Bayesian Marginal Inference

Tom Leonard; John S. J. Hsu; Kam-Wah Tsui

Bayesian Marginal Inference

TOM LEONARD, JOHN S. J. HSU, and KAM-WAH TSUI*

A method is proposed for approximating the marginal posterior density of a continuous function of several unknown parameters, thus permitting inferences about any parameter of interest for nonlinear models when the sample size is finite. Possibly tedious numerical integrations are replaced by conditional maximizations, which are shown to be quite accurate in a number of special cases. There are similarities with the profile likelihood ideas originated by Kalbfleisch and Sprott (1970), and the method is contrasted with a Laplacian approximation recommended by Kass, Tierney, and Kadane (1988, in press), referred to here as the "KTK procedure." The methods are used to approximate the marginal posterior densities of the log-linear interaction effects and an overall measure of association in a two-way contingency table. Snee's (1974) hair/eye color data are reanalyzed, and adjustments are proposed to Goodman's (1964) analysis for the full-rank interaction model. Another application concerns marginalization problems for a discrete p-parameter exponential family distribution, and inferences are considered for the probability of a zero count.

KEY WORDS: Contingency table; Discrete exponential family; Interaction effect; Marginal posterior density; Measure of association; Laplace's method; Profile likelihood; Saddle-point accuracy; Zero count.

1. DESCRIPTION OF PROBLEM

Bayesian methodology provides a general paradigm for inference regarding any function of the unknown parameters in nonlinear models, even when the sample size is finite, and whether or not prior information is available. The complex computations required, however, for exact marginal posterior densities appear to reduce the availability of this approach to applied statisticians. In this article we show that marginal posterior densities can be approximated in many situations via easily accessible conditional maximization procedures. Our aim is to emphasize Bayesian marginal inference as one of the few convincing applied approaches to the analysis of nonlinear models, and thereby to broaden the type of nonlinear model that may conveniently be analyzed.

Consider an $n \times 1$ vector $\mathbf{y}$ of observations with joint density or probability mass function

$$p(\mathbf{y} \mid \boldsymbol{\theta}), \quad \mathbf{y} \in R^n, \boldsymbol{\theta} \in R^p, \quad (1.1)$$

given an unknown $p \times 1$ vector of parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^T$ that possesses a positive prior density $\pi(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in R^p$. Assume that it is required to approximate the marginal posterior density $\pi(\eta \mid \mathbf{y})$ of a parameter of interest

$$\eta = g(\boldsymbol{\theta}), \quad \eta \in \Omega \subseteq R^1, \quad (1.2)$$

where g is a continuous real-valued function on R^p. In many situations it will be virtually impossible to perform the tedious $(p - 1)$-dimensional numerical integrations for computing the exact density

$$\pi(\eta \mid \mathbf{y}) = \lim_{\gamma \to 0} \gamma^{-1} \int_D \pi(\boldsymbol{\theta} \mid \mathbf{y}) \, d\boldsymbol{\theta}, \quad \eta \in \Omega, \quad (1.3)$$

where D denotes the region

$$D = D(\eta, \gamma) = \{\boldsymbol{\theta} : \eta \le g(\boldsymbol{\theta}) \le \eta + \gamma\} \quad (1.4)$$

and

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta})p(\mathbf{y} \mid \boldsymbol{\theta}), \quad \boldsymbol{\theta} \in R^p, \quad (1.5)$$

denotes the posterior density of $\boldsymbol{\theta}$.

In some situations, such as when $\theta_1, \dots, \theta_p$ are a posteriori independent, it may be possible to compute the exact density of η by computer simulations. A good approximation may, however, still be more useful in terms of shorter computer time. In this spirit Kass, Tierney, and Kadane (1988, in press) and Tierney, Kass, and Kadane (1988) proposed the elegant approximation

$$\tilde{\pi}(\eta \mid \mathbf{y}) \propto \pi_M(\eta \mid \mathbf{y}) / [|\mathbf{R}_\eta|^{1/2} (\mathbf{b}_\eta^T \mathbf{R}_\eta^{-1} \mathbf{b}_\eta)^{1/2}], \quad (1.6)$$

where

$$\pi_M(\eta \mid \mathbf{y}) = \sup_{\boldsymbol{\theta} : g(\boldsymbol{\theta}) = \eta} \pi(\boldsymbol{\theta} \mid \mathbf{y}), \quad (1.7)$$

$$\mathbf{b}_\eta = \partial g(\boldsymbol{\theta}_\eta) / \partial \boldsymbol{\theta}_\eta, \quad (1.8)$$

$$\mathbf{R}_\eta = \partial^2 \log \pi(\boldsymbol{\theta} \mid \mathbf{y}) / \partial(\boldsymbol{\theta}\boldsymbol{\theta}^T)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\eta}, \quad (1.9)$$

and $\boldsymbol{\theta}_\eta$ conditionally maximizes (1.5) with respect to $\boldsymbol{\theta}$ for each fixed η. Theorem 1 of Tierney et al. (1988) refers to regularity conditions that require six times continuous differentiability of the transformation in (1.2), within particular regions that remain fixed as $n \to \infty$. Since first derivatives of $g(\boldsymbol{\theta})$ appear in the denominator of (1.6), zero derivatives are likely to cause a problem. Many useful nonlinear transformations possess zero first derivatives for some $\boldsymbol{\theta}$.

We will refer to (1.6)–(1.9) as defining the "KTK procedure." The quadratic term in the denominator of (1.6) is included to slightly extend a suggestion made by Leonard (1982), Leonard and Novick (1986), and Tierney and Kadane (1986), which works well, but mainly in the special cases where $\eta = \theta_j$, for some j, or where η is a linear transformation of $\boldsymbol{\theta}$. The numerical integrations in (1.3) are replaced by conditional maximizations, which can often be completed in straightforward fashion by using standard Newton–Raphson procedures, employing a hill-

climbing package, or using the general algorithms developed at the University of Minnesota by Luke Tierney (see Tierney, Kass, and Kadane 1987).

