed in terms of the third-derivative measure $\tilde{F}$. The tail probability criterion for adequate normality of the posterior distribution can then be used to provide an upper bound for the size of $\tilde{F}$ permissible for acceptable normality in the tails of the posterior.

The DiCiccio, Field, and Fraser (DFF) tail probability approximation for a one-dimensional random variable θ is

$$P(\theta \leq t) = \Phi(r) + \varphi(r)\left(\frac{1}{r} + \frac{[-\tilde{l}''(\tilde{\theta})]^{1/2}}{\tilde{l}'(t)}\right) + O(n^{-3/2}), \tag{1}$$

where $\tilde{l}$ is the logposterior, $r = \operatorname{sgn}(t - \tilde{\theta})\{2[\tilde{l}(\tilde{\theta}) - \tilde{l}(t)]\}^{1/2}$, Φ is the standard normal cumulative distribution function, φ denotes the standard normal density, n is the sample size, and it is assumed that $\tilde{l}(\theta)$ is $O(n)$ for each fixed θ and $(t - \tilde{\theta}) = O(n^{-1/2})$. By expanding both $r(t)$ and $\tilde{l}'(t)$ about $\tilde{\theta}$, these quantities can be expressed in terms of the standardized third derivative, as in the following theorem, which is proved in the Appendix.

Theorem. Under the conditions necessary for the validity of the DFF tail probability approximation, the probability of the modal approximation to the p-probability tail region is

$$
\begin{aligned}
P(\theta \le \Phi^{-1}(p)\tilde{\sigma} + \tilde{\theta}) \\
&= \Phi\left(\Phi^{-1}(p)\left[1 - \frac{1}{3}\Phi^{-1}(p)\tilde{F}\right]^{1/2}\right) \\
&+ \frac{\varphi\left(\Phi^{-1}(p)\left[1 - \frac{1}{3}\Phi^{-1}(p)\tilde{F}\right]^{1/2}\right)}{\Phi^{-1}(p)} \\
&\cdot \left\{\frac{\left[1 - \frac{1}{2}\Phi^{-1}(p)\tilde{F}\right] - \left[1 - \frac{1}{3}\Phi^{-1}(p)\tilde{F}\right]^{1/2}}{\left[1 - \frac{1}{2}\Phi^{-1}(p)\tilde{F}\right]\left[1 - \frac{1}{3}\Phi^{-1}(p)\tilde{F}\right]^{1/2}}\right\} \\
&+ O(n^{-1}). \qquad (2)
\end{aligned}
$$

Because the right side of (2) is monotone in $\tilde{F}$ (see Fig. 1), one may find bounds on $\tilde{F}$ corresponding to the bounds on the .05 tail regions specified in Section 2.2. By considering $p = .05$ and $p = .95$ in (2), bounds on the magnitude of $\tilde{F}$ required for acceptable normality according to the guideline may be found. The value of $|\tilde{F}|$ that makes the probability of either of the approximate .05 tail regions twice the nominal value is .384, with $\tilde{F}$ positive for the upper tail and negative for the lower tail. The value of $|\tilde{F}|$ that makes the coverage of either approximate .05 tail region one-half the nominal value is .470, the sign positive for the lower tail and negative for the upper tail. Whether .38 or the more lenient .47 may be used as the upper bound for $\tilde{F}$ for both the upper and lower tails depends on the skewness and tail behavior of the posterior.

Regardless, as shown in Figure 1, the more stringent bound of .38 for $|\tilde{F}|$ will ensure that both the upper and lower nominal .05 tail regions have probability between .025 and .10, to $O(n^{-1})$. Note that the upper bound of .10 on the tail probability is the operational bound, leading to the cutoff of .38 for $|\tilde{F}|$.

Remark 1. It is possible to express the right side of (2) as a cubic in p so that bounds for $|\tilde{F}|$ satisfying more general forms of the guidelines given in Section 2.2 may be easily found for arbitrary p. This was done by Slate (1991, sec. 8.2.1).

Remark 2. Under the conditions assumed thus far, $\tilde{F} - \hat{F} = O(n^{-3/2})$, so that $\tilde{F}$ may be replaced by $\hat{F}$ in (2) and the result remains true. Thus, whether or not a flat prior is used, the upper bound given for $\tilde{F}$ applies equally well to $\hat{F}$, which is somewhat easier to compute. The results given in Section 4 are based on the likelihood, whereas the extensions in Section 5 account for information contained in a conjugate prior.

3. THE MODELS AND PARAMETERIZATIONS

The models considered in this article consist of a random sample of n observations from an NEF-QVF distribution with a conjugate prior on the mean μ. Many interesting properties of the distributions in this family were given by Morris (1982, 1983), only a few of which are reported in Section 3.1. The reparameterizations of μ considered are discussed in Section 3.2.

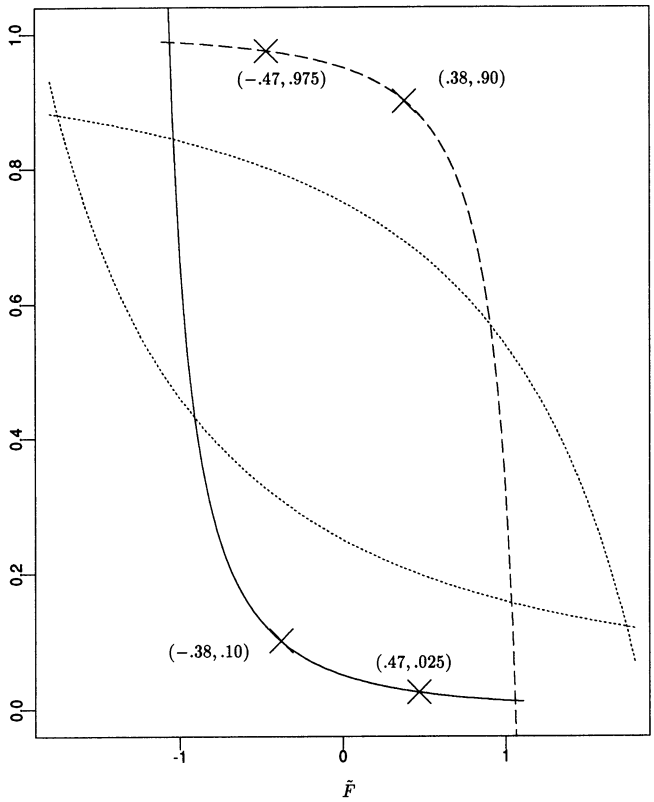

Figure 1. The Behavior of the Right Side of (2) as a Function of $\tilde{F}$. The solid line is (2) with $p = .05$, and the dashed line is for $p = .95$. The crosses mark the points where the approximate tail probabilities achieve the bounds specified in Section 2.2. The dotted lines show the approximation with $p = .25$ (lower line) and $p = .75$. The curves extend beyond the "football-shaped" regions, which rotate counter-clockwise and flatten as p nears .5.

3.1 Natural Exponential Families with Quadratic Variance Functions

A random variable X has an NEF-QVF distribution with mean μ and variance function $V(\mu)$ if the distribution of X given μ is a natural exponential family and the variance of X is at most a quadratic function of the mean, $V(\mu) = \nu_0 + \nu_1\mu + \nu_2\mu^2$. The six NEF-QVF's are the normal(μ, σ^2) with $V(\mu) = \sigma^2\ (=\nu_0)$, Poisson$(\mu)$ with $V(\mu) = \mu$, gamma (a, b) with $\mu = a/b$ and $V(\mu) = \mu^2/a$, binomial(r, p) with $\mu = rp$ and $V(\mu) = \mu - \mu^2/r$, negative binomial(r, p) with $\mu = r(1 - p)/p$ and $V(\mu) = r\mu + \mu^2$, and the generalized hyperbolic secant with $V(\mu) = r + \mu^2/r$, $r > 0$. Of particular interest are what Morris calls the elemental forms of these distributions, the cases that have the leading coefficient of the variance function equal to ± 1: the exponential for the gamma ($a = 1$), Bernoulli for the binomial ($r = 1$), geometric for the negative binomial ($r = 1$), and the case with $r = 1$ for the generalized hyperbolic secant distributions. (If in addition to $r = 1$, $\mu = 0$, the generalized hyperbolic secant distribution becomes the hyperbolic secant distribution, as discussed by Johnson and Kotz [1970]. The modifier "generalized" will often be dropped when reference is made to the generalized hyperbolic secant distribution.)

