Subset Third Derivative Diagnostics of Nonnormality

by

Elizabeth H. Slate

1Research partially supported by National Science Foundation grant DMS 9505065.
Abstract

As more complicated parametric models are investigated, it becomes increasingly more difficult to anticipate regions where the likelihood function or posterior distribution will be poorly behaved and cause computational or inferential techniques to fail. For univariate parameters, Slate (1994), Albert (1989), Hills (1989) and others have proposed a measure based on the third derivative of the log-posterior density that can indicate when the posterior is poorly behaved in the sense that it is far from normal. This measure of nonnormality was generalized to a multidimensional posterior distribution by Kass (1989), who obtained two affine-invariant scalar summaries of the third derivative array for use as diagnostics. This paper extends these scalar summaries for the multivariate parameter to subsets of the parameter. The subset measures produce the summaries of Kass when the full parameter is of interest and approximate those summaries for the marginal distribution of a subset of the parameter. The subset diagnostics can help to identify problematic components of the parameter when the joint nonnormality measures are large, and also may indicate poor behavior in the posterior distribution that is not detected by the joint measures.

KEY WORDS: multivariate likelihood, marginal distribution, normal approximation, multivariate posterior, reparameterization
1. INTRODUCTION

Recent developments in computationally intensive techniques for fitting statistical models (such as Monte Carlo Markov chain methods) and rapid advances in computing speed are enabling statisticians to fit increasingly more complicated statistical models. This trend toward more complicated, high dimensional parametric models is further fueled by the extensive data sets that are becoming ever more routinely available. Understanding the behavior of statistical models, however, is clearly more difficult as the parameter dimension grows. It is hard to anticipate regions where the likelihood function or posterior distribution will be poorly behaved and cause computational or inferential techniques to fail. Indeed, even recognizing that these failures have occurred can be problematic.

This paper proposes diagnostics that can facilitate the exploration of multivariate likelihood functions and distributions. For both computational and inferential purposes, “well-behaved” distributions are typically those that are close to normal. Thus the diagnostics proposed in this paper are measures of nonnormality and are calibrated according to the reliability of normality-based inferences. More specifically, the diagnostics indicate when inference based on asymptotic normality may be unreliable, and, consequently, when posterior distributions (or likelihood functions) may be problematic.

Let $Y = (Y_1, \ldots, Y_n)$ be a random sample from a five-times continuously differentiable distribution depending on an unknown $m$-dimensional parameter $\theta = (\theta_1, \ldots, \theta_m)$. Suppose that a smooth prior distribution is specified for $\theta$ and the corresponding log-posterior distribution is denoted as $\ell(\theta)$. Let the posterior mode be $\hat{\theta}$ and denote the $k$-th derivative of the log posterior evaluated at the mode as $D^k_\theta \ell(\hat{\theta})$. The modal normal approximation to the posterior distribution is the normal
distribution with mean $\hat{\theta}$ and precision $-D_2^2 \ell(\hat{\theta})$. The third derivative that is the focus of this paper is $D_3^3 \ell(\hat{\theta})$.

Section 2 describes the third derivative diagnostic $F^2$ for a scalar parameter $\theta$ and its generalizations to the measures $m^2 B^2$ and $B^2$ for $\theta \in \mathbb{R}^m$. The larger the values of these diagnostics, the greater the anticipated deviation of the posterior distribution from its modal normal approximation. The measures $m^2 B^2$ and $B^2$ can be helpful in diagnosing deviations from normality of the $m$-dimensional posterior for $\theta$, but they can be misleading. It is possible, for example, that $m^2 B^2$ and $B^2$ are large and yet a normal approximation for subsets of the parameter is adequate. Similarly, there may be some subsets of the parameter for which normal approximations are unreasonable while the joint measures $m^2 B^2$ and $B^2$ are small. An example of the first type of behavior is introduced in Section 2, and additional examples in Section 4 illustrate both kinds of behavior.

These observations support the importance of supplementing the measures $m^2 B^2$ and $B^2$ with the subset diagnostics introduced in Section 3. The development is similar to that of Cook and Goldberg (1986) for subset curvature measures in non-linear regression. (The procedure of Cook and Goldberg also applies, with obvious modifications, to the more general curvature measures presented by Kass and Slate, 1994.) Examples in Section 4 illustrate the use of the subset diagnostics. Additional discussion is in Section 5.

2. THIRD DERIVATIVE DIAGNOSTICS

Extending the notation given in the Introduction, let $G = -D_2^2(\hat{\theta})$ be minus the Hessian of the log posterior evaluated at the mode, and let $W = D_3^3(\hat{\theta})$. The quantities $G$ and $W$ are called the observed Hessian and the observed third derivative
because they are evaluated at the mode, similar in meaning to "observed information" for the likelihood function. This section summarizes the third derivative diagnostics first for a scalar parameter $\theta$ and then for $\theta \in \mathbb{R}^m$. An example is introduced to illustrate the need for the subset third derivative diagnostics that are developed in Section 3.

2.1. Scalar $\theta$

For scalar $\theta$, $G = -\ell''(\hat{\theta})$, and $W = \ell'''(\hat{\theta})$. For arguments $\theta$ such that $(\theta - \hat{\theta}) = O(n^{-1/2})$, the log-posterior distribution for $\theta$ may be written as

$$
\ell(\theta) - \ell(\hat{\theta}) = -\frac{1}{2} G(\theta - \hat{\theta}) + \frac{1}{6} W(\theta - \hat{\theta})^3 + O(n^{-1}).
$$

The first term on the right side of (1) is the modal normal approximation to the posterior distribution. The next term involves the third derivative $W$, and so any nonzero value for $W$ indicates a departure of the posterior from the modal normal approximation.

