Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models)

Robert E. Kass; Duane Steffey

Journal of the American Statistical Association, Vol. 84, No. 407 (Sep., 1989), 717-726.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28198909%2984%3A407%3C717%3AABIICI%3E2.0.CO%3B2-Q

Journal of the American Statistical Association is currently published by American Statistical Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/astata.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org/
Thu May 5 20:40:07 2005

Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models)

ROBERT E. KASS and DUANE STEFFEY*

We consider two-stage models of the kind used in parametric empirical Bayes (PEB) methodology, calling them *conditionally independent hierarchical models*. We suppose that there are k "units," which may be experimental subjects, cities, study centers, etcetera. At the first stage, the observation vectors Y_i for units $i = 1, \ldots, k$ are independently distributed with densities $p(y_i \mid \theta_i)$, or more generally, $p(y_i \mid \theta_i, \lambda)$. At the second stage, the unit-specific parameter vectors θ_i are iid with densities $p(\theta_i \mid \lambda)$. The PEB approach proceeds by regarding the second-stage distribution as a prior and noting that, if λ were known, inference about θ could be based on its posterior. Since λ is not known, the simplest PEB methods estimate the parameter λ by maximum likelihood or some variant, and then treat λ as if it were known to be equal to this estimate. Although this procedure is sometimes satisfactory, a well-known defect is that it neglects the uncertainty due to the estimation of λ. In this article we suggest that approximate Bayesian inference can provide simple and manageable solutions to this problem. In Bayesian inferences, a prior density $\pi(\cdot)$ on λ is introduced, the posterior $p(\lambda \mid \mathbf{y})$ is calculated, and the posterior density of θ_i is then equal to the expectation, with respect to $p(\lambda \mid \mathbf{y})$, of the conditional posterior $p(\theta_i \mid y_i, \lambda)$. From the Bayesian point of view, the PEB estimate is of interest because it is a first-order approximation to the posterior mean [having an error of order $O(k^{-1})$]. Letting E_λ and V_λ denote the expectation and variance with respect to $p(\lambda \mid \mathbf{y})$, we may write the posterior variance of θ_i as $V(\theta_i \mid \mathbf{y}) = E_\lambda\{V(\theta_i \mid y_i, \lambda)\} + V_\lambda\{E(\theta_i \mid y_i, \lambda)\}$. The conditional posterior variance $V(\theta_i \mid y_i, \hat{\lambda})$, where $\hat{\lambda}$ is the maximum likelihood estimator, approximates only the first term. When we include an approximation to the second term we obtain a first-order approximation to the posterior variance itself. In many examples, this elementary method, incorporating approximations to both terms, will substantially account for the estimation of λ. We briefly consider second-order approximations, noting that the work of Deely and Lindley (1981) may be extended using expansions derived by Lindley (1980), Mosteller and Wallace (1964), Tierney and Kadane (1986), and Tierney, Kass, and Kadane (1989). We suggest that second-order approximations provide rough and, often, easily computed assessments of accuracy of first-order approximations. Although we confine our data-analytical examples to simple models, we believe the methods will be useful in general settings. An important area of application is longitudinal data analysis.

KEY WORDS: Asymptotic posterior; Asymptotic variance; Bayes empirical Bayes; Hyperparameters; Laplace's method; Longitudinal analysis; Mixed models; Random-effects models.

1. INTRODUCTION

When data are collected from many units that are somehow similar, such as subjects, animals, cities, etcetera, the statistical problem is to combine the information from the various units to understand better the phenomenon under study. Usually there is substantial variability among units, and a natural way to approach the problem is to build a two-stage "hierarchical model" and then use it to make inferences. In general, a two-stage hierarchical model for a random vector Y is a specification of a first-stage family of densities $\{p(y \mid \theta, \lambda) : \theta \in \Theta, \lambda \in \Lambda\}$ for Y conditional on θ and λ and a second-stage family of densities $\{p(\theta \mid \lambda) : \lambda \in \Lambda\}$ for θ conditional on a second-stage parameter λ; the parameter spaces Θ and Λ may be multidimensional. Here, we are concerned with two-stage hierarchical models for observation vectors Y_i of length n_i on units $i = 1, \ldots, k$, in which the first-stage parameters θ, which have their distribution modeled in the second stage, are unit-specific, that is, they are indexed by a subscript i.

We assume that the vector pairs (Y_i, θ_i) are independent across units, conditionally on λ. Put differently, we have the following specification for what we will call a *conditionally independent hierarchical model* (CIHM).

Stage 1. Conditionally on $(\theta_1, \ldots, \theta_k)$ and λ, the vectors Y_i are independent with densities $p(y_i \mid \theta_i, \lambda)$ $(i = 1, \ldots, k)$ belonging to a family $\{p(y \mid \theta, \lambda) : \theta \in \Theta, \lambda \in \Lambda\}$.

Stage 2. Conditionally on λ, the vectors θ_i are iid with density belonging to a family $\{p(\theta \mid \lambda) : \lambda \in \Lambda\}$.