Two pertinent results from Morris' papers are summarized as follows:

1. $X_i \overset{\text{iid}}{\sim}$ NEF-QVF(μ, $V(\mu)$), $i = 1, \ldots, n$ implies that the sample mean $\bar{X}$ has a NEF-QVF(μ, $V(\mu)/n$) distribution. Hence the log-likelihood for μ based on the random sample has the form $l(\mu) = n\bar{x}\theta(\mu) - n\psi(\theta(\mu))$, where θ is the natural parameter and $\psi(\theta)$ is the cumulant generating function.

2. The conjugate prior on μ has the form $f(\mu) \propto \exp\{m\mu_0\theta(\mu) - m\psi(\theta(\mu)) - \log V(\mu)\}$ with mean μ_0 and variance $V(\mu_0)/(m - \nu_2)$. The prior parameter m is a "convolution" parameter and may be thought of in terms of the sample size equivalent of the information in the conjugate prior. (Recall that analyses under conjugate priors are often equivalent to analyses using a flat prior with additional, contrived data; see Novick and Hall 1965.) The posterior for μ under the conjugate prior has the same form, with $N = n + m$ replacing m and $x_0 = (m\mu_0 + n\bar{x})/N$ replacing μ_0. The posterior mean for μ is x_0, and the posterior variance of μ is $V(x_0)/(N - \nu_2)$.

Morris (1983) also gave the form for the marginal distribution of the sample mean $\bar{X}$ for each of the six NEF-QVF distributions under the conjugate prior for μ. This result is used in Section 5 when considering the average behavior of $\tilde{F}$.

3.2 Reparameterizations

Because the posterior for μ is normal under the normal model, reparameterizations will not be considered for this model. For the other five models, four parameterizations will be of particular interest. These are the vanishing third-derivative parameterization, the variance-stabilizing parameterization, the original mean-value parameterization, and parameterizations within a larger class of power-like transformations.

3.2.1 Reparameterizations for Vanishing Third Derivative. Anscombe (1964) (see also DiCiccio 1984) gave the following expression for the data-independent reparameterization that makes the third derivative of the log-likelihood zero at the MLE: $\gamma(\mu) = \int_a^\mu \exp\{\frac{1}{3}\int_b^t h(u)\,du\}\,dt$, where the function h is defined by $h(\hat{\mu}) = l'''(\hat{\mu})/l''(\hat{\mu})$ and a and b are constants. The sufficiency of $\bar{x}$ in the NEF-QVF models ($\bar{x} = \hat{\mu}$) guarantees that $l'''(\hat{\mu})/l''(\hat{\mu})$ is a function of the data through $\hat{\mu}$ only, although this is generally not the case. (See Hougaard 1982 and Kass 1984 for derivations of γ and other parameterizations in curved exponential and more general families.)

When the data are a random sample from a NEF-QVF(μ, $V(\mu)$) distribution, the function h is $h(u) = -2V'(u)/V(u)$. In this case, then, the vanishing third derivative (of the log-likelihood) reparameterization, $\gamma(\mu)$, satisfies $\gamma(\mu) \propto \int^\mu \{V(t)\}^{-2/3}\,dt$. The form of this parameterization in the six NEF-QVF families is $\gamma(\mu) \propto \mu$ for normal data, $\gamma(\mu) \propto \mu^{1/3}$ for Poisson data, $\gamma(\mu) \propto \mu^{-1/3}$ for gamma data, $\gamma(\mu) \propto \int_0^\mu [t(1-t)]^{-2/3}\,dt$ for Bernoulli data (the incomplete beta (1/3, 1/3) function), $\gamma(\mu) \propto \int_0^\mu [t(1+t)]^{-2/3}\,dt$ for geometric data, and $\gamma(\mu) \propto \int_{-\infty}^\mu [r + t^2/r]^{-2/3}\,dt$ for the hyperbolic secant distribution. As Anscombe (1964) discussed for the Bernoulli model, taking $\gamma(\mu) = \frac{1}{3}\int_0^\mu [t(1-t)]^{2/3}\,dt$ produces a transformation that behaves like $\mu^{1/3}$ when μ is near zero. The parameterization for the geometric distribution may be expressed in terms of an $F_{2/3,2/3}$ cumulative distribution function, whereas that for the hyperbolic secant density may be expressed in terms of a $t_{1/3}$ cumulative distribution function. Note that under the conjugate prior, these parameterizations will also make the third derivative of the logposterior vanish at the mode.

3.2.2 Variance-Stabilizing Parameterization. Although they will clearly not minimize the third-derivative measures, variance-stabilizing parameterizations are of interest because of the promising results obtained by Achcar (1989) and Achcar and Smith (1990). These authors observed that the variance-stabilizing parameterization produces very accurate approximations to posterior moments via Laplace's method in a number of models.

Given the variance function $V(\mu)$, the variance-stabilizing parameterization, η, is found by solving $[\eta'(\mu)]^2V(\mu) = c$ for c a nonzero constant. The solution is $\eta(\mu) \propto \int^\mu V(u)^{-1/2}\,du$. The form of this variance-stabilizing parameterization in each of the six NEF-QVF distributions is $\eta(\mu) \propto \mu$ for normal data, $\eta(\mu) \propto \mu^{1/2}$ for Poisson data, $\eta(\mu) \propto \log\mu$ for exponential data, $\eta(\mu) \propto \arcsin(\mu^{1/2})$ for Bernoulli data, $\eta(\mu) \propto \log(2\mu + 1 + 2(\mu + \mu^2)^{1/2})$ for geometric data, and $\eta(\mu) \propto \log\{\mu + (1 + \mu^2)^{1/2}\}$ for hyperbolic secant data.

3.2.3 Expressions for the Third-Derivative Measures. The parameterization that makes the third derivative of the log-likelihood vanish, the mean value, and the variance-stabilizing parameterization are all of the form $\phi^z(\mu) \propto \int^\mu V(u)^z\,du$, with z taking the value $-\frac{1}{2}$ for the variance-stabilizing parameterization, $-\frac{2}{3}$ for the vanishing $\hat{F}$ parameterization, and zero for the mean-value parameterization. Some results are given here for parameterizations of this form, with explicit expressions shown for the variance-stabilizing and vanishing third-derivative parameterizations.

Because of equivalence of MLE's, it is straightforward to express the standardized third derivative of the log-likelihood in one parameterization in terms of the standardized third derivative in another parameterization. This relationship simplifies the comparison of $\hat{F}$ across reparameterizations. The correspondence for $\mu = \mu(\phi)$ is (Sprott 1973)

$$\hat{F}_\phi = \hat{F}_\mu - \frac{3\left(\dfrac{d^2\mu}{d\phi^2}\right)}{\left(\dfrac{d\mu}{d\phi}\right)^2\left(-\dfrac{d^2l}{d\mu^2}\right)^{1/2}} = \hat{F}_\mu + \frac{3\left(\dfrac{d^2\phi}{d\mu^2}\right)}{\left(\dfrac{d\phi}{d\mu}\right)\left(-\dfrac{d^2l}{d\mu^2}\right)^{1/2}},$$

where the derivatives are evaluated at the appropriate MLE's. For the reparameterization ϕ^z, this expression for the standardized third derivative becomes

$$\hat{F}_{\phi^z} = \frac{V'(\bar{x})(2 + 3z)}{n^{1/2}V(\bar{x})^{1/2}}. \tag{3}$$

Deriving results for the posterior is more complicated due to the lack of invariance of posterior modes. The third derivative of the logposterior in ϕ^z is given by

$$\tilde{F}_{\phi^z} = \frac{V' \cdot (2 + 3z)}{V^{1/2}[N + V'' \cdot (1 + z)]^{1/2}},$$

where V and its derivatives are evaluated at $\mu = \mu(\tilde{\phi}^z)$. Because of the use of the conjugate prior, $\tilde{F}_{\phi^z}$ is $\hat{F}_{\phi^z}$ based on a sample of size $n = N + V''(\mu(\tilde{\phi}))(1 + z)$ with mean $\bar{x} = \mu(\tilde{\phi})$.