Kass et al. (1988) justified (1.6) primarily by asymptotic ($n \to \infty$) theory similar to Laplace's (see, e.g., De Bruijn 1970). They showed that, under fairly general regularity conditions, (1.6) differs from the exact posterior density by a term behaving like $O(n^{-1})$ times the true density. Under their regularity conditions, $n^{-1} \log \pi(\theta \mid \mathbf{y})$ converges, as $n \to \infty$, to a twice differentiable function $U(\theta)$. Tierney et al. (1988) discussed the "saddle-point accuracy" of this procedure.

In the next four sections we will see that the KTK procedure is open to further development in some situations when n is finite and $g(\theta)$ is a nonlinear transformation. Furthermore, although straightforward modifications may be available when p is small, these are less obvious when p is large. Although the KTK procedure is well formulated in asymptotic situations as $n \to \infty$, we can find no reasonably general justification in the literature for this procedure when n is finite, unless (1.2) is linear (see Comment 1 of Section 3) or approximately linear. In particular, (1.6) is likely to perform less well in situations where the first derivatives of (1.2) are close to 0 in regions of moderate posterior probability. This problem is illustrated in the next section.

2. EXAMPLES OF THE KTK PROCEDURE

Consider a very simple situation where $p = 1$ and the posterior density of the single parameter θ takes the beta-logistic form

$$\pi(\theta \mid \mathbf{y}) = \pi_y(\theta)$$

$$= \frac{\Gamma(n)}{\Gamma(y)\Gamma(n-y)} \exp\{\theta y\}/(1 + e^\theta)^n,$$

$$-\infty < \theta < \infty, \quad \tfrac{1}{2} < y/n < 1, \quad (2.1)$$

which arises, for example, in the Bayesian analysis of the logistic parameter θ of the binomial distribution.

Consider the two-to-one transformation

$$\eta = g(\theta) = \theta^2. \qquad (2.2)$$

Under the choice in (2.2) the exact posterior density of η is

$$\pi(\eta \mid \mathbf{y}) = \frac{1}{2} \eta^{-1/2} \pi_y(\eta^{1/2}) + \frac{1}{2} \eta^{-1/2} \pi_y(-\eta^{1/2}),$$

$$0 < \eta < \infty, \qquad (2.3)$$

whereas the KTK approximation (1.6) is

$$\tilde{\pi}(\eta \mid \mathbf{y}) \propto \eta^{-1/2} \pi_y(\eta^{1/2}), \qquad 0 < \eta < \infty, \qquad (2.4)$$

which omits the extra term in (2.3) needed for a two-to-one transformation. Although the first term on the right side of (2.3) will predominate when n is large, the second term will noticeably contribute when n is finite, unless the posterior probability, that $\theta < 0$, is negligible. A parallel result is, for example, available for the two-to-one entropy transformation $\eta = \phi \log \phi + (1 - \phi)\log(1 - \phi)$, where

$\phi = e^\theta/(1 + e^\theta)$. There are substantial numerical differences between the corresponding cumulative distribution functions for sample sizes as high as $n = 100$; numerical comparisons were described by Hsu, Leonard, and Tsui (1988). Whenever η in (2.2) is close to 0, the derivative $\partial \eta / \partial \theta = 2\theta$ will be close to 0 (see discussion at the end of Sec. 1).

Similar examples, omitting possibly important terms, are available in a spectrum of single-parameter situations when the transformation $\eta = g(\theta)$ is many-to-one. An example, when $p = 2$, may be constructed when the posterior density of a normal mean θ and log-variance α takes the form

$$\pi(\theta, \alpha \mid \mathbf{y})$$

$$\propto \exp\left\{ -\frac{1}{2} \alpha n - \frac{1}{2} S^2 e^{-\alpha} - \frac{1}{2} n e^{-\alpha}(\theta - \bar{y})^2 \right\},$$

$$-\infty < \theta < \infty, \quad -\infty < \alpha < \infty, \quad (2.5)$$

where $S^2 > 0$. In this case the exact posterior density of $\eta = \sin \theta$ is

$$\pi(\eta \mid \mathbf{y}) \propto \sum_{k=-\infty}^{\infty} [S^2 + n\{\sin^{-1}(\eta) + 2k\pi - \bar{y}\}^2]^{-(1/2)n}$$

$$\times |\cos(\sin^{-1}(\eta) + k)|^{-1}, \quad (2.6)$$

where $\sin^{-1}(\eta)$ denotes the value of $\theta \in (0, \pi)$ for which $\eta = \sin \theta$. However, (1.6) now reduces to

$$\tilde{\pi}(\eta \mid \mathbf{y}) \propto [S^2 + n\{\sin^{-1}(\eta) + 2k_\eta\pi - \bar{y}\}^2]^{-(1/2)n}$$

$$\times |\cos(\sin^{-1}(\eta) + k_\eta)|^{-1}, \quad (2.7)$$

where k_η is the integer minimizing $|\sin^{-1}(\eta) + 2k_\eta\pi - \bar{y}|$. Infinitely many possibly important terms from the exact density (2.6) are now omitted, and the transformation has derivatives close to 0 at many points.

Consider next the multiparameter situation where $\theta_1, \dots, \theta_p$ are a posteriori independent and each has been suitably transformed to possess a standard normal distribution. Then the KTK approximation (1.6) should not be applied to the transformation

$$\eta = \sum_{i=1}^{p} \theta_i^2, \qquad (2.8)$$

since η possesses derivatives close to 0 at many points. Approximation (1.6) is now a chi-squared distribution with 1 df, and the exact answer is chi-squared with p df. The deviations of (1.6) from the exact answer, therefore, compound themselves as the dimensionality increases.