The third derivative diagnostic is $F^2 = W^2 G^{-3}$. The standardization by $G^3$ makes $F^2$ invariant under affine transformations of $\theta$. This measure has also been proposed by Hills (1989, sec. 2.2.1) and Albert (1989) in the Bayesian context and by Sprott (1973) and Anscombe (1964) for indicating departures from normality in the likelihood (replace $\ell$ by the log likelihood and $\hat{\theta}$ by the maximum likelihood estimate).

Slate (1994, 1997a) used a tail probability approximation to show that $F^2 > 0.16$ may indicate problems in the tail behavior of the posterior distribution in the sense that the tail regions that have probability 0.05 under the modal normal approximation may have quite different coverage under the exact posterior distribution. By
considering arbitrary probability tail regions, this calibration of \( F^2 \) can be extended to a desired precision. The diagnostic \( F^2 \), however, and the tail probability rule used for calibration are restricted to univariate distributions.

2.2. Multivariate \( \theta \)

Suppose now that \( \theta \in \mathbb{R}^m \). Then the observed Hessian \( G \) is an \( m \times m \) matrix with entries \( g_{ab} = \partial^2 \ell(\hat{\theta})/\partial \theta_a \partial \theta_b \), and the observed third derivative of the log posterior is an \( m \times m \times m \) array, \( W = (w_{abc}) \) with \( w_{abc} = \partial^3 \ell(\hat{\theta})/\partial \theta_a \partial \theta_b \partial \theta_c \), \( a, b, c = 1, \ldots, m \).

There are two scalar summaries of \( W \) that are invariant under affine transformations of \( \theta \). These are (Kass, 1989; McCullagh, 1987, sec. 2.8)

\[
m^2 \bar{B}^2 = \sum_{a,b,c,d,e,f} g^{ab}g^{de}g^{cf}w_{abc}w_{def}
\]

(2)

\[
B^2 = \sum_{a,b,c,d,e,f} g^{ad}g^{be}g^{cf}w_{abc}w_{def},
\]

(3)

where \( G^{-1} = (g^{ab}) \). Both \( m^2 \bar{B}^2 \) and \( B^2 \) are generalizations of \( F^2 \): when \( m = 1 \), \( m^2 \bar{B}^2 = B^2 = F^2 \).

Slate (1997a) suggested that these summary third derivative measures are large when \( m^2 \bar{B}^2 > m(0.16) \) and \( B^2 > m^3(0.16) \). See Slate (1992) for an alternative calibration of \( m^2 \bar{B}^2 \) based on the tail probabilities of the chi-square distribution.

2.3. Curvature Viewpoint

Kass and Slate (1994) noted that \( m^2 \bar{B}^2 \) and \( B^2 \) are formed from the three-way array of observed third derivatives of the log-posterior density in precisely the way that the curvature summaries \( \bar{\omega}^2 \) and \( \omega^2 \) (also \( \bar{\gamma}^2 \) and \( \gamma^2 \)) are formed from the three-way array of observed second derivatives of the expectation function in nonlinear regression.
(For definition and discussion of these curvature summaries, see Bates and Watts, 1980, 1988; Kass, 1989; Kass and Slate, 1994; and Efron, 1975.) Specifically, a linear combination of \( m^2 \bar{B}^2 \) and \( B^2 \) may be obtained as a root-mean-square of a "curvature" of the directed first derivative of the log-posterior density. Let \( u(t) = D_\theta \ell(\hat{\theta} + tv) \) for \( t \in \mathbb{R} \) and \( v \in \mathbb{R}^m \). Form the "curvature" in the direction \( v \) by

\[
\kappa(v) = \frac{\|u''(0)\|_{G^{-1}}}{\|u'(0)\|_{G^{-1}}^2}.
\]  

(4)

An average third derivative curvature may be defined as the root-mean-square average of \( \kappa(v) \):

\[
(B^2)_{RMS} = \frac{1}{S} \int_R \kappa^2(v) \, dS,
\]  

(5)

where \( R = \{ v : \|u'(0)\|_{G^{-1}} = 1 \} = \{ v : \|v\|_{G} = 1 \} \) and \( S \) is the surface area of the \( G \)-unit sphere. By the same methods used by Bates and Watts (1980, following equation (2.30)) it follows that

\[
(B^2)_{RMS} = \frac{m^2 \bar{B}^2 + 2B^2}{m(m + 2)}.
\]