3.2.4 Power Transformations. Let the class of power transformations be given by

$$\phi_v(\mu) = \frac{\mu^v - 1}{v}, \qquad v \neq 0$$
$$= \log \mu, \qquad v = 0 \qquad (4)$$

for $\mu > 0$. This class of reparameterizations is clearly applicable in the gamma, Poisson, and negative binomial cases, but variants of the power transformations will be studied for the Bernoulli and generalized hyperbolic secant models.

The standardized third derivative of the log-likelihood at $\hat{\phi}_v$ is $\hat{F}_{\phi_v} = [2\bar{x}V'(\bar{x}) - 3(1 - v)V(\bar{x})]/[V(\bar{x})^{1/2}n^{1/2}\bar{x}]$. The power transformation that will make the standardized third derivative vanish at the MLE may be data dependent. The value of v that yields $\hat{F}_{\phi_v} = 0$ is $v^\diamond = 1 - [2\bar{x}V'(\bar{x})]/[3V(\bar{x})]$. The only variance functions for which $v^\diamond$ is independent of the data are those for which $V'(u)/V(u) = k/u$ for some constant k. Such variance functions satisfy $V(u) \propto u^k$. Because it is assumed that $V(\mu)$ is quadratic, k must be one of 0, 1, or 2. Referring to the variance functions in Section 3.1, only the normal, Poisson, and gamma models have variance functions of this form and hence have reparameterizations within the power class that satisfy the third-derivative criterion and do not depend on the data. For the normal, of course, the power is $v^\diamond = 1$ (no reparameterization necessary); for the Poisson model, the power is $v^\diamond = \frac{1}{3}$; and for the gamma model, the power is $v^\diamond = -\frac{1}{3}$. The behavior of the optimal data-dependent power $v^\diamond$ for the geometric and generalized hyperbolic secant distributions is deferred to Sections 4.3 and 4.5.

Power reparameterizations are not well defined for the hyperbolic secant mean, and they are unsatisfactory for the Bernoulli mean (μ = probability of success $\in (0, 1)$) due to the asymmetric treatment of μ and $1 - \mu$. Therefore, two variants of the power transformation are introduced for these models. A "symmetric" power transformation, given by

$$\psi_v(\mu) = \frac{\left(\dfrac{\mu}{1-\mu}\right)^v - 1}{v}, \qquad v \neq 0$$
$$= \log\left(\frac{\mu}{1-\mu}\right), \qquad v = 0, \qquad (5)$$

is considered for the Bernoulli model. This class includes as a special case when $v = 0$ the natural parameterization, the logit. The "symmetric" property of ψ is that $\psi_v(\mu) = -\psi_{-v}(1 - \mu)$.

The second variant of the usual power transformation is the "exponential" transformation, which handles the problem of $\mu \in \mathbb{R}$ for the hyperbolic secant model by applying the power transformation to $\exp(\mu)$:

$$\xi_v(\mu) = \frac{e^{\mu v} - v}{v}, \qquad v \neq 0$$
$$= \mu, \qquad v = 0. \qquad (6)$$

Here no reparameterization corresponds to $v = 0$ exactly, rather than to an affine transformation of the case $v = 1$ as in the standard power transformation.

4. LIKELIHOOD BEHAVIOR

In this section, the five nonnormal NEF-QVF models are addressed in turn. For each model, the parameterizations described in Section 3 are considered in terms of the magnitude of the third-derivative measure $\hat{F}$ and the probability under the normalized likelihood of the tail regions that have coverage .05 under the normal approximation. For the power-like class of reparameterizations (ϕ_v, ψ_v, or ξ_v), the behavior of $\hat{F}$ as a function of v is explored. Except for the exponential model, the behavior of a parameterization depends not only on the value of the sample size n, but also on the value of $\bar{x}$, and this dependence is also exhibited. Let n^* denote the minimum sample size that is required for acceptable normality in the .05 tails, as defined in Section 2.2. Illustrative tables of n^* are given in the following subsections for each model and parameterization. These values of n^* were computed to an accuracy of .001 but are reported to the next-larger integer, in keeping with the sample size interpretation. The extension of these results to the Bayesian model with a conjugate prior is explained in Section 5.

4.1 The Exponential Model

For the exponential model, the class of power transformations (4) contains both the variance-stabilizing parameterization, $\eta = \phi_0$, and the vanishing third-derivative parameterization, $\gamma = \phi_{-1/3}$. The standardized third-derivative measure for ϕ_v is given by $\hat{F}_{\phi_v} = (1 + 3v)/n^{1/2}$. Because this is a scale model and not a location model, $\hat{F}_{\phi_v}$ does not depend on the location of the data, but only on the sample size.

Similarly, the values of the sample size required for adequate normality in the .05 tails do not depend on $\bar{x}$. A plot of the required sample sizes as a function of the power used for reparameterization (not shown) indicates that n^* is approximately a quadratic function of v. (This relationship may be predicted using the constraint $|\hat{F}_{\phi_v}| \leq \varepsilon$.) The values of n^* for the four parameterizations of primary interest for this model are shown in Table 1. The vanishing third-derivative parameterization is the best of the parameterizations in the sense of requiring the smallest value of n for acceptable normality, needing just one observation. The variance-stabilizing parameterization $\eta = \log \mu$ also performs quite well, with $n^* = 7$. In fact, as long as $n \geq 7$, any power transformation with $v \in [-2/3, 0]$ produces an acceptably normal likelihood. The canonical parameterization, $\phi_{-1} \propto \mu^{-1}$, requires 33 observations for the normal approximation to the likelihood to be accurate in the .05 tails. Each of these parameterizations is a marked improvement over the original mean-value parameterization, for which 114 observations are required.

Table 1. Selected Values of the Sample Sizes n^* and N^* Required for Adequate Normality in the .05 Tails of the Likelihood and Posterior for the NEF-QVF Models Under Various Parameterizations

Distribution	n^* γ	n^* η	n^* μ	n^* Other	N^* γ	N^* η	N^* μ	N^* Other
Exponential	1	7	114	33 ($\phi_{-1} \propto \mu^{-1}$)	2	7	113	34 ($\phi_{-1} \propto \mu^{-1}$)
Poisson ($\bar{x}$ = 1)	1	4	33	7 ($\phi_0 = \log \mu$)	1	4	34	7 ($\phi_0 = \log \mu$)
Geometric ($\bar{x}$ = 2)	1	8	119	2 ($\phi_0 = \log \mu$)	2	8	119	2 ($\phi_0 = \log \mu$)
Geometric ($\bar{x}$ = .5)	2	12	159	4 ($\phi_0 = \log \mu$)	4	13	159	5 ($\phi_0 = \log \mu$)
Bernoulli ($\bar{x}$ = .25)	3	10	65	9 (logit)	4	11	65	8 (logit)
Bernoulli ($\bar{x}$ = .05)	15	63	599	109 (logit)	20	72	599	108 (logit)
GHS ($\bar{x}$ = 1)	1	4	47	10 ($\xi_{-2/3}$)	1	3	36	3 ($\xi_{-2/3}$)
GHS ($\bar{x}$ = 10)	1	7	111	>650 ($\xi_{-2/3}$)	1	7	>5000	>5000 ($\xi_{-2/3}$)

NOTE: The value reported is the smallest integer larger than the computed value of n^* and N^*.

4.2 The Poisson Model

For the Poisson model, the various parameterizations are again all members of the class of power transformations (4) with $\eta = \phi_{1/2}$ and $\gamma = \phi_{1/3}$. The standardized third-derivative measure is $\hat{F}_{\phi_v} = (3v - 1)/(n\bar{x})^{1/2}$. Because the Poisson is a convolution family, the data affect the value of $\hat{F}_{\phi_v}$ through $n\bar{x}$, the product of the sample size and the sample mean, and so this product governs when ϕ_v will be poorly behaved. Thus large values of n can compensate for small values of $\bar{x}$, and vice versa.