3. AN ALTERNATIVE MARGINALIZATION PROCEDURE

Under fairly general regularity conditions, we may approximate the posterior density $\pi(\theta \mid \mathbf{y})$ in (1.5), for values of θ lying in the region D in (1.4), via a Taylor series expansion of $\log \pi(\theta \mid \mathbf{y})$ about the conditional maximum θ_η of θ, given that $\eta = g(\theta)$. Note that θ_η always lies in D, which is the region that should be considered when performing the integrations in (1.3). Neglecting cubic and

higher terms in this expansion provides the approximation $\pi^*(\boldsymbol{\theta} \mid \mathbf{y})$ to $\pi(\boldsymbol{\theta} \mid \mathbf{y})$, where

$$\log \pi^*(\boldsymbol{\theta} \mid \mathbf{y}) = \log \pi_M(\eta \mid \mathbf{y})$$
$$+ \mathbf{l}_\eta^T(\boldsymbol{\theta} - \boldsymbol{\theta}_\eta) - \tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_\eta)^T \mathbf{R}_\eta(\boldsymbol{\theta} - \boldsymbol{\theta}_\eta) \quad (3.1)$$

with

$$\mathbf{l}_\eta = \partial \log \pi(\boldsymbol{\theta} \mid \mathbf{y})/\partial \boldsymbol{\theta} \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}_\eta} \quad (3.2)$$

and $\pi_M(\eta \mid \mathbf{y})$ and $\mathbf{R}_\eta$ defined in (1.7) and (1.9).

Note that $\mathbf{R}_\eta$ is the posterior information matrix of $\boldsymbol{\theta}$, evaluated at the conditional maximum $\boldsymbol{\theta} = \boldsymbol{\theta}_\eta$. Our method can only be applied if $\mathbf{R}_\eta$ exists and is positive definite for all values of η. The approximation in (3.1) is recommended in situations where the unconditional posterior distribution of $\boldsymbol{\theta}$ is initially judged to be not too far from multivariate normal; suitable transformations such as log or logit may facilitate this.

Under the approximation in (3.1), the integrations in (1.3) can be performed explicitly over the region D, and the marginal posterior density of η reduces, after some rearrangement, to

$$\pi^*(\eta \mid \mathbf{y}) \propto \pi_M(\eta \mid \mathbf{y})|\mathbf{R}_\eta|^{-1/2}$$
$$\times \exp\left\{\frac{1}{2}\mathbf{l}_\eta^T\mathbf{R}_\eta^{-1}\mathbf{l}_\eta\right\}f(\eta \mid \boldsymbol{\theta}_\eta^*, \mathbf{R}_\eta^{-1}), \qquad \eta \in \Omega, \quad (3.3)$$

where the constant of proportionality may be calculated via a one-dimensional numerical integration over $\eta \in \Omega$,

$$\boldsymbol{\theta}_\eta^* = \boldsymbol{\theta}_\eta + \mathbf{R}_\eta^{-1}\mathbf{l}_\eta, \quad (3.4)$$

and $f(\eta \mid \boldsymbol{\mu}, \mathbf{C})$ denotes the density of $\eta = g(\boldsymbol{\theta})$ when $\boldsymbol{\theta}$ possesses a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$. All terms on the right side of (3.3) are important numerically. In particular, the density f should either be obtained analytically, approximated using further techniques, or perhaps calculated by computer simulation, whichever is easiest for the user in special cases. For example, note the following comments.

1. If $\eta = \mathbf{a}^T\boldsymbol{\theta}$ is a linear transformation, f is just a normal density with mean $\mathbf{a}^T\boldsymbol{\mu}$ and variance $\mathbf{a}^T\mathbf{C}\mathbf{a}$. In this special case our suggestion (3.3) and the KTK approximation (1.6) are both algebraically equivalent to the Leonard (1982) Laplacian approximation.

2. If instead $\eta = \boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta}$, with $\mathbf{A}$ denoting a $p \times p$ positive-semidefinite matrix of constants, then the exact f is the density of a linear combination of independent non-central chi-squared variates. For many such linear combinations, however, a gamma approximation with correct mean $\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} + \text{tr}(\mathbf{C}\mathbf{A})$ and variance $2 \, \text{tr}(\mathbf{C}\mathbf{A})^2 + 4\boldsymbol{\mu}^T\mathbf{A}\mathbf{C}\mathbf{A}\boldsymbol{\mu}$ is reasonable.

3. If η is the rth largest of $\theta_1, \ldots, \theta_p$, then f is the density of the corresponding order statistic of a multivariate normal distribution.

4. In situations where $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ is so complex that it is inefficient to simulate $\pi(\eta \mid \mathbf{y})$ based upon random realizations from $\pi(\boldsymbol{\theta} \mid \mathbf{y})$, it may be easier to simulate $f(\eta \mid \boldsymbol{\mu}, \mathbf{C})$ based upon random realizations from a multivariate

normal distribution. Hence all of the terms in (3.3) may be usefully calculated in a wide variety of complex situations.

5. In the single-parameter situation of Section 2, we have, using the correct Jacobian transformation procedure,

$$f(\eta \mid \mu, C) = \tfrac{1}{2}\eta^{-1/2}\Psi_{\eta^{1/2}}(\mu, C) + \tfrac{1}{2}\eta^{-1/2}\Psi_{-\eta^{1/2}}(\mu, C),$$
$$0 < \eta < \infty, \quad (3.5)$$

where $\Psi_\theta(\mu, C)$ denotes a normal density for θ with mean μ and variance C. The extra terms in (3.3) then provide a close approximation to the exact density (2.3), as demonstrated numerically by Hsu et al. (1988), in terms of the cumulative distribution function. For all other transformations considered in Section 2, it is possible to similarly incorporate the correct Jacobian transformation procedure into the approximation. For the multiparameter example, discussed in the penultimate paragraph of Section 2, our method gives the exact chi-squared distribution with p df (in this situation the choice of $\boldsymbol{\theta}_\eta$ is not unique, but our method is exact for any possible choice).

Using theory parallel to that of Kass et al. (1988) it is possible to show that, under regularity conditions similar to those stated in their paper, the error term of (3.3) tends as $n \to \infty$ to $O(n^{-1/2})$ times the true density. Cubic error terms that vanish under the KTK justification of (1.6) do not vanish when integrating over the region (1.4) for our alternative in (3.3). Our procedure, therefore, appears to possess asymptotic properties inferior to KTK. By including the f contribution to (3.3), however, we avoid the omission of terms that may be large when n is finite.