2.4. A Motivating Example

To motivate the need for subset third derivative diagnostics, consider the growth data for orange tree three given in Draper and Smith (1981, p. 524, exercise N). The trunk circumference was measured at seven times within a five-year period. One model proposed for these data is the Gompertz nonlinear regression model, for which the circumference measurements, \( Y_i \), are independent normal random variables with variance \( \sigma^2 \) and means \( \mu_i = \theta_1 \exp\left(-e^{\theta_2 x_i} - \theta_3 x_i\right) \), where \( x_i \) is the time of the \( i \)th measurement. Fixing \( \sigma \) at its maximum likelihood estimate (MLE) and
using a flat prior for $\theta = (\theta_1, \theta_2, \theta_3)$, the values of the third derivative summaries are $m^2 \bar{B}^2 = 10.83$, $B^2 = 33.54$, and $(B^2)_{RMS} = 5.19$. Large values for these diagnostics, based on the guidelines reported in Section 2.2, are 0.48, 4.32 and 0.61, respectively. Thus it is apparent that the modal normal approximation to the joint posterior distribution for $\theta$ is unreliable. Indeed, as an illustration, the posterior probability of the region about $\hat{\theta}$ that has 90% coverage under the modal normal distribution is approximately 0.67 ($se = 0.001$) (computed using importance sampling based on 100000 samples from a $t_2$ distribution). Yet, as shown in Figure 4, the marginal posterior distribution for $\theta_3$ is close to normal. The approximate coverage of the

![Figure 1: The marginal posterior density for $\theta_3$ in the orange tree example (solid line) and its modal normal approximation (dotted line).](image)

nominal central 90% region for $\theta_3$ is 0.92 (0.001), also computed by importance sampling. Section 4.1 returns to this example.

3. SUBSET THIRD DERIVATIVE DIAGNOSTICS

To supplement the diagnostic $F^2$ for scalar $\theta$ and the summaries $m^2 \bar{B}^2$ and $B^2$ for $\theta \in \mathbb{R}^m$, subset third derivative measures are developed in a manner analogous to
that for subset curvature measures in the context of nonlinear regression by Cook and Goldberg (1986).

Partition \( \theta \) as \( \theta = (\theta_r, \theta_s) \) where \( \theta_r \in \mathbb{R}^{m_r} \) and \( \theta_s \in \mathbb{R}^{m_s} \) with \( m_r + m_s = m \). Here, \( \theta_s \) represents the components of \( \theta \) of interest, for which the subset third derivative diagnostics will be developed. Partition the joint mode in the same way, as \( \tilde{\theta} = (\tilde{\theta}_r, \tilde{\theta}_s) \), and write

\[
G = \begin{pmatrix} G_{rr} & G_{rs} \\ G_{sr} & G_{ss} \end{pmatrix} \quad \text{and} \quad W = \begin{pmatrix} W_{rr} & W_{rs} \\ W_{sr} & W_{ss} \end{pmatrix},
\]

where \( G_{ij} \) is \( m_i \times m_j \) and \( W_{ijk} \) is \( m_i \times m_j \times m_k \), \( i, j, k \in \{ r, s \} \).

The subset third derivative diagnostics are computed using the log-profile posterior \( h(\theta_s) = \ell(\tilde{\theta}(\theta_s)) \), where \( \tilde{\theta}(\theta_s) \) maximizes \( \ell \) for the given value of \( \theta_s \). Assuming that \( \tilde{\theta}(\cdot) \) has continuous second derivatives, \( h \) may be approximated using a third-order Taylor series expansion about \( \tilde{\theta}_s \), giving

\[
h(\theta_s) = \ell(\tilde{\theta}) + \frac{1}{2} \phi_s^T D_{\tilde{\theta}_s}^2 h(\tilde{\theta}_s) \phi_s + \frac{1}{6} \phi_s^T D_{\tilde{\theta}_s}^3 h(\tilde{\theta}_s) \phi_s + O(n^{-1}),
\]

for \( \theta_s \) such that \( \phi_s = (\theta_s - \tilde{\theta}_s) \) is \( O(n^{-1/2}) \). The brackets in the third derivative term denote the column multiplication of Bates and Watts (1980), which is explained briefly in the Appendix. The subset Hessian and subset third derivative are the Hessian and third derivative of \( h \) evaluated at \( \tilde{\theta}_s \). Define the subset Hessian as \( G_s = -D_{\tilde{\theta}_s}^2 h(\tilde{\theta}_s) \) and the subset third derivative as \( W_s = D_{\tilde{\theta}_s}^3 h(\tilde{\theta}_s) \). Let \( \Delta_1 = D_{\theta_s}[\tilde{\theta}(\theta_s)] \) evaluated at \( \tilde{\theta}_s \). Then (see Appendix) the subset Hessian and third derivative may
be expressed as

$$G_s = \Delta_1^T G \Delta_1$$  \hspace{1cm} (6)

and $$W_s = [\Delta_1^T][\Delta_1^T W \Delta_1].$$  \hspace{1cm} (7)

It is straightforward to show that $$G_s = G_{ss} - G_{sr} G_{rr}^{-1} G_{rs},$$ the inverse of the part of $$G^{-1}$$ corresponding to the subset of interest. Expanding the expression for $$W_s$$ gives

$$W_s = - [(G_{rr}^{-1} G_{rs})^T] [(G_{rr}^{-1} G_{rs})^T W_{rrr} (G_{rr}^{-1} G_{rs})]$$

$$+ 3 (G_{rr}^{-1} G_{rs})^T W_{srr} (G_{rr}^{-1} G_{rs})$$

$$- 3 (G_{rr}^{-1} G_{rs})^T W_{srs}$$

$$+ W_{sss}.$$  

Of course, when $$\theta_s = \theta, G_s = G$$ and $$W_s = W.$$ 

The subset third derivative diagnostics are analogous to the full-parameter summaries given in (2) and (3), but use the subset Hessian $$G_s$$ and subset third derivative $$W_s$$:

$$m_s^2 B_s^2 = \sum_{a,b,c,d,e,f} \frac{m_s}{g_s g_{de} g_{cf} (w_s)_{abc} (w_s)_{def}}$$  \hspace{1cm} (8)

$$B_s^2 = \sum_{a,b,c,d,e,f} \frac{m_s}{g_s g_{de} g_{cf} (w_s)_{abc} (w_s)_{def}},$$  \hspace{1cm} (9)

where now $$G_s^{-1} = (g_s^{ab}).$$ When $$\theta_s = \theta,$$ these subset summaries yield the summaries for the full third derivative array given in (2) and (3). These measures are also considered by DiCiccio and Stern (1993) in the context of Bartlett corrections in Bayesian inference.
Like \( m^2 \bar{B}^2 \) and \( B^2 \), the subset third derivative summaries are "large" when \( m^2 \bar{B}^2_s > m_s(0.16) \) and \( B^2_s > m^3_s(0.16) \).

3.1. Affine Invariance

A desirable property of \( m^2 \bar{B}^2 \) and \( B^2 \) is their affine invariance. The subset third derivative summaries are also invariant under appropriate affine transformations. Specifically, the subset measures \( m^2 \bar{B}^2_s \) and \( B^2_s \) are invariant under transformations of \( \theta \) to \( \psi \) of the form \( \psi = E\theta + v \) with

\[
\begin{pmatrix}
\psi_r \\
\psi_s
\end{pmatrix} = \begin{pmatrix}
E_{rr} & E_{rs} \\
0 & E_{ss}
\end{pmatrix}
\begin{pmatrix}
\theta_r \\
\theta_s
\end{pmatrix} + \begin{pmatrix}
v_r \\
v_s
\end{pmatrix}
\]  

(10)

for \( \psi = (\psi_r, \psi_s)^T \), \( E \in \mathbb{R}^{m \times m} \) partitioned as \( \theta \). (Thus \( E_{rr} \) is \( m_r \times m_r \), \( E_{ss} \) is \( m_s \times m_s \), and \( E_{rs} \) is \( m_r \times m_s \).) Hence the subset third derivative summaries are invariant when \( \psi_s \) is an affine transformation of the subset of interest, \( \theta_s \).

To establish this invariance property, recall first that \( m^2 \bar{B}^2 \) and \( B^2 \) are affine invariant, and note that under the transformation \( \psi = A\theta + v \) with \( A \) arbitrary (but full rank), the observed Hessian and third derivative array for \( \psi \) are

\[
G^\psi = A^{-T} \theta A^{-1} \quad \text{and} \quad W^\psi = [A^{-T}] [A^{-T} \theta A^{-1}],
\]

(11)

where \( G^\theta \) and \( W^\theta \) are the observed Hessian and third derivative array in the \( \theta \) parameterization. Consequently, all that is needed to demonstrate the invariance of \( m^2 \bar{B}^2_s \) and \( B^2_s \) under transformations of the form (10) is to show that the subset Hessian and third derivative array transform in the same manner: that there is an
\[ G_s^\psi = A^{-T} G_s^\theta A^{-1} \quad \text{and} \quad W_s^\psi = \begin{bmatrix} A^{-T} \\ A^{-T} W_s^\theta A^{-1} \end{bmatrix}. \] (12)

That such a matrix \( A \) exists follows from noting that \( \Delta_1^\psi = E \Delta_1^\theta E^{ss}, \) where \( E^{ss} = E_{ss}^{-1} \) is the lower right \( m_s \times m_s \) portion of \( E^{-1}. \) Using this relationship between \( \Delta_1^\psi \) and \( \Delta_1^\theta, \) it is easy to verify that equations (12) hold for \( A = E_{ss}. \)

The affine invariance of \( m_s^2 \tilde{B}_s^2 \) and \( B_s^2 \) can be exploited to simplify their computation. One possibility is to choose the parameterization so that the Hessian is block diagonal, \( i.e. \) so that \( G_{rs} \) is zero. Then, in this parameterization, \( G_s = G_{ss} \) and \( W_s = W_{sss}, \) the portions of the Hessian and third derivative array corresponding to the components of the parameter of interest. A transformation of \( \theta \) that makes the Hessian block diagonal is (10) with \( v = 0 \) and

\[ E = \begin{pmatrix} G_{rr} & G_{rs} \\ 0 & G_{ss} \end{pmatrix}, \]

where the entries in \( E \) are from the observed Hessian in the original parameter \( \theta. \) Denoting this parameterization by \( \psi, \) \( G_s^\psi \) and \( W_s^\psi \) are obtained by extracting the parts of \( G^\psi \) and \( W^\psi \) corresponding to the components of interest. Thus, given the observed Hessian \( G^\theta \) and third derivative array \( W^\theta \) for the original parameterization \( \theta, \) the subset scalar summaries may be computed by forming \( G^\psi \) and \( W^\psi \) by using (11) with \( A = E, \) extracting the portions corresponding to the components of interest, \( G_s^\psi = G_{ss}^\psi, W_s^\psi = W_{sss}^\psi, \) and then using these in (8) and (9).

3.2. Marginal Distribution

The observed subset third derivative summaries \( m_s^2 \tilde{B}_s^2 \) and \( B_s^2 \) are \( O(n^{-2}) \) approx-
imations to the measures $m^2 B^2$ and $B^2$ for the marginal posterior distribution of $\theta_s$. This relationship follows from consideration of the Laplace approximation to the marginal distribution of $\theta_s$ (Tierney, Kass and Kadane, 1989).