The sample sizes needed for acceptable coverage of the approximate .05 tail regions when the sample mean is $\bar{x} = 1$ are shown in Table 1. Although not illustrated in the table, of the parameterizations γ, η and μ, γ produces the smallest values of n^* for all values of $\bar{x}$ considered (in the range .1 to 50). But the reparameterization $\phi_{1/4}$ performs slightly better than γ in this respect, and $\phi_{5/18}$ is better yet.

Note that the bound $|\hat{F}| < \varepsilon$ implies that

$$n\bar{x} > (3v - 1)^2/\varepsilon^2. \qquad (7)$$

As may be expected from the discussion in Section 2.3, for each v the product $n^*\bar{x}$ is approximately constant (the range over the $\bar{x}$'s tried is less than .5 for each v). Using the value of $\varepsilon = .38$ from Section 2.3, (7) may be used to predict values of n^*. The results are given in Table 2, where the computed values of $n^*\bar{x}$ are taken to be the maximum values over all $\bar{x}$. The predicted values typically underestimate the actual values of $n^*\bar{x}$, suggesting that the bound for $|\hat{F}_{\phi_v}|$ obtained from the tail probability approximation (2) is a little too large for the Poisson model.

4.3 The Geometric Model

The geometric distribution has probability mass function given by $P(X = x|p) = p(1 - p)^x$ for $p \in (0, 1)$ and x a nonnegative integer. The mean is given by $\mu = (1 - p)/p$. For this model, the power reparameterizations (4) are again considered, but the variance-stabilizing and vanishing third-derivative parameterizations are not included in this class. The variance-stabilizing parameterization is $\eta(\mu) = \log(2\mu + 1 + 2(\mu + \mu^2)^{1/2})$, and the vanishing third-derivative parameterization is taken to be $\gamma(\mu) = [\Gamma(1/3)^2/\Gamma(2/3)]G(\mu)$, where G is the $F_{2/3,2/3}$ cumulative distribution function.

The standardized third derivatives for η and ϕ_v are given by

$$\hat{F}_\eta = \frac{1 + 2\bar{x}}{2n^{1/2}\bar{x}^{1/2}(1 + \bar{x})^{1/2}}$$

and

$$\hat{F}_{\phi_v} = \frac{2(1 + 2\bar{x}) - 3(1 - v)(1 + \bar{x})}{\bar{x}^{1/2}(1 + \bar{x})^{1/2}n^{1/2}}.$$

The expressions for $\hat{F}$ show that small values of $\bar{x}$ will produce standardized third derivatives with large magnitude. One noteworthy relationship is that the standardized third derivative for the mean-value parameterization (an affine transformation of ϕ_1) has an $\hat{F}$ value four times that for η. The data-dependent power $v^\diamond = (1/3)(1 - \bar{x})/(1 + \bar{x})$ gives $\hat{F}_{\phi_{v^\diamond}} = 0$, so that values of $\bar{x}$ near zero suggest the cube root parameterization, values near 1 indicate that the log transformation will perform well, and the inverse cube root parameterization is recommended for large values of $\bar{x}$.

Some values of the sample size needed for adequate coverage of the approximate .05 tails are given in Table 1. The vanishing third-derivative parameterization, γ, performs the best of the parameterizations considered, requiring only $n^* = 3$ for $\bar{x}$ as small as $\frac{1}{2}$. Again, the variance-stabilizing parameterization is next best, and both of these parameterizations are marked improvements over the original mean-value parameterization.

4.4 The Bernoulli Model

The Bernoulli model consists of a random sample of size n from the Bernoulli distribution with mean μ (the probability of success). The parameterizations of interest for the Bernoulli model are the variance-stabilizing parameteriza-

Table 2. Lower Bounds on the Product $n^*\bar{x}$ for Adequate Likelihood Normality for the Poisson Model

v	$n^*\bar{x}$ from (7)	Computed $n^*\bar{x}$
0	6.9	7.0
1/4	.43	.5
1/3 (γ)	0	1.0
1/2 (η)	1.7	3.5
2/3	6.9	9.5
3/4	10.8	14.0
1 (μ)	27.7	33.0

NOTE: The values in the first column are predicted using the lower bound of .38 for $\hat{F}$ while those in the second column satisfy the coverage criterion.

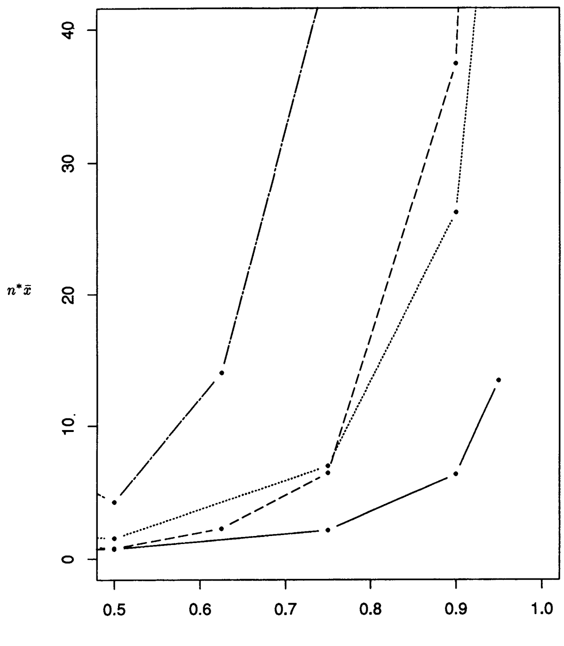

Figure 2. Required Number of Successes for Adequate Coverage of Both .05 Tails for the Bernoulli Model. The values are shown for the mean-value (dashed and dotted line), incomplete beta (solid line), arcsine square root (dotted line), and logit (dashed line) parameterizations. The values of $n^*\bar{x}$ are symmetric about $\bar{x}$ = .5.

tion, the vanishing third-derivative parameterization, and the class of symmetric power transformations (5). The variance-stabilizing parameterization is taken to be $\eta(\mu) = \arcsin(\mu^{1/2})$, which behaves like $\mu^{1/2}$ for μ near zero. The vanishing third-derivative parameterization is an incomplete beta integral given by $\gamma(\mu) = (1/3)[\Gamma(1/3)^2/\Gamma(2/3)]G(\mu)$, where G is the beta(1/3, 1/3) cumulative distribution function. When μ is near zero, $\gamma(\mu)$ behaves like $\mu^{1/3}$. Recall that the logit transformation is a member of the class of symmetric transformations given by ψ_0.

The standardized third-derivative diagnostics for the variance-stabilizing and symmetric power transformations are given by

$$\hat{F}_\eta = \frac{1 - 2\bar{x}}{2n^{1/2}\bar{x}^{1/2}(1 - \bar{x})^{1/2}}$$

and

$$\hat{F}_{\psi_v} = \frac{1 - 2\bar{x} - 3v}{n^{1/2}\bar{x}^{1/2}(1 - \bar{x})^{1/2}}.$$

The power $v^\diamond = (1 - 2\bar{x})/3$ will make $\hat{F}_{\psi_{v^\diamond}}$ vanish. The optimal power for ψ_v ranges from $\frac{1}{3}$ to $-\frac{1}{3}$ as $\bar{x}$ moves from zero to 1. Because $\hat{F}_\eta = \frac{1}{2}\hat{F}_{\psi_0}$, there is evidence from the third-derivative diagnostic that the arcsine square root parameterization is better behaved than the logit (though both are inferior to the incomplete gamma). Note that for both the variance-stabilizing parameterization and the logit parameterization (ψ_0), the value of $\hat{F}$ is zero when $\bar{x}$ = .5. For all of these reparameterizations, the extreme values of $\bar{x}$ (those near zero and 1) will be problematic.

Values of the sample size required for adequate normality in both .05 tails of the normalized likelihood for $\bar{x}$ = .05 and $\bar{x}$ = .25 are given in Table 1; Figure 2 shows the required number of successes $n^*\bar{x}$. These values are symmetric about $\bar{x}$ = .5. The vanishing third-derivative parameterization requires the smallest sample size, except when $\bar{x}$ = .5, in which case the logit transformation requires slightly less. Whether the arcsine square root or the logit transformation is better depends on how extreme $\bar{x}$ is. For $|\bar{x} - .5| < .3$, roughly, the logit parameterization performs better than the variance-stabilizing parameterization, whereas when $|\bar{x} - .5| > .3$, the arcsine square root parameterization is better.