Note that the approximation in (3.3) is well defined even when the transformation $\eta = g(\boldsymbol{\theta})$ is nondifferentiable, or differentiable with zero derivatives. Our approach is somewhat more tedious to apply than KTK, since careful attention needs to be paid to our f contribution, but we wish our approach to work well for finite n and many-to-one transformations (whether or not these are differentiable). As Tierney et al. (1988, sec. 4) implied, there is a trade-off between the complexity of the procedure employed and the potential accuracy of the results.

Our method provides an alternative to Bayesian versions of Edgeworth expansions (e.g., Johnson and Ladalla 1979; Zellner and Rossi 1984), based upon unconditional maximization. In many applications, the approximation in (3.3) will suffice, without further terms, since a conditional maximization has been employed, permitting greater accuracy in regions of interest.

The approximation in (3.3) is related to the profile likelihood ideas of Kalbfleisch and Sprott (1970), who also advocated marginalization based upon conditional maximization. If the prior distribution for $\boldsymbol{\theta}$ is taken to be uniform over R^p, then (3.3) could be interpreted as an "integrated likelihood" for η. Our approach is, therefore, related to the likelihood philosophy.

4. TWO-WAY CONTINGENCY TABLES

Following Leonard (1977) and Leonard and Novick (1986), consider an $r \times s$ contingency table where the cell

counts $\{y_{ij}: i = 1, \ldots, r; j = 1, \ldots, s\}$ are taken to be independent and Poisson distributed, given corresponding cell means of $\{\theta_{ij}: i = 1, \ldots, r; j = 1, \ldots, s\}$. If the θ_{ij} are a priori independent and gamma distributed with respective means α_{ij}/β and variances α_{ij}/β^2, then the $\gamma_{ij} = \log \theta_{ij}$ possess joint posterior density

$$\pi(\boldsymbol{\gamma} \mid \mathbf{y}) \propto \exp\left\{\sum_{ij} (y_{ij} + \alpha_{ij})\gamma_{ij} - (\beta + 1) \sum_{ij} \exp(\gamma_{ij})\right\},$$

$$\boldsymbol{\gamma} = \{\gamma_{ij}\} \in R^{rs}, \quad (4.1)$$

which approaches the likelihood of the γ_{ij} as β and all of the α_{ij} tend to 0. This limiting situation will be assumed in our practical examples, and we will, therefore, effectively marginalize the likelihood function.

Important parameters of interest include the linear transformations (see Comment 1 of Sec. 3)

$$\lambda_{ij}^{AB} = \gamma_{ij} - \gamma_{i.} - \gamma_{.j} + \gamma_{..},$$

$$i = 1, \ldots, r, \quad j = 1, \ldots, s \quad (4.2)$$

(with the dot notation denoting average with respect to that subscript), which denote the interaction effects for the rs cells under Goodman's (1964) full-rank log-linear interaction model, and the many-to-one nonlinear transformation

$$\eta = \sum_{ij} (\gamma_{ij} - \gamma_{i.} - \gamma_{.j} + \gamma_{..})^2/rs, \quad (4.3)$$

which provides an overall measure of association recommended by Altham (1970) for investigating independence of rows and columns in the $r \times s$ table. The usual independence model occurs when $\eta = 0$, so each of the interaction effects in (4.2) is 0.

The marginal posterior density of the linear transformation $\eta = \lambda_{ij}^{AB}$ may be approximated for each (i, j) by application of (3.3) or (1.6) to the joint density (4.1), following Comment 1 of Section 3. For technical details of the conditional maximization, see Leonard and Novick (1986, p. 46). Note that this is simpler than the hierarchical prior techniques proposed by Leonard and Novick (1986, p. 44). Hsu et al. (1988) reported comparisons of our approximations, based on (3.3) or (1.6) with the exact curves, for a series of 2×2 tables and showed that the approximations are virtually identical, to within three significant digits, of the exact results. For example, Fisher (1935) reported cell counts of 2, 15, 10, and 3 for a set of criminal twin data. For these data, (3.3) was accurate, similar to procedures proposed by Barndorff-Nielsen and Cox (1979), for which they reported "saddle-point accuracy."

The hair/eye color data introduced by Snee (1974) are described in Table 1. A key conclusion concerning the (2, 4)th cell was first suggested to us by John Wood; people in the sample with blue eyes are about five times as likely to possess blonde hair. Hsu et al. (1988) reported that approximation (3.3) to the marginal posterior density of $\eta = \lambda_{24}^{AB}$ is extremely close to a histogram representing the exact posterior density and, based upon 40,000 simulations, form 16 independent log-gamma distributions [possessing joint-posterior density (4.1)]. The approximations

Table 1. Cross-Classification of 592 Students According to Hair and Eye Color

	Black hair	Brunette hair	Red hair	Blonde hair	Total
Brown eyes	68	119	26	7	220
Blue eyes	20	84	17	94	215
Hazel eyes	15	54	14	10	93
Green eyes	5	29	14	16	64
Total	108	286	71	127	592

were similarly accurate for the posterior densities of the interaction effects for all 16 cells in Table 1, though Lindley's (1964) normal approximation was only slightly less accurate.

As the posterior distribution of λ_{24}^{AB} was almost completely concentrated on the positive part of the real line, we can conclude that the interaction effect for this cell is almost certainly positive. In general, we recommend an applied assessment of the full posterior densities of all of the interactions, taking into account posterior tail probabilities, practical significance, and patterns across the table. Our interaction/residual analysis, summarized in Table 2, is intended as a practical diagnostic, helping the user to decide on the next stage of his analysis (e.g., immediate practical conclusions, collapsing to 2×2 tables, or setting some of the interactions equal to 0). Our methodology permits more precise finite sample inferences than are, for example, available under asymptotic maximum likelihood theory (Goodman 1964).