Let $f(\theta_s)$ be the marginal posterior density for $\theta_s$, $\hat{\theta}_s$ be the marginal posterior mode, and $l(\theta_s) = \log f(\theta_s)$ be the log marginal density. Then, with $G^{\theta_s} = -D_{\theta_s}^2 l(\hat{\theta}_s)$ and $W^{\theta_s} = D_{\theta_s}^3 l(\hat{\theta}_s)$, the marginal scalar third derivative summaries for $\theta_s$ are given by (2) and (3) as

\begin{align*}
(m_s^2 B^2)_{\theta_s} &= \sum_{a,b,c,d,e,f} (G^{\theta_s})^{ad} (G^{\theta_s})^{bc} (G^{\theta_s})^{df} (W^{\theta_s})_{abc} (W^{\theta_s})_{def} \\
(B^2)_{\theta_s} &= \sum_{a,b,c,d,e,f} (G^{\theta_s})^{ad} (G^{\theta_s})^{be} (G^{\theta_s})^{cf} (W^{\theta_s})_{abc} (W^{\theta_s})_{def}
\end{align*}

where $(G^{\theta_s})^{-1} = ((G^{\theta_s})^{ab})$ and $W^{\theta_s} = ((W^{\theta_s})_{abc})$, $a,b,c = 1, \ldots, m_s$.

When $f(\theta_s)$ is not known precisely, Tierney et al. (1989) provide the approximation

\[ \hat{f}(\theta_s) = c \left\{ \frac{\det \Sigma(\theta_s)}{\det \Sigma} \right\}^{-1/2} \exp \left\{ h(\theta_s) - h(\hat{\theta}_s) \right\}, \]

where $h(\theta_s) = \ell(\theta_s)$ as before, $\Sigma = G^{-1}$, $\Sigma(\theta_s)$ is minus the inverse of the $m_s \times m_s$ Hessian of $h(\theta_s)$, and $c$ is a normalizing constant. This approximation has relative error of order $O(n^{-3/2})$ for $\theta_s$ such that $\theta_s - \hat{\theta}_s$ is order $O(n^{-1/2})$. Thus,

\[ l(\theta_s) = h(\theta_s) - h(\hat{\theta}_s) - \frac{1}{2} \log \left\{ \frac{\det \Sigma(\theta_s)}{\det \Sigma} \right\} + \log c + O(n^{-3/2}). \]

There are two differences between the subset scalar summaries $m_s^2 \bar{B}_s^2$ and $B_s^2$ and the marginal scalar third derivative summaries $(m_s^2 \bar{B}_s^2)_{\theta_s}$ and $(B_2^2)_{\theta_s}$. For the subset summaries, $l(\theta_s)$ has been approximated by $h(\theta_s) = \ell(\theta_s)$ and $\hat{\theta}_s$ has been approx-
imated by \( \hat{\theta}_s \). Note that \( l(\theta_s) = h(\theta_s) + O(1) \), so that \( D_{\theta_s} l(\hat{\theta}_s) = (\hat{\theta}_s - \hat{\theta}_s) D_{\theta_s}^2 l(\hat{\theta}_s) + O(1) \) and \( \hat{\theta}_s - \hat{\theta}_s = O(n^{-1}) \). From this it follows that \( D_{\theta_s}^k l(\hat{\theta}_s) = D_{\theta_s}^k h(\hat{\theta}_s) + O(1) \), for \( k = 2, 3 \). Thus, \((G^{\theta_s})_{ab} = (G_s)_{ab} + O(n^{-1}) \), \((W^{\theta_s})_{abc} = (W_s)_{abc} + O(n^{-1}) \) and \((G^{\theta_s})_{ab} = g_{ab}^{\theta_s} + O(n^{-2}) \). Consequently the marginal measures \((m^2 \tilde{B}^2)_{\theta_s} \) and \((B^2)_{\theta_s} \) differ from \(m^2 \tilde{B}^2_s \) and \(B^2_s \) by \(O(n^{-2}) \). The relationship between the exact marginal and subset third derivative summaries was also noted by DiCiccio and Stern (1993) in the context of obtaining Bartlett corrections for Bayesian inference.

3.3. Curvature Viewpoint

Section 2.3 described how a root-mean-square third derivative summary may be obtained as a linear combination of the summaries \(m^2 \tilde{B}^2 \) and \(B^2 \). By replacing the function \( u \) and Hessian \( G \) in that section with the directed first derivative of \( h \) and subset Hessian \( G_s \), a root-mean-square subset third derivative summary results. Specifically, let \( u_s(t) = D_{\theta_s} h(\hat{\theta}_s + tv) \) for \( t \in \mathbb{R} \) and \( v \in \mathbb{R}^{m_s} \) and form the “curvature” in the direction \( v \) by (4) with \( G \) replaced by \( G_s \). Then the root-mean-square subset third derivative summary is \((B^2_s)_{RMS} \), given by (5) with the new definition of \( \kappa \) and the region of integration being determined by \( G_s \). The result is that \((B^2_s)_{RMS} = (m^2 \tilde{B}^2_s + 2B^2_s)/(m_s(m_s + 2)) \).

4. EXAMPLES

This section returns to the orange tree example of Section 2.4 and then presents two additional examples that illustrate the subset third derivative diagnostics.