Figure 3 shows these four parameterizations in a way that highlights these relationships. Each reparameterization has first been standardized so that at μ = .5, the value of the transformation is .5 and the slope is 1. Then the deviations of these standardized transformations from the standardized incomplete beta parameterization have been displayed as deviations from zero; that is, the curves plotted are $g(\mu) = \gamma_1^{-1}(f(\mu)) - \mu$, where γ_1 is the standardized incomplete beta transformation for $f(\mu)$ the standardized versions of $\gamma(\mu)$, $\eta(\mu)$, logit(μ), and μ, with the exception that $-g(\mu)$ is shown when $f(\mu)$ is the standardized logit function. Because $g(\mu) = -g(\mu - .5)$ for $\mu > .5$, these curves are only shown for $\mu \leq .5$. In this figure the incomplete beta parameterization appears as a horizontal line at zero. Again, the arcsine square root shows the least deviation and the original mean-value parameterization shows the most deviation from the incomplete beta parameterization.

4.5 The Generalized Hyperbolic Secant Model

The natural exponential family with variance function $V(\mu) = 1 + \mu^2$ is a class of generalized hyperbolic secant

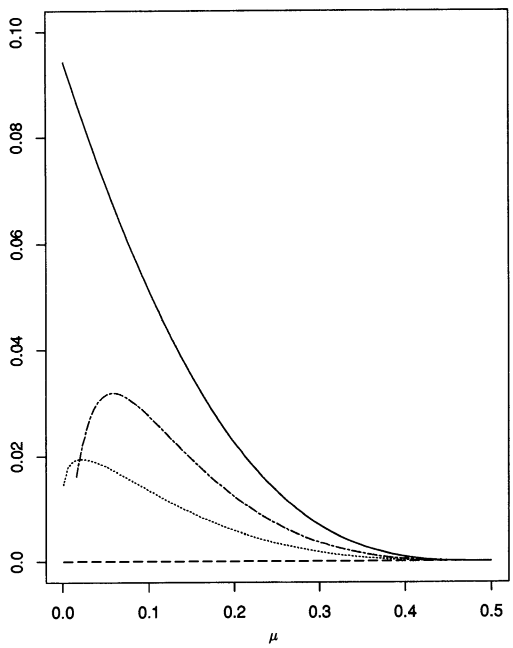

Figure 3. The curves are $g(\mu) = \gamma^{-1}(f(\mu)) - \mu$ where γ is the incomplete beta function and f is the identity (solid line), incomplete beta (dashed line), arcsine square root (dotted line), or logit function (dashed and dotted line), except that $-g(\mu)$ is shown when f is the logit function. The functions f have been standardized so that at μ = .5 the value is .5 and the slope is 1.

distributions. The density with respect to Lebesgue measure of a single observation from this distribution is

$$f_1(x|\mu) = \frac{\exp\{x \arctan(\mu)\}\cos(\arctan(\mu))}{2\cosh\left(\dfrac{\pi x}{2}\right)}$$

for μ, $x \in \mathbb{R}$. When $\mu = 0$, this density is known as the hyperbolic secant distribution. Although the density of $\bar{X}$ must be expressed as an infinite product, the likelihood can be written directly using the result of Morris (Sec. 3.1), which states that the density of the sample mean from a NEF-QVF(μ, $V(\mu)$) is NEF-QVF(μ, $V(\mu)/n$). Hence the likelihood for μ based on the observed sample mean $\bar{x}$ is $L(\mu) \propto \exp\{n\bar{x}\arctan(\mu)\}(1+\mu^2)^{-n/2}$.

The variance-stabilizing parameterization is given by $\eta(\mu) = \log\{\mu + (1+\mu^2)^{1/2}\}$, which takes values in $\mathbb{R}$. The vanishing third-derivative parameterization is given by $\gamma(\mu) = \text{beta}(1/2, 1/6)P[T \le (\mu/3^{1/2})]$, where $T \sim t_{1/3}$. The vanishing third-derivative parameterization γ takes values in $[0, \text{beta}(1/2, 1/6)]$, which is approximately $[0, 7.29)$. The third parameterization considered for this model is a generalization of the class of power transformations to the real line, given by $\xi_v(\mu)$ in (6). The range of ξ_v is $(-1/v, \infty)$, $\mathbb{R}$, or $(-\infty, -1/v)$, depending on whether v is greater than, equal to, or less than zero.

The expression for $\bar{F}_\eta$ is complicated and not very enlightening and so is not reported here (interested readers may consult Slate 1991). For the ξ_v parameterization, however, the standardized third derivative of the log-likelihood is given by

$$\hat{F}_{\xi_v} = \frac{4\bar{x} + 3v(1+\bar{x}^2)}{n^{1/2}(1+\bar{x}^2)^{1/2}},$$

which produces as optimal data-dependent power $v^\diamond = -4\bar{x}/[3(1+\bar{x}^2)]$. The behavior of this power is that $v^\diamond$ is $\pm 2/3$ when $\bar{x} = \mp 1$ and zero when $\bar{x}$ is zero and as $|\bar{x}|$ approaches infinity. Thus it is expected that the $-2/3$ power will perform well for positive $\bar{x}$, the $2/3$ power will work well for negative $\bar{x}$, and the original mean-value parameterization ($v = 0$) will perform well when $\bar{x}$ is very near or very far from zero.

Selected values of n^* for the parameterizations are shown in Table 1. For the variance-stabilizing parameterization, n^* is symmetric in $\bar{x}$, so that $\bar{x}$ and $-\bar{x}$ produce the same required sample size. As $\bar{x}$ moves away from zero, the value of n^* grows, but $n = 7$ is sufficient for adequate coverage of both .05 tails as long as $|\bar{x}| \le 20$. The variance-stabilizing parameterization appears to work quite well, in the sense of small n^*, for the hyperbolic data.

The vanishing third-derivative parameterization again performs better than the variance stabilizing parameterization. The values of n^* were found to be less than one for every value of $\bar{x}$ tried. The remarkable performance of the vanishing third-derivative parameterization for this model is further illustrated in Figure 4. Normalized likelihoods are shown for both γ and the original parameterization μ for $n = 1$ and $\bar{x} = 2$.

Investigation of the values of n^* needed for the ψ_v parameterization shows that the third-derivative criteria predicted the best powers quite accurately. The original mean-value parameterization is optimal when $\bar{x} = 0$ and the $-2/3$ power works well for positive values of $\bar{x}$, corresponding to the $2/3$ power for negative $\bar{x}$. Hendl (1987) also considered the ψ_v transformation for random variables following a Pearson Type IV distribution and focused on two estimates of v originally proposed by Manly (1976). A comparison of the Manly estimates of v to those obtained using the criteria discussed here shows that their behavior is very similar.

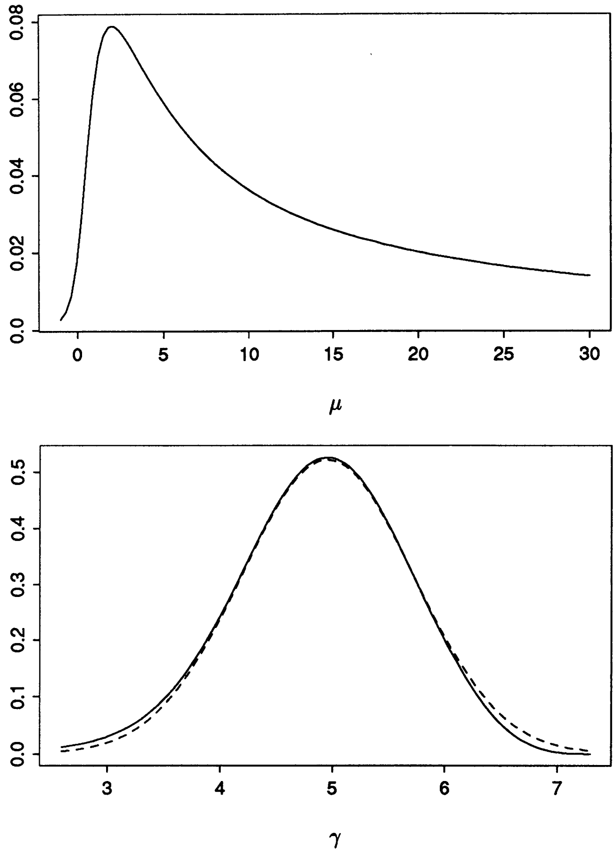

Figure 4. Likelihood and Normal Approximation to It for the Vanishing Third Derivative Parameterization (γ) of the Generalized Hyperbolic Secant Model. The top graph shows the likelihood in the original parameter μ for $n = 1$ and $\bar{x} = 2$. In the lower plot, the solid line is the corresponding likelihood in the γ parameterization and the dashed line is the normal approximation.