5. THE OVERALL MEASURE OF ASSOCIATION

Now consider applying (3.3) to approximate the marginal posterior density of the measure of association in (4.3). This permits the user to make an overall inference regarding the closeness of the data to the independence model and, therefore, provides an alternative to a goodness-of-fit test.

Maximizing (4.1) subject to the constraint in (4.3) yields the nonlinear equations

$$(\beta + 1)\exp\{\gamma_{ij}^{(\eta)}\} = y_{ij} + \alpha_{ij} - k_{\eta}$$

$$\times [(1 - r^{-1} - s^{-1} + r^{-1}s^{-1})\gamma_{ij}^{(\eta)} - \mu_{ij}],$$

$$i = 1, \ldots, r, \quad j = 1, \ldots, s, \quad (5.1)$$

for the conditional maxima $\gamma_{ij}^{(\eta)}$ of the γ_{ij} given (4.3). Here k_{η} denotes a Lagrange multiplier, and

$$\mu_{ij} = r^{-1} \sum_{k:k\neq i} \gamma_{kj}^{(\eta)} + s^{-1} \sum_{g:g\neq j} \gamma_{ig}^{(\eta)} - r^{-1}s^{-1} \sum_{(k,g)\neq(i,j)} \gamma_{kg}^{(\eta)}.$$

$$(5.2)$$

Table 2. Interaction Analysis for Eye/Hair Color Data

	Black hair	Brunette hair	Red hair	Blonde hair
Brown eyes	+* (1.000)	+ (.989)	0 (.648)	−* (.000)
Blue eyes	− (.012)	− (.024)	− (.001)	+* (1.000)
Hazel eyes	0 (.717)	0† (.824)	0 (.654)	−† (.077)
Green eyes	− (.005)	0† (.121)	+ (.973)	+ (.992)

NOTE: + denotes positive interaction; 0 denotes no interaction; − denotes negative interaction. An * means that the interaction is very noticeable; a † means that the conclusion regarding an interaction is unclear. The figures in parentheses denote our approximate posterior probabilities that the interactions are positive.

Equation (5.1) may be solved, for fixed μ_{ij}, using one-dimensional Newton–Raphsons for each $\gamma_{ij}^{(\eta)}$. Then cyclic substitution on the μ_{ij} in (5.2) will suffice for convergence. For each k_η, the corresponding η may be calculated from (4.3), and this function may be inverted computationally to express k_η in terms of η. Then substitution of the $\gamma_{ij}^{(\eta)}$ for the corresponding γ_{ij} in (4.1) provides the maximized posterior density $\pi_M(\eta \mid \mathbf{y})$. In the extreme right tail of this function there are some slight problems, because an η may yield two different solutions for the $\gamma_{ij}^{(\eta)}$. In this case, it is best to choose the $\gamma_{ij}^{(\eta)}$ maximizing (4.1).

For the hair/eye color data, $\pi_M(\eta \mid \mathbf{y})$ is described by curve b of Figure 1 and is too skewed to the right when compared with the histogram representing the exact results (based upon the same 40,000 simulations as described in Sec. 3). Under this nonlinear, many-to-one, transformation the KTK adjustments in (1.6) are insufficient to distinguishably change $\pi_M(\eta \mid \mathbf{y})$ in this particular case. Therefore, KTK does not well approximate the true posterior density of η, presumably because the problems highlighted in Section 2, for example, surrounding the transformation in (2.8), become compounded as the dimensionality increases. Our suggestion in (3.3), however, with the density f for η approximated by a gamma curve, with correct first two moments (see Comment 2 of Sec. 3), leads to curve a of Figure 1. This is reassuringly close to the simulated exact results.

Hsu et al. (1988) described a method for interpreting the marginal density of η. They also provided a further example of the procedure described in this section, where the KTK approximation is totally different from the exact result, overheavily skewed to the right, and adjusts $\pi_M(\eta \mid \mathbf{y})$ in an incorrect direction. It would appear that KTK does not always fare well under the particular transformation (4.3).

In this particular situation our approximate analytic procedure is easier to apply than computer simulations, since (a) this involves much less (about 70% less) computer time and (b) the final result is already a smooth curve, rather than a histogram requiring somewhat arbitrary smoothing. In the more complicated situation described in Section 6, a slower "importance sampling" procedure is needed to simulate the exact posterior density. However, in some
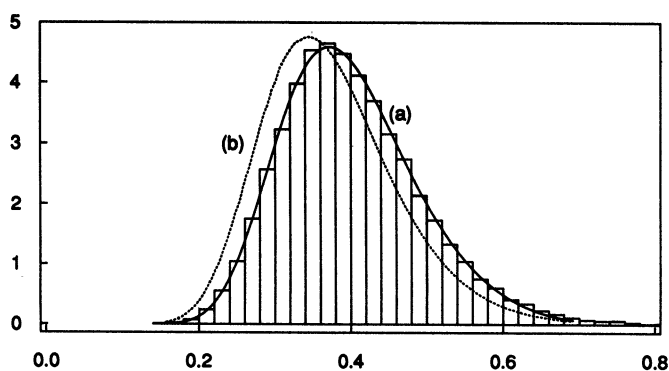

Figure 1. Posterior Densities and Histogram of Overall Measure of Association for the Eye/Hair Color Data. The histogram is based on 40,000 observations generated from the true posterior distribution: (a) ——, our approximation; (b) ···, maximized posterior density and KTK approximation.

situations (e.g., research in progress concerning the posterior density of the mean of a mixture of exponential distributions), it is simpler to simulate rather than conditionally maximize. Therefore, in any particular situation, the user will need to make a practical judgment. In any case, it is always important to check the accuracy of the conditional maximization procedure, for example, by at least simulating the first few posterior moments using the method outlined in Section 6.