4.1. Orange Tree

Returning to the orange tree example of Section 2.4, recall that a flat prior distribution is used for the parameter \( \theta = (\theta_1, \theta_2, \theta_3) \), so that the diagnostics are
evaluating nonnormality of the likelihood function. Because the likelihood factors, consider the error standard deviation fixed at its MLE, $\hat{\sigma} = 5.09$. The joint posterior mode and approximate standard errors are $\hat{\theta}_1 = 188.4$ (25.87), $\hat{\theta}_2 = 0.83$ (0.076), and $\hat{\theta}_3 = 1.35$ (0.29). Approximate posterior correlations under the flat prior are $\text{corr}(\theta_1, \theta_2) = -0.21$, $\text{corr}(\theta_1, \theta_3) = -0.96$, and $\text{corr}(\theta_2, \theta_3) = 0.44$.

Figure 2 shows the conditional log-likelihood contours for these data, obtained by fixing each component of $\theta$ at its MLE in turn. Note that the $(\theta_1, \theta_3)$ contours are “banana-shaped,” a property often associated with numerical problems with the likelihood function. Table 1 gives the subset third derivative summaries for these data. For each subset size (1, 2 or 3), “large” values of the diagnostics are also given based on the guidelines given in Section 2.2. From the diagnostics for the full parameter $(\theta_1, \theta_2, \theta_3)$, it is clear that this likelihood function is poorly behaved. The subset measures further indicate that it is indeed $(\theta_1, \theta_3)$, associated with the banana contours, that is most poorly behaved. The univariate subset diagnostics show that $\theta_1$ is most problematic. Additionally, these diagnostics reveal that inferences based on the modal normal approximation to the marginal distribution of $\theta_3$ raise no concern, supporting the observation of Section 2.4. For this model and data, it may be possible to reparameterize just $(\theta_1, \theta_3)$ or perhaps even just $\theta_1$ and greatly improve the behavior of the joint likelihood function.

4.2. Guinea Pig

Johansen (1984) discusses data obtained from eight guinea pigs. Five tissue samples from each pig were assigned to each of ten different concentrations of B-methyl-glucoside. The response is the average uptake volume of the five tissue samples in micromoles per milligram of fresh tissue per two minutes. These data have also
Figure 2: Conditional log-likelihood contours for orange tree 3. The contours are at 1, 2 and 3 units from the maximum log likelihood value.

been considered by Lindstrom and Bates (1990) and by Bennett, Rancine-Poon and Wakefield (1996, p. 345). For illustrative purposes here, consider the data for pig six only. Let $x_i$ be the concentration of B-methyl-glucoside and $y_i$ be the average uptake volume of the 5 tissue samples, $i = 1, \ldots, 10$. The model is a nonlinear regression model with

$$\log(Y_i) \sim N(\mu_i, \sigma^2),$$

independently
Table 1: The subset third derivative measures for orange tree 3.

<table>
<thead>
<tr>
<th>Subset</th>
<th>(m^2\bar{B}_x^2)</th>
<th>(B_x^2)</th>
<th>((B_x^2)_{RMS})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\theta_1)</td>
<td>2.47</td>
<td>2.47</td>
<td>2.47</td>
</tr>
<tr>
<td>(\theta_2)</td>
<td>0.13</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>(\theta_3)</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>Large</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td>((\theta_1, \theta_2))</td>
<td>1.82</td>
<td>3.54</td>
<td>1.11</td>
</tr>
<tr>
<td>((\theta_1, \theta_3))</td>
<td>4.88</td>
<td>13.45</td>
<td>3.97</td>
</tr>
<tr>
<td>((\theta_2, \theta_3))</td>
<td>0.28</td>
<td>0.93</td>
<td>0.27</td>
</tr>
<tr>
<td>Large</td>
<td>0.32</td>
<td>1.28</td>
<td>0.36</td>
</tr>
<tr>
<td>((\theta_1, \theta_2, \theta_3))</td>
<td>10.83</td>
<td>33.54</td>
<td>5.19</td>
</tr>
<tr>
<td>Large</td>
<td>0.48</td>
<td>4.32</td>
<td>0.61</td>
</tr>
</tbody>
</table>

\[
\mu_i = \log \left( \frac{\phi_1 x_i}{\phi_2 + x_i} + \phi_3 x_i \right).
\]

Take a flat prior (so that \(\ell\) is the log-likelihood function) and again consider \(\sigma\) fixed at its MLE, 0.097. Thus the dimension of the parameter is \(m = 3\). The posterior modes (MLEs) and approximate standard errors are \(\tilde{\phi}_1 = 0.23 (0.03), \tilde{\phi}_2 = 2.98 (0.57), \tilde{\phi}_3 = 0.0016 (0.0009)\). First order posterior correlations are \(\text{corr}(\phi_1, \phi_2) = 0.92, \text{corr}(\phi_1, \phi_3) = -0.88\) and \(\text{corr}(\phi_2, \phi_3) = -0.76\).