5. EXTENSION TO THE BAYESIAN MODEL

The equivariance of the likelihood under reparameterization makes the study of parameterizations from a likelihood perspective quite appealing. But a study of parameterization behavior from a Bayesian viewpoint is especially relevant, because the choice of parameterization is particularly critical for small sample sizes, which is often the situation when prior information most matters. As noted by Tibshirani and Wasserman (1989), parameterizations that produce a well-behaved likelihood need not produce a well-behaved posterior in the same sense. But a correspondence between the behavior of a parameterization under likelihood and Bayesian analyses may be expected if a conjugate prior is assumed. This section considers the nature of this correspondence.

5.1 The Effective Sample Size

The sample size necessary for adequate posterior normality under a conjugate prior analysis for the NEF-QVF distributions clearly depends on the amount of information contained in the prior distribution. The use of a conjugate prior is equivalent to using a flat prior and an augmented data set for a wide class of reparameterizations (see Novick and Hall 1965). Thus a Bayesian study may rank the parameterizations for the NEF-QVF models according to the value of an *effective sample size* that accounts for both the observed data and the prior information.

The effective sample size is directly related to $N = n + m$, as defined in Section 3.1, but an origin must be chosen so that a value of zero for the effective sample size corresponds to an absence of information in some sense. The definition of the effective sample size used here relies on the fact that Jeffreys's prior for each of these models is a member, or at least a limit of members, of the class of conjugate priors for μ. Jeffreys's prior, which is given by $\pi(\mu) \propto i(\mu)^{1/2}$ where $i(\mu)$ is the Fisher information, is taken to be the prior corresponding to a prior sample size of zero. Hence the definition of the effective sample size adopted here is as follows:

> The *effective sample size,* denoted by N^*, is the size of the data set that produces the same posterior under Jeffreys's prior as is produced under the conjugate prior with convolution parameter m when a sample of size n is observed.

As an illustration, consider the Bernoulli model with $\mu = p$, the probability of success. Jeffreys's prior is $\pi(p) \propto p^{-1/2}(1 - p)^{-1/2}$, whereas the conjugate prior is $f(p) \propto p^{m\mu_0-1}(1 - p)^{m(1-\mu_0)-1}$. Consequently, if a sample of size n is observed with sample mean $\bar{x}$, then the posterior under Jeffreys's prior is $\pi(p \mid n, \bar{x}) \propto p^{n\bar{x}-1/2}(1 - p)^{n(1-\bar{x})-1/2}$, whereas the posterior under the conjugate prior is $f(p \mid n, \bar{x}) \propto p^{n\bar{x}+m\mu_0-1}(1 - p)^{n+m-n\bar{x}-m\mu_0-1}$. The effective sample size is the size of the data set with mean $\bar{x}^*$, say, for which $\pi(p \mid N^*, \bar{x}^*) = f(p \mid n, \bar{x})$. Thus N^* and $\bar{x}^*$ must satisfy $N^*\bar{x}^* - \frac{1}{2} = n\bar{x} + m\mu_0 - 1$ and $N^*(1 - \bar{x}^*) - \frac{1}{2} = n + m - n\bar{x} - m\mu_0 - 1$. The solution is $\bar{x}^* = (n\bar{x} + m\mu_0 - \frac{1}{2})/N^*$ and $N^* = n + m - 1$. Hence for the Bernoulli model, $N^* = N - 1$.

Result. The value of N^* for each of the NEF-QVF models is given by $N^* = N + \nu_2$, where ν_2 is the coefficient of μ^2 in the variance function $V(\mu)$.

Thus N^* is $N + 1$ for the exponential, geometric, and generalized hyperbolic secant models, $N - 1$ for the Bernoulli model, and N for the Poisson model.

Although the conjugate prior analysis yielding a posterior indexed by $N = n + m$ and posterior mean x_0 is (under the "likelihood reparameterizations" of Novick and Hall) equivalent to a flat prior analysis for some data set of size n_* with mean $\bar{x}_*$, say, the values for both n_* and $\bar{x}_*$ depend on the parameterization, in general. For example, it is straightforward to show that the conjugate posterior for the power parameterization ϕ_v of the exponential model based on the values $N + 1 = N^*$ and x_0 is proportional to the likelihood with $n_* = N^* + v$ and $\bar{x}_* = Nx_0/N^*$. The correspondences for the other models and parameterizations are shown in Table 3, with the exception of the ξ_v parameterization for the generalized hyperbolic secant model, which is not a likelihood parameterization.

Table 3. A Conjugate Prior Analysis With Effective Sample Size N^* and Posterior Mean x_0 Equivalent to a Likelihood Analysis Using a Sample of Size n_* With Mean $\bar{x}_*$

Model		n_*	$\bar{x}_*$
Exponential	ϕ_v	$N^* + v$	Nx_0/n_*
Poisson	ϕ_v	N^*	$x_0 - v/n_*$
Geometric	ϕ_v	$N^* + v$	$(Nx_0 - v)/n_*$
	γ	$N^* - 1/3$	$(Nx_0 - 1/3)n_*$
	η	N^*	$(Nx_0 - 1/2)/n_*$
Bernoulli	ψ_v	$N^* + 1$	$(Nx_0 - v)/n_*$
	γ	$N^* + 1/3$	$(Nx_0 - 1/3)/n_*$
	η	N^*	$(2Nx_0 - 1)/(2n_*)$
Generalized Hyperbolic Secant	γ	$N^* - 1/3$	Nx_0/n_*
	η	N^*	Nx_0/n_*

NOTE: Recall that $N^* = N + \nu_2$.

5.2 A Bayesian View of the NEF-QVF Parameterizations

Because a change in parameterization necessitates a change in the data set, the translation of the likelihood results of Section 4 does not provide a satisfying description of the behavior of these parameterizations in a conjugate Bayesian analysis. Some illustrative results from a Bayesian study of the parameterizations using a conjugate prior for the mean-value parameterization are given here.

The Bayesian results use as criteria the third-derivative measure $\tilde{F}$ and posterior tail probabilities as defined in Section 2. For each of the models and parameterizations, values of the effective sample size N^* required for adequate posterior normality were computed with x_0 fixed at the same values chosen for $\bar{x}$ in the likelihood analyses. Selected results, shown in Table 1, are similar to those for the likelihood models, with N^* differing from n^* by no more than 5 for the γ parameterization, but by potentially much more for the more poorly behaved parameterizations.