6. THE SEMIPARAMETRIC ESTIMATION OF A DISCRETE DISTRIBUTION

In this section we consider a procedure for the semiparametric estimation of a discrete distribution concentrated on the integers $0, 1, \ldots, m$. Let $y_1, \ldots, y_n$ denote a random sample from a discrete probability distribution, with probability mass function

$$p(y = j \mid \boldsymbol{\theta}) = \phi_j, \qquad j = 0, \ldots, m, \qquad (6.1)$$

where multivariate logits $\gamma_1, \ldots, \gamma_n$ satisfying

$$\phi_j = e^{\gamma_j} \Big/ \sum_{h=0}^{m} e^{\gamma_h} \qquad (6.2)$$

are taken to possess the functional form

$$\gamma_j = a_j + \theta_1 j + \theta_2 j^2 + \cdots + \theta_p j^p, \qquad p < m, \qquad (6.3)$$

with the a_j and p specified and $\theta_1, \ldots, \theta_p$ denoting unknown parameters.

When $a_j = \log{}^m C_j$, with ${}^m C_j = m!/j!(m - j)!$, and $\theta_2 = \cdots = \theta_p = 0$, this specification reduces to a binomial distribution for each y_i, with probability $e^{\theta_1}/(1 + e^{\theta_1})$ and number of trials m. With this choice of a_j, (6.3) more generally permits investigations of deviations from a binomial assumption. This model could, therefore, be used to investigate reasonability of an assumption of simple random sampling from a large population by n sample surveyors, each questioning m people, with yes/no responses. Alternatively, with each a_j set equal to 0 and both p and m large, (6.3) permits an effectively nonparametric fit to a discrete distribution. It would be possible to use other basis functions, for example, B-splines or terms from a Fourier series, instead of $j, j^2, \ldots, j^p$, and still remain within the discrete p-parameter exponential family.

Parameters of interest include the sampling probabilities $\phi_0, \phi_1, \ldots, \phi_m$, each of which is a nonlinear many-to-one function of $\theta_1, \ldots, \theta_p$. For ease of presentation, attention is confined here to

$$\eta = \phi_0 = e^{a_0} \Big/ \sum_{h=0}^{m} \exp\{a_h + \theta_1 h + \theta_2 h^2 + \cdots + \theta_p h^p\},$$

$$(6.4)$$

which represents the probability that any particular observation is 0. This probability is of particular interest in machine component reliability situations where the m components may not fail independently, or with equal probability, and where ϕ_0 is the probability that the machine does not fail (this event occurs when none of the m components fail, and $y_1, \ldots, y_n$ denote the number of component failures on n separate runs).

If prior information is available, then a multivariate normal prior distribution for $\theta_1, \ldots, \theta_p$ could be employed. In the absence of prior information, we assume a uniform prior distribution for $\theta_1, \ldots, \theta_p$ over R^p, in which case the posterior density, if it exists, is the likelihood function. That is,

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) \propto \eta^n \exp\{\theta_1 t_1 + \theta_2 t_2 + \cdots + \theta_p t_p\}, \quad \boldsymbol{\theta} \in R^p, \tag{6.5}$$

where η satisfies (6.4), and

$$t_k = \sum_{j=1}^{m} n_j j^k, \quad k = 1, \ldots, p, \tag{6.6}$$

with n_j representing the number of observations out of $y_1, \ldots, y_n$ that equal j.

It is straightforward to conditionally maximize the density in (6.5) with respect to $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$, given the constraint in (6.4). For each fixed η, solve the $p + 1$ equations

$$\sum_{h=0}^{m} \exp\{a_h + \theta_1 h + \theta_2 h^2 + \cdots + \theta_p h^p\} = e^{a_0}/\eta \tag{6.7}$$

and

$$\sum_{h=0}^{m} h^k \exp\{\lambda + a_h + \theta_1 h + \theta_2 h^2 + \cdots + \theta_p h^p\} = t_k,$$
$$k = 1, \ldots, p, \tag{6.8}$$

with respect to $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ and the quantity λ (which is a function of the Lagrange multiplier). The solution for $\boldsymbol{\theta}$ provides the conditional maximum $\boldsymbol{\theta}_\eta$.

Hence, by including adjustment terms, we may readily compute the KTK approximation (1.6) together with our proposal in (3.3). The f contribution to (3.3), however, demands careful attention. Note that when $\boldsymbol{\theta}$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$, it is possible to calculate all moments of the contribution

$$\zeta = \zeta(\boldsymbol{\theta}) = \sum_{h=1}^{m} \exp\{a_h + \boldsymbol{\theta}^T \boldsymbol{\epsilon}_h\} \tag{6.9}$$

to (6.4), where $\boldsymbol{\epsilon}_h = (h, h^2, \ldots, h^p)^T$. For example, the mean and variance of ζ are

$$\zeta^* = \sum_{h=1}^{m} \exp\{a_h + \boldsymbol{\mu}^T \boldsymbol{\epsilon}_h + \tfrac{1}{2} \boldsymbol{\epsilon}_h^T \mathbf{C} \boldsymbol{\epsilon}_h\} \tag{6.10}$$

and

$$v^* = \sum_{h=1}^{m} \sum_{k=1}^{m} \exp\left\{ a_h + a_k + \boldsymbol{\mu}^T(\boldsymbol{\epsilon}_h + \boldsymbol{\epsilon}_k) \right.$$
$$\left. + \frac{1}{2} \boldsymbol{\epsilon}_h^T \mathbf{C} \boldsymbol{\epsilon}_h + \frac{1}{2} \boldsymbol{\epsilon}_k^T \mathbf{C} \boldsymbol{\epsilon}_k \right\} [\exp(\boldsymbol{\epsilon}_h^T \mathbf{C} \boldsymbol{\epsilon}_k) - 1]. \tag{6.11}$$

Therefore, the distribution of ζ may be approximated, within a suitable family, by requiring the first several moments to be correct. One possibility is to take ζ to possess

a lognormal distribution with means ζ^* and variance v^*. Then the density f of η may be approximated via the transformation $\eta = e^{a_0}/(e^{a_0} + \zeta)$. Alternatively, the density f of η in (6.4) could be simulated directly, based upon random realizations for $\boldsymbol{\theta}$ from a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$.