Figure 3 shows the conditional log-likelihood contours for these data. The \((\phi_1, \phi_3)\) and \((\phi_1, \phi_2)\) contours reflect the strong correlation between these parameters, but there are not the banana-shaped contours as for the orange tree example. Table 2 gives the subset third derivative summaries for these data. Again, for each subset size (1, 2 or 3), “large” values of the diagnostics are also given. With the exception of a small exceedance of the guideline by \(m^2\bar{B}^2\), there is no indication from the joint \((m = 3)\) diagnostics of poor behavior in the likelihood. Consideration
Figure 3: Conditional log-likelihood contours for guinea pig 6. The contours are at 1, 2 and 3 units from the maximum log likelihood value.

of the subset measures, however, reveals that \((\phi_1, \phi_2)\) may be poorly behaved and, furthermore, that much of this poor behavior may be attributable to \(\phi_2\). Figure 4 shows the marginal posterior distribution for \(\phi_2\) and its modal normal approximation. The tails of these distributions differ substantially. Indeed the lower and upper 5% regions under the normal approximation have posterior probabilities of 0.015 (0.0004) and 0.12 (0.001).
Table 2: The subset third derivative measures for guinea pig 6. “Large” values for each subset size are also given.

<table>
<thead>
<tr>
<th>Subset</th>
<th>$m_s^2 B_s^2$</th>
<th>$B_s^2$</th>
<th>$(B_s^2)_{RMS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi_1$</td>
<td>0.18</td>
<td>0.18</td>
<td>0.18</td>
</tr>
<tr>
<td>$\phi_2$</td>
<td>0.26</td>
<td>0.26</td>
<td>0.26</td>
</tr>
<tr>
<td>$\phi_3$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Large</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td>$(\phi_1, \phi_2)$</td>
<td>0.49</td>
<td>0.39</td>
<td>0.16</td>
</tr>
<tr>
<td>$(\phi_1, \phi_3)$</td>
<td>0.27</td>
<td>0.31</td>
<td>0.11</td>
</tr>
<tr>
<td>$(\phi_2, \phi_3)$</td>
<td>0.33</td>
<td>0.32</td>
<td>0.12</td>
</tr>
<tr>
<td>Large</td>
<td>0.32</td>
<td>1.28</td>
<td>0.36</td>
</tr>
<tr>
<td>$(\phi_1, \phi_2, \phi_3)$</td>
<td>0.56</td>
<td>0.50</td>
<td>0.10</td>
</tr>
<tr>
<td>Large</td>
<td>0.48</td>
<td>4.32</td>
<td>0.61</td>
</tr>
</tbody>
</table>

4.3. Survival Model

Consider the two-parameter model used by Feigl and Zelen (1965) to represent the relationship between survival time for leukemia patients and white blood cell count (WBC). Let $Y_i$ and WBC$_i$ be the survival time in weeks and the white blood cell count for the $i$-th patient. The model is

$$ Y_i \sim \text{Exponential with mean } \mu_i, \text{ independently} $$

$$ \mu_i = \theta_1 \exp(-\theta_2 x_i), $$

where $\theta = (\theta_1, \theta_2)$, and $x_i = x'_i - \bar{x}'$ with $x'_i = \log(\text{WBC}_i)$ is the centered logarithm of the white blood cell count for patient $i$, $i = 1, \ldots, n$. Here the data for the $n = 17$ AG-positive patients are used. In addition to $\theta$, the parameterizations $\lambda$ and $\beta$ may be studied, where $\mu_i = (\beta_1 x'_i)^{\beta_2}$, with $x'_i = \exp(-x_i)$, and $\mu_i = \exp(\lambda_1 - \lambda_2 x_i)$. Kass and Slate (1994) examine likelihood-based curvature measures.
for these parameterizations and conclude that $\beta$ is very poorly behaved (has more severely banana-shaped contours than orange tree three), $\theta$ is considerably better and $\lambda$ is the best of the three.

The subset third derivative summaries of the posterior distribution are shown in Table 3, based on a uniform prior on $\lambda$. For the univariate subsets, $m_\star^2 \tilde{B}_\star^2 = B_\star^2$, and these measures approximate $F^2$ for the corresponding univariate marginal distributions. By referring the diagnostic values to the guideline of 0.16, one notes that the second component of the parameter is well-behaved in all three parameterizations, but the measures for $\theta_1$ and, much more so, $\beta_1$, indicate nonnormality. Thus the large values for $m^2 \tilde{B}^2$ and $B^2$ for the joint posterior of $\beta$ are due primarily to the first component $\beta_1$. This example illustrates how the third derivative diagnostics can aid in the choice of a well-behaved parameterization.

5. DISCUSSION

The subset third derivative measures described here extend the diagnostics $F^2$ for scalar parameters and $m^2 \tilde{B}^2$ and $B^2$ for multivariate parameters to provide addi-
Table 3: Subset scalar third derivative summaries for the Feigl and Zelen example.

<table>
<thead>
<tr>
<th>Subset</th>
<th>( m^2 B^2 )</th>
<th>( B^2 )</th>
<th>( (B^2)_{RMS} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta_1 )</td>
<td>0.89</td>
<td>0.89</td>
<td>0.89</td>
</tr>
<tr>
<td>( \theta_2 )</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>( \lambda_1 )</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>( \lambda_2 )</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>8.67</td>
<td>8.67</td>
<td>8.67</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>Large</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
</tr>
</tbody>
</table>

(\( \theta_1, \theta_2 \)) | 1.39 | 1.06 | 0.44 |
(\( \lambda_1, \lambda_2 \)) | 0.24 | 0.24 | 0.09 |
(\( \beta_1, \beta_2 \)) | 20.55 | 46.99 | 14.32 |
Large | 0.32 | 1.28 | 0.61 |

Informational information about the behavior of the posterior distribution for subsets of the parameter. As illustrated in the orange tree example, the subset diagnostics can help to identify problematic components of the parameter when the joint diagnostics \( m^2 B^2 \) and \( B^2 \) are large. Also the subset measures may indicate poor behavior in the posterior distribution that is not detected by the joint diagnostics, as seen in the model for guinea pig six, for example.