Because of the use of the conjugate prior, $\tilde{F}(N, x_0) = \hat{F}(n_*, \bar{x}_*)$ and the results of Section 4 for $\hat{F}$ apply for $\tilde{F}$ with the strong caveat that n_* and $\bar{x}_*$ may depend on the parameterization. The Bayesian analysis permits additional observations about $\tilde{F}$. First, there is a bonafide marginal distribution for the data that can be used to evaluate the average behavior of $\tilde{F}$. Consider the geometric model, for example, where the marginal distribution of $n\bar{X}$ is a Beta-Pascal distribution. Figure 5 shows the average value of $|\tilde{F}_{\phi_v}|$ as the power chosen for reparameterization and the prior mean (μ_0) vary. The "sample sizes" have been set to $n = 9$ and $m = 2$, giving $N^* = 12$. As μ_0 increases, the best power moves from near $1/4$ to near $-1/4$. This behavior suggests that the log transformation can be expected to perform well, though not optimally, in most situations. For comparison, Table 4 gives the corresponding values of the average standardized third derivative of the logposterior for the variance-stabilizing parameterization. The log transformation has consistently lower values of the average standardized third derivative than

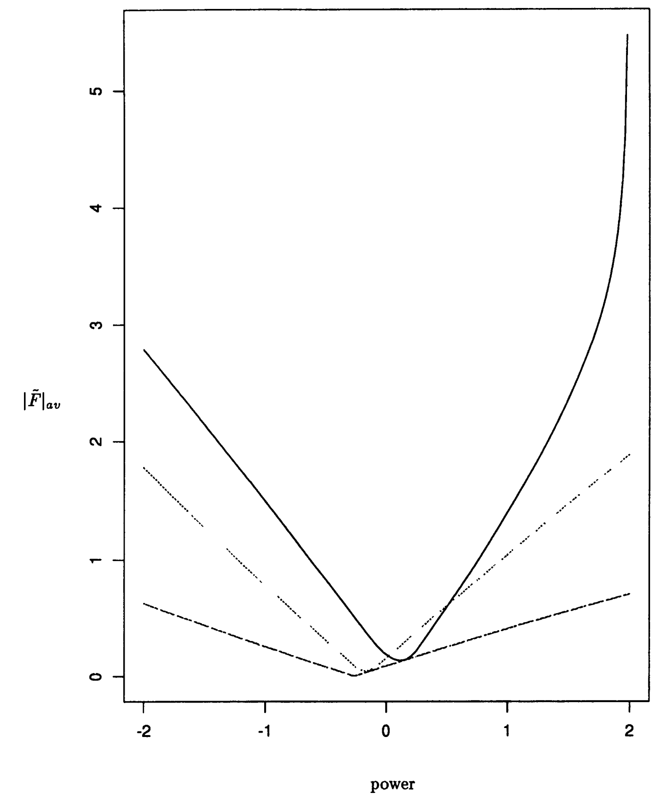

Figure 5. The Standardized Third Derivative of the Logposterior Averaged Over the Marginal Distribution of the Data for the Geometric Data Model. The solid line shows the behavior when $\mu_0 = 1$, the dotted line is for $\mu_0 = 5$, and the dashed line illustrates the relationship when $\mu_0 = 20$. The sample size is $n = 9$, and the prior parameter m is 2 in each case. The power 1 corresponds to no transformation.

the variance-stabilizing parameterization, supporting the smaller values for N^* shown in Table 1.

A second observation of interest is empirical investigation of the relationship between $\tilde{F}$ and the minimum effective sample size that satisfies the tail probability criterion. For the NEF-QVF models, actual coverage probabilities of .025 and .10 for the nominal .05 probability tails correspond to typical values of $\tilde{F}$ between .38 and .43. A consequence is that the minimum effective sample size required for adequate normality in the .05 tails may be predicted using the constraint $|\tilde{F}| < .38$. Prediction of n^* from $|\hat{F}| < .38$ was described for the Poisson model in Section 4.2; the results are similar using $|\tilde{F}| < .38$.

6. DISCUSSION

This article has used a comparison of tail probabilities and a third-derivative measure to evaluate parameterizations according to the nearness to normality of the posterior distribution. The theorem given in Section 2.3 relates these two criteria, enabling a definition of adequate normality in terms of tail probabilities to be expressed in terms of an upper bound on the magnitude of the third-derivative diagnostic. The parameterization criteria were investigated for models based on the class of NEF-QVF distributions with an emphasis on the variance-stabilizing and vanishing third-derivative parameterizations. The parameterizations were ranked according to the minimum values of an effective sample size required for acceptable normality according to the guideline specified in Section 2.2.

For each of the NEF-QVF models, the vanishing third-derivative parameterization was found to give excellent results under the tail probability criterion for the range of data values, though there was often another parameterization among those considered that dominated it for particular sample locations. In the cases where the vanishing third-derivative parameterization (γ) was within the class of power transformations defined for the model, powers near the power corresponding to γ were the next-best parameterizations. For these models, the third derivative is an accurate predictor of these parameterizations' performance under the coverage criterion. The variance-stabilizing parameterization generally took a very respectable third place, still satisfying the coverage criterion with reasonably small values of the effective sample size. This observation agrees with those of Achar (1989) and Achcar and Smith (1990), who found that the variance-stabilizing parameterization tended to enhance performance of Laplace's method (Tierney and Kadane 1986) for approximating posterior moments. It may be possible to formalize this relationship by showing that the difference between the third derivative of the logposterior and the derivative of the information tends to be small.

One would expect that the results would be at least qualitatively the same if the coverage criterion were changed to specify that the probability in the approximate .025 tails be between .0125 and .05. Results for the Poisson model (not given here) support this claim, although the values of N^* needed for adequate normality in both .025 tails were found to be roughly twice the values required for the .05 tails.

Although this article has focused on the NEF-QVF distributions, one may extend the results to natural exponential family distributions with cubic variance functions as discussed by Letac and Mora (1990) or to the more general power variance function families of Bar-Lev and Enis (1986). The cubic variance function families include the inverse Gaussian, for example, which is arguably of more interest than the generalized hyperbolic secant distribution. A conjugate Bayesian analysis for the inverse Gaussian distribution was discussed by Banerjee and Bhattacharyya (1976). Using a prior for μ proportional to $\mu^{-2}\exp\{-\gamma\beta(\mu^{-1} - \beta^{-1})^2/2\}$ shows that the vanishing third-derivative parameterization ($\gamma \propto \mu^{-1}$) requires approximately 40% of the observations the variance-stabilizing ($\eta \propto \mu^{-1/2}$) needs to meet the .05 tail probability criterion (n = 3, 18, 37 for γ versus n = 10,

Table 4. Values of the Standardized Third Derivative of the Logposterior Averaged Over the Marginal Distribution of the Data for the Geometric Model

Parameter	$\mu_0 = 1$	$\mu_0 = 5$	$\mu_0 = 20$
η	.344	.269	.105
log	.186	.154	.088
$v = -1/4$	.485	.075	.006
$v = 1/4$	.219	.379	.169

NOTE: The parameterization η is variance stabilizing, and v indicates the power for a power reparameterization. In all cases, $n = 9$ and $m = 2$.

36, 94 for η for $\bar{x} = 1, 5, 10$). In addition, the mean-value parameterization requires approximately 25 times as many observations as η ($n = 237$, 1,185, 2,371).

An important extension of the sample size requirements and calibration for the third-derivative measure presented in this article is their application in more complex multi-parameter models. Research currently in progress indicates that there are simple but nevertheless useful heuristics based on the one-parameter results that provide suggested sample sizes and third-derivative calibrations for a fairly wide class of multidimensional models.

APPENDIX: PROOF OF THE THEOREM

The relationship given by (2) in Section 2.3 is established in this Appendix. By expanding $\tilde{l}(t)$ and $\tilde{l}'(t)$ about $\tilde{\theta}$, $r(t)$ and $\tilde{l}'(t)$ can be expressed in terms of the standardized third derivative:

$$r = \operatorname{sgn}(t-\tilde{\theta})\cdot\left\{2\left[-\frac{1}{2}(t-\tilde{\theta})^2\tilde{l}''(\tilde{\theta}) - \frac{1}{6}(t-\tilde{\theta})^3\tilde{l}'''(\tilde{\theta}) - \frac{1}{24}(t-\tilde{\theta})^4\tilde{l}^{(4)}(\tilde{\theta}) + O(n^{-3/2})\right]\right\}^{1/2}$$
$$= \left(\frac{t-\tilde{\theta}}{\tilde{\sigma}}\right)\left[1 - \frac{1}{3}\left(\frac{t-\tilde{\theta}}{\tilde{\sigma}}\right)\tilde{\sigma}^3\tilde{l}'''(\tilde{\theta}) - \frac{1}{12}\left(\frac{t-\tilde{\theta}}{\tilde{\sigma}}\right)^2\tilde{\sigma}^4\tilde{l}^{(4)}(\tilde{\theta}) + O(n^{-3/2})\right]^{1/2}$$
$$= \left(\frac{t-\tilde{\theta}}{\tilde{\sigma}}\right)\left[1 - \frac{1}{3}\left(\frac{t-\tilde{\theta}}{\tilde{\sigma}}\right)\tilde{F} - \frac{1}{12}\left(\frac{t-\tilde{\theta}}{\tilde{\sigma}}\right)^2\tilde{\sigma}^4\tilde{l}^{(4)}(\tilde{\theta})\right]^{1/2} + O(n^{-3/2})$$

$$\tilde{\sigma}\tilde{l}'(t) = -\left(\frac{t-\tilde{\theta}}{\tilde{\sigma}}\right)\left[1 - \frac{1}{2}\left(\frac{t-\tilde{\theta}}{\tilde{\sigma}}\right)\tilde{F} - \frac{1}{6}\left(\frac{t-\tilde{\theta}}{\tilde{\sigma}}\right)^2\tilde{\sigma}^4\tilde{l}^{(4)}(\tilde{\theta})\right] + O(n^{-3/2}). \quad \text{(A.1)}$$