It is still possible to simulate the exact posterior distribution of $\eta = \phi_0$ in (6.4) under the complicated posterior distribution of $\boldsymbol{\theta}$ in (6.5). Let $w(\boldsymbol{\theta})$ denote the function of $\boldsymbol{\theta}$ defined by the right side of (6.5), where $\eta = \eta(\boldsymbol{\theta})$ is the expression on the right side of (6.4). Note that if the maximum likelihood vector $\hat{\boldsymbol{\theta}}$ and likelihood information matrix

$$\mathbf{R} = -\left. \frac{\partial^2 \log w(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} \tag{6.12}$$

exist, then the posterior distribution in (6.5) is roughly multivariate normal with mean vector $\hat{\boldsymbol{\theta}}$ and covariance matrix $\mathbf{R}^{-1}$. Consider l simulations $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_l$ from this multivariate normal distribution. Then, using standard results for importance sampling, the kth posterior moment of η is exactly

$$E(\eta^k \mid \mathbf{y}) = \lim_{l \to \infty} \sum_{j=1}^{l} [\eta^k(\boldsymbol{\theta}_j) w(\boldsymbol{\theta}_j)/\Psi_{\boldsymbol{\theta}_j}(\hat{\boldsymbol{\theta}}, \mathbf{R}^{-1})]$$
$$\times \left[\sum_{j=1}^{l} w(\boldsymbol{\theta}_j)/\Psi_{\boldsymbol{\theta}_j}(\hat{\boldsymbol{\theta}}, \mathbf{R}^{-1}) \right]^{-1}, \tag{6.13}$$

where $\Psi_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}, \mathbf{R}^{-1})$ denotes our approximate multivariate normal density for $\boldsymbol{\theta}$. See, for example, Rubinstein (1981).

It is possible to similarly compute the posterior cumulative distribution function of η by referring to the posterior expectations of appropriate indicator functions. The posterior density of η may, therefore, be closely approximated, for example, by a histogram. (See Tanner and Wong 1987.)

7. POSTERIOR INFERENCES FOR THE PROBABILITY OF A ZERO COUNT

We now consider two examples of the methodology discussed in Section 6 for inferences about ϕ_0. Two hundred observations were first randomly generated from a mixture of two binomial distributions, each with sample size $m = 10$, where the binomial probability parameter was equal to .2 with probability $\frac{1}{3}$, and equal to .7 with probability $\frac{2}{3}$, yielding a true value of $\phi_0 = .0358$. The observed frequencies of the integers $0, 1, 2, \ldots, 10$ were 8, 12, 17, 18, 12, 23, 27, 34, 31, 14, and 4, respectively.

These data may be closely fitted by the model in (6.3) with $p = 4$ and $a_j = 0$. For example, the maximum likelihood fit provides a value $X^2 = 2.94$ for the usual chi-squared goodness-of-fit statistic, with 6 df and a maximum likelihood estimate $\hat{\phi}_0 = .040$ for ϕ_0, which perfectly fits the observed relative frequency. The maximum likelihood estimates for $\theta_1, \theta_2, \theta_3$, and θ_4 were $\hat{\theta}_1 = 7.9$, $\hat{\theta}_2 = -34.0$, $\hat{\theta}_3 = 63.4$, and $\hat{\theta}_4 = -38.2$ with respective approximate

standard errors 4.5, 18.4, 28.2, and 14.3, which use the approximate covariance matrix $\mathbf{R}^{-1}$ with $\mathbf{R}$ defined in (6.12).

The histogram in Figure 2 represents 20,000 simulations for the exact posterior density of ϕ_0, using the importance sampling procedure outlined at the end of Section 6 and a uniform prior for $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)^T$. Curve b normalizes the maximized posterior density (1.7) for $\eta = \phi_0$ or, equivalently, Kalbfleisch and Sprott's (1970) profile likelihood, to ensure that this integrates to unity. Curve a adjusts curve b using our modified conditional maximization procedure outlined in Section 6. This involves the simplest approximation described in Section 6 to the f contribution to (3.3), based upon a lognormal distribution for the expression in (6.9). Even with this further approximation, curve a seems to be remarkably close to the exact posterior density. However, the KTK curve c adjusts curve b, based on Kalbfleisch and Sprott's (1970) idea, in an opposite direction. Although the mode of curve a decreases $\hat{\phi}_0 = .04$ to $\phi_0 = .0354$, the mode of curve c lies at $\phi_0 = .042$.

Comparisons of the corresponding approximations to the posterior mean, standard deviation, and coefficients of skewness and kurtosis are described in the first four rows of Table 3. The quantities in the second, third, and fourth rows correspond to curves a, b, and c, respectively, in Figure 2. Our approximation again fares well when compared with the simulated exact results.

For our second example, 100 observations were randomly generated from a mixture of two binomial distributions, again with sample size $m = 10$, but where the binomial probability parameter was equal to .1 with probability .25 and equal to .8 with probability .75, giving $\phi_0 = .0872$. The observed frequencies of the integers 0, 1, 2, . . . , 10 were now 8, 9, 6, 1, 1, 1, 7, 13, 27, 18, 9, so that an approach accommodating small counts seems desirable.

With $p = 4$ and $a_j = 0$, the maximum likelihood fit now provides $X^2 = 7.38$ with 6 df and the maximum likelihood estimate $\hat{\phi}_0 = .089$ compares with an observed relative frequency of .08. We computed the various approximations to the posterior density of ϕ_0, and the results were
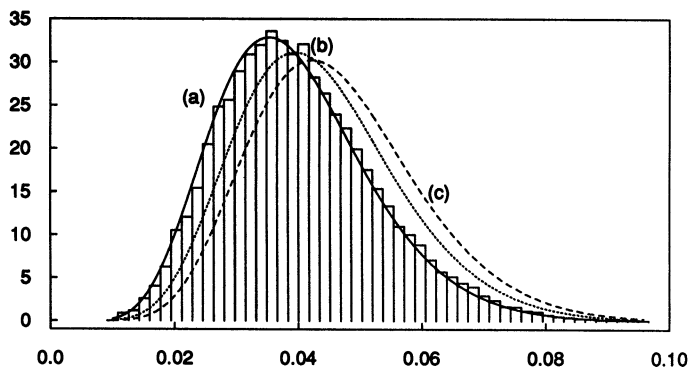

Figure 2. Posterior Densities and Histogram of ϕ_0 for the Binomial Mixture Example. The histogram is based on 20,000 simulations for exact posterior density using importance sampling: (a) ——, our approximation; (b) ···, maximized posterior density; (c) ---, KTK approximation.