The subset diagnostics approximate the third derivative diagnostics that could be computed from the marginal posterior distribution of the subset of interest. Thus, by considering all possible subsets of the parameter, the subset third derivative diagnostics can be used to evaluate the degree of nonnormality of all component marginal distributions of the parameter.

For the examples shown, exact third derivatives have been computed. However the subset scalar summaries can be computed efficiently using numerical derivatives in a manner similar to that described by DiCiccio and Stern (1993). Such calculation
avoids specifying derivatives of higher than second order and avoids performing summations beyond third order.

If the behavior of the full $m$-dimensional posterior distribution of is primary interest, then conditional distributions (rather than marginal) may be more appealing. This is because of the central-limit-type effect of the averaging that is required to compute marginal distributions. Slate (1997a) considers diagnostics based on conditional distributions and links them to dynamic displays that can facilitate exploration of multivariate likelihood and posterior functions.
APPENDIX: THE SUBSET HESSIAN AND THIRD DERIVATIVE

ARRAY

A1. The derivatives of $\tilde{\theta}(\theta_s)$

Write $\tilde{\theta}(\theta_s) = (\rho(\theta_s), \theta_s)^T$ and let $H = -G = D^2_\theta \ell(\tilde{\theta})$. Because $\rho$ maximizes $\ell\{[\rho(\theta_s), \theta_s]\}$ for each fixed $\theta_s$,

$$
\frac{\partial \ell\{[\rho(\theta_s), \theta_s]\}}{\partial \rho_a} \bigg|_{\rho=\rho(\theta_s)} = 0 \tag{A.1}
$$

for $a = 1, 2, \ldots, m_r$ and all $\theta_s$. Differentiating (A.1) with respect to $\theta_s$ and evaluating at $\tilde{\theta}_s$ gives $(H_{rr} H_{rs}) \Delta_1 = 0$, from which it follows that

$$
\Delta_1 = \begin{pmatrix} -H_{rr}^{-1} H_{rs} \\ I_{m_s} \end{pmatrix}. \tag{A.2}
$$

The Hessian of $\tilde{\theta}(\theta_s)$ evaluated at $\tilde{\theta}_s$, denoted as $\Delta_2 = D^2_{\theta_s} \tilde{\theta}(\theta_s)$, may be obtained similarly. Differentiation of (A.1) twice with respect to $\theta_s$ (the $u$ and $v$-th components, say) and evaluating at $\tilde{\theta}_s$ yields

$$
\sum_{b=1}^{m_r} (H_{rr})_{ab} \Delta_2_{bu} + \sum_{b,c=1}^{m} (\Delta_1)_{bu} W_{abc} \Delta_1_{cv} = 0,
$$

$u, \ v = 1, \ldots, m_s$. From this,

$$
\Delta_2 = - \begin{bmatrix} H_{rr}^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta_1^T \\ W \Delta_1 \end{bmatrix}.
$$
Here the brackets indicate the column multiplication as defined by Bates and Watts (1980). Briefly, if \( A \) is an \( a \times b \) array and \( B \) is a \( b \times c \times d \) array, then \( C = [A][B] \) is an \( a \times c \times d \) array obtained by premultiplying each of the \( cd \) \( b \)-columns of \( B \) by \( A \).

**A2. The Hessian and third derivative**

The Hessian of \( h \) with respect to \( \theta_s \) given in (6) is immediate. Simplification using (A.2) produces \( G_s = G_{ss} - G_{sr}G_{rr}^{-1}G_{rs} \), as reported in Section 3.

The third derivative array given in (7) requires some justification. The third derivative term is

\[
D_{\theta_s}^3 h(\theta_s) \bigg|_{\theta_s} = -3 \left[ \Delta_2 \right] \left[ H \Delta_1 \right] + \left[ \Delta_1^T \right] \left[ \Delta_1^T W \Delta_1 \right]. \tag{A.3}
\]

Using the forms for \( \Delta_1 \) and \( \Delta_2 \) given above,

\[
\left[ \Delta_2 \right] \left[ H \Delta_1 \right] = -\begin{bmatrix} H_{rr}^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta_1^T W \Delta_1 \\ 0 \end{bmatrix} = -\begin{bmatrix} H_{rr}^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta_1^T W \Delta_1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ -H_{sr}H_{rr}^{-1}H_{rs} + H_{ss} \end{bmatrix}
\]

Thus the first term in (A.3) is zero and (7) follows.
REFERENCES


Captions for Figures

**Figure 1** The marginal posterior density for $\theta_3$ in the orange tree example (solid line) and its modal normal approximation (dotted line).

**Figure 2** Conditional log-likelihood contours for orange tree 3. The contours are at 1, 2 and 3 units from the maximum log likelihood value.

**Figure 3** Conditional log-likelihood contours for guinea pig 6. The contours are at 1, 2 and 3 units from the maximum log likelihood value.

**Figure 4** The marginal posterior density for $\phi_2$ in the guinea pig example (solid line) and its modal normal approximation (dotted line).