Using these expressions in the DiCiccio, Field, and Fraser (DFF) tail probability approximation (1) gives

$$P(\theta \le t) = \Phi\left(\zeta\left[1 - \frac{1}{3}\zeta\tilde{F} - \frac{1}{12}\zeta^2\tilde{\sigma}^4\tilde{l}^{(4)}(\tilde{\theta})\right]^{1/2}\right) + \varphi\left(\zeta\left[1 - \frac{1}{3}\zeta\tilde{F} - \frac{1}{12}\zeta^2\tilde{\sigma}^4\tilde{l}^{(4)}(\tilde{\theta})\right]^{1/2}\right)$$
$$\cdot\frac{1}{\zeta}\left\{\frac{1}{\left[1 - \frac{1}{3}\zeta\tilde{F} - \frac{1}{12}\zeta^2\tilde{\sigma}^4\tilde{l}^{(4)}(\tilde{\theta})\right]^{1/2}} - \frac{1}{\left[1 - \frac{1}{2}\zeta\tilde{F} - \frac{1}{6}\zeta^2\tilde{\sigma}^4\tilde{l}^{(4)}(\tilde{\theta})\right]}\right\} + O(n^{-3/2}),$$

where $\zeta = (t-\tilde{\theta})/\tilde{\sigma}$. Thus, to maintain the accuracy of $O(n^{-3/2})$ of the DFF approximation, the fourth derivatives of the logposterior must be included in the expansion. This result emphasizes the local nature of the third derivative as a measure of nonnormality.

When the terms involving the fourth derivative of the logposterior are dropped from the expansions, the order of the approximation reduces to $O(n^{-1})$ and takes the form

$$P(\theta \le t) = \Phi\left(\zeta\left[1 - \frac{1}{3}\zeta\tilde{F}\right]^{1/2}\right) + \varphi\left(\zeta\left[1 - \frac{1}{3}\zeta\tilde{F}\right]^{1/2}\right)\cdot\frac{1}{\zeta}\left\{\frac{1}{\left[1 - \frac{1}{3}\zeta\tilde{F}\right]^{1/2}} - \frac{1}{\left[1 - \frac{1}{2}\zeta\tilde{F}\right]}\right\} + O(n^{-1}). \quad \text{(A.2)}$$

Note that when $\tilde{F} = 0$, (A.2) gives

$$P(\theta \le t) = \Phi\left(\frac{t-\tilde{\theta}}{\tilde{\sigma}}\right) + O(n^{-1}),$$

which illustrates the improvement over the usual $O(n^{-1/2})$ normal approximation gained when $\tilde{F} = 0$. If the third derivative of $\tilde{l}$ is forced to vanish in (1), then the result is

$$P(\theta \le t) = \Phi(r) + O(n^{-1}).$$

Thus, as can be seen from (A.1) as well, $r = (t-\tilde{\theta})/\tilde{\sigma} + O(n^{-1})$ when the third derivative vanishes.

[Received September 1992. Revised December 1993.]

REFERENCES

Achcar, J. A. (1989), "Reparameterizations and Laplace Approximations for Posterior Moments in Binomial Models," *Revista de Matematica Estatistica,* 7, 73–86.

Achcar, J. A., and Smith, A. F. M. (1990), "Aspects of Reparameterization in Approximate Bayesian Inference," in *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard,* eds. S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner, Amsterdam: North-Holland.

Albert, J. H. (1989), "On the Use of Power Transformations to Simplify Posterior Distributions," technical report, Bowling Green State University.

Anscombe, F. J. (1964), "Normal Likelihood Functions," *Annals of the Institute of Statistical Mathematics,* 16, 1–19.

Banerjee, A. K., and Bhattacharyya, K. (1976), "A Purchase Incidence Model With Inverse Gaussian Interpurchase Times," *Journal of the American Statistical Association,* 71, 823–829.

Bar-Lev, S. K., and Enis, P. (1986), "Reproducibility and Natural Exponential Families With Power Variance Functions," *The Annals of Statistics,* 14, 1507–1522.

DiCiccio, T. J. (1984), "On Parameter Transformations and Interval Estimation," *Biometrika,* 71, 477–485.

DiCiccio, T. J., Field, C. A., and Fraser, D. A. S. (1990), "Approximations of Marginal Tail Probabilities and Inference for Scalar Parameters," *Biometrika,* 77, 77–95.

Hendl, J. (1987), "A Note on Exponential Transformations," *The Statistician,* 36, 55–57.

Hills, S. (1989), "The Parameterisation of Statistical Models," Ph.D. thesis, University of Nottingham.

Hougaard, P. (1982), "Parameterizations of Nonlinear Models," *Journal of the Royal Statistical Society,* Ser. B, 44, 244–252.

Johnson, N. I., and Kotz, S. (1970), *Continuous Univariate Distributions,* Vol. 2, Boston: Houghton-Mifflin.

Kass, R. E. (1989), "The Geometry of Asymptotic Inference," *Statistical Science,* 4, 188–234.

——— (1984), "Canonical Parameterizations and Zero Parameter-Effects Curvature," *Journal of the Royal Statistical Society,* Ser. B, 46, 86–92.

Kass, R. E., and Slate, E. H. (1994), "Some Diagnostics of Maximum Likelihood and Posterior Non-Normality," submitted to *The Annals of Statistics.*

Letac, G., and Mora, M. (1990), "Natural Real Exponential Families with Cubic Variance Functions," *The Annals of Statistics,* 18, 1–37.

Manly, B. F. J. (1976), "Exponential Data Transformations," *The Statistician,* 25, 37–42.

Morris, C. N. (1983), "Natural Exponential Families With Quadratic Variance Functions: Statistical Theory," *The Annals of Statistics,* 11, 515–529.

——— (1982), "Natural Exponential Families With Quadratic Variance Functions," *The Annals of Statistics,* 10, 65–80.

Naylor, J. C., and Smith, A. F. M. (1982), "Applications of a Method for the Efficient Computation of Posterior Distributions," *Applied Statistics,* 31, 214–225.

Novick, M. R., and Hall, W. J. (1965), "A Bayesian Indifference Procedure," *Journal of the American Statistical Association,* 60, 1104–1117.

Ross, G. J. S. (1978), "Exact and Approximate Confidence Regions for Functions of Parameters in Non-Linear Models," in *COMPSTAT* 78, Third Symposium on Computation, Vienna: Physica-Verlag.

Slate, E. H. (1991), "Reparameterization of Statistical Models," Ph.D. thesis, Carnegie Mellon University, Dept. of Statistics.

Sprott, D. A. (1980), "Maximum Likelihood in Small Samples: Estimation in the Presence of Nuisance Parameters," *Biometrika,* 67, 515–523.

——— (1975), "Application of Maximum Likelihood Methods to Finite Samples," *Sankhya,* 37, 259–270.

——— (1973), "Normal Likelihoods and Their Relation to Large-Sample Theory of Estimation," *Biometrika,* 60, 457–465.

Tierney, L., and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association,* 81, 82–86.

Viveros, R. (1990), "An Approximate Normalizing Power Transformation for the *F* Distribution," Communications in Statistics, Part B—Simulation and Computation, 19, 57–69.