Table 3. Comparisons of Posterior Approximations for the Probability of a Zero Count

	Mean	Standard deviation	Coefficient of skewness	Coefficient of kurtosis
First example				
Exact (simulated)	.0394	.0128	.659	.636
Our approximation	.0395	.0126	.625	.438
Normalized profile likelihood	.0436	.0132	.545	.233
KTK approximation	.0464	.0135	.496	.105
Second example				
Exact (simulated)	.0900	.0276	.555	.435
Our approximation	.0895	.0272	.492	.104
Normalized profile likelihood	.0972	.0277	.455	−.140
KTK approximation	.1026	.0281	.341	−.130

similar in spirit to those described for the first example of this section. For example, our approximation was again much closer to the simulated exact posterior density than either normalized profile likelihood or the KTK approximation. We approximated the posterior mode to lie at $\phi_0 = .081$, when compared with the KTK value of $\phi_0 = .095$. Further comparisons are described in the last four rows of Table 3.

8. CONCLUDING REMARKS

The marginalization procedure described in Section 3 has worked well in practice in all examples that we have considered so far, though some care needs to be taken in special situations, and the conditions discussed in the second paragraph of Section 3 should be seriously evaluated. For example, Leonard, Tsui, and Hsu (in press) discussed a problem concerning hyperparameters in hierarchical models and instead recommended a single-dimensional numerical integration, together with some conditional maximizations. The implications of (3.3) for the analysis of nonlinear models appear to be important. It is similarly possible to approximate the predictive distributions of future observations or statistics. Methods available for prediction were reviewed by Bjornstad (in press) and Leonard et al. (in press).

REFERENCES

Altham, P. M. E. (1970), "The Measure of Association of Rows and Columns for an $r \times s$ Contingency Table," Journal of the Royal Statistical Society, Ser. B, 32, 63–73.

Barndorff-Nielsen, O., and Cox, D. R. (1979), "Edgeworth and Saddlepoint Approximation" (with discussion), Journal of the Royal Statistical Society, Ser. B, 4, 279–312.

Bjornstad, Jan (in press), "Predictive Likelihood—A Review," Statistical Science, 4.

De Bruijn, N. G. (1970), Asymptotic Methods in Analysis, Amsterdam: North-Holland.

Fisher, R. A. (1935), "The Logic of Inductive Inference," Journal of the Royal Statistical Society, Ser. A, 98, 39–54.

Goodman, L. A. (1964), "Interactions in Multidimensional Contingency Tables," The Annals of Mathematical Statistics, 35, 632–646.

Hsu, J. S. J., Leonard, T., and Tsui, K. W. (1988), "Bayesian and Likelihood Marginal Inference With Applications to Contingency Table Analysis," Technical Report 825 (revised version), University of Wisconsin–Madison, Dept. of Statistics.

Johnson, R. A., and Ladalla, J. N. (1979), "The Large Sample Behavior of Posterior Distributions Which Sample From Multiparameter Ex-

ponential Family Models, and Allied Results," *Sankyhā*, Ser. B, 41, 196–215.

Kalbfleisch, J. D., and Sprott, D. A. (1970), "Application of Likelihood Methods to Models Involving Large Numbers of Parameters," *Journal of the Royal Statistical Society*, Ser. B, 32, 175–208.

Kass, R. E., Tierney, L., and Kadane, J. B. (1988), "The Validity of Posterior Expansions Based on Laplace's Method," Technical Report 396, Carnegie-Mellon University, Dept. of Statistics.

——— (in press), "Asymptotics in Bayesian Computation," in *Bayesian Statistics 3*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Valencia, Spain: University of Valencia Press.

Leonard, T. (1977), "A Bayesian Approach to Some Multinomial Estimation and Pretesting Problems," *Journal of the American Statistical Association*, 72, 865–868.

——— (1982), Comment on "A Simple Predictive Density Function," by M. Lejeune and G. D. Faulkenberry, *Journal of the American Statistical Association*, 77, 657–658.

Leonard, T., and Novick, M. R. (1986), "Bayesian Full Rank Marginalization for Two-Way Contingency Tables," *Journal of Educational Statistics*, 11, 33–56.

Leonard, T., Tsui, K. W., and Hsu, J. S. J. (in press), Comment on "Predictive Likelihood—A Review," *Statistical Science*, 4.

Lindley, D. V. (1964), "The Bayesian Analysis of Contingency Tables," *The Annals of Mathematical Statistics*, 35, 1622–1643.

Rubinstein, R. Y. (1981), *Simulation and the Monte Carlo Method*, New York: John Wiley.

Snee, R. (1974), "Graphical Display of Two-Way Contingency Tables," *The American Statistician*, 38, 9–12.

Tanner, T. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–549.

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

Tierney, L., Kass, R. E., and Kadane, J. B. (1987), "Interactive Bayesian Analysis Using Accurate Asymptotic Approximations," in *Proceedings of the 19th Symposium on the Interface*, ed. R. M. Heiberger, Alexandria, VA: American Statistical Association, pp. 15–21.

——— (1988), "Approximate Marginal Densities of Non-linear Functions," Technical Report 513, University of Minnesota, Dept. of Statistics (submitted to *Biometrika*).

Zellner, A., and Rossi, P. E. (1984), "Bayesian Analysis of Dichotomous Quantal Response Models," *Journal of Econometrics*, 25, 265–393.