We take the dimension of Θ to be p and that of Λ to be m. In this article, using an approach similar to that taken previously by Deely and Lindley (1981) for one-parameter problems, we discuss approximate Bayesian inference about the unit-specific parameters $\theta_1, \ldots, \theta_k$, which accounts for the uncertainty introduced through the estimation of λ. We use several examples to illustrate elementary asymptotic results, focusing much of our attention on a "delta method" approximation to the posterior variance.

CIHM's are sometimes called "parametric empirical Bayes models" because the obvious interpretation of the second-stage densities as priors led to the development of "parametric empirical Bayes" (PEB) methodology (Efron

* Robert E. Kass is Associate Professor, Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA 15213. His work was supported by National Science Foundation Grants DMS-8503019 and DMS-8705646. Duane Steffey is Assistant Professor, Department of Mathematical Sciences, San Diego State University, San Diego, CA 92182. The authors gratefully acknowledge the many helpful comments and suggestions of the referees and the associate editor, which led to substantial improvements in the article.

© 1989 American Statistical Association
Journal of the American Statistical Association
September 1989, Vol. 84, No. 407, Theory and Methods

and Morris 1973, 1975), following on the earlier nonparametric empirical Bayes approach of Robbins (1964). Since that terminology should refer to statistical methods rather than probabilistic models, however, we prefer the more descriptive label we have introduced here. In Bayesian inference, a prior $\pi(\cdot)$ on λ is introduced, and inferences are based on the posterior distributions of the θ_i's and λ. Our Stage 2 is often called a prior, rather than being part of "the model," but from the Bayesian point of view this distinction is unimportant. We prefer our terminology because, in our treatment, the role played by the distribution on λ is analogous to that of the prior in a single-stage model. We also note that when CIHM's are applied in the context of repeated-measures analysis, the usual terminology (e.g., of Laird and Ware 1982) distinguishes the parameters θ_i and λ by saying that they refer to "random" and "fixed" effects. This makes sense from the non-Bayesian point of view, but the connotation would rarely be appropriate within the Bayesian paradigm. The more fundamental characterization is that the θ_i's vary among units and λ is common to all units, so we prefer to call the θ_i's *unit-specific* parameters and λ the *common* parameter. See Lindley (1971) for an early discussion of the Bayesian approach to CIHM's [see, too, the comment on that paper by Kempthorne, as well as Good (1980), for historical remarks about hierarchical models].

To emphasize the purpose of our discussion, we remind the reader that in PEB methodology, once λ is estimated by maximum likelihood or some variant, inferences about the θ_i's are based on their conditional posterior means, conditionally on λ, with λ set equal to its estimated value; that is, in familiar notation, $E(\theta_i \mid y_i, \hat{\lambda})$ is used to estimate θ_i. Other related estimators are sometimes used as well. For a review of the basic methodology, see Morris (1983). From the Bayesian point of view, the PEB approach effectively substitutes the conditional posterior density $p(\theta_i \mid y_i, \lambda)$, with $\lambda = \hat{\lambda}$, for the correct posterior density given by

$$p(\theta_i \mid \mathbf{y}) = \int p(\theta_i \mid \mathbf{y}, \lambda) p(\lambda \mid \mathbf{y})\, d\lambda, \tag{1.1}$$

where the posterior density $p(\lambda \mid \mathbf{y})$ is based on the prior $\pi(\lambda)$. Thus the PEB approximation $p(\theta_i \mid y_i, \lambda)$, with $\lambda = \hat{\lambda}$, fails to take account of the uncertainty about λ, which enters (1.1) through the posterior density $p(\lambda \mid \mathbf{y})$. On the one hand, it is well known that the PEB estimate $E(\theta_i \mid y_i, \hat{\lambda})$ is approximately equal to the posterior mean $E(\theta_i \mid \mathbf{y})$. On the other hand, it is equally well recognized that the corresponding variance estimate $V(\theta_i \mid y_i, \hat{\lambda})$ is too small. In Section 3 we formalize these observations with asymptotic statements that furnish approximations as $k \to \infty$, and we obtain a correction term for the variance. This follows a list of several important consequences of the CIHM structure, in Section 2. We provide some interpretation of the corrected first-order variance approximation, having relative error of order $O(k^{-1})$, and then briefly discuss second-order approximations, recommending their use to check accuracy of first-order approximations. In Section 4 we exemplify and interpret the approximations in four examples, three of which are data-analytical, and in Section 5 we comment on several remaining issues.

Though our treatment is Bayesian throughout, we believe the first-order variance approximation we present in Section 3 could also be justified as a variance estimate in non-Bayesian theory. For a recent non-Bayesian alternative, based on the bootstrap, see Laird and Louis (1987).

2. CONDITIONALLY INDEPENDENT HIERARCHICAL MODELS

For subsequent use we review here several immediate consequences of the conditional independence structure of CIHM's. The first is due to the independence of $Y_1, \ldots, Y_k$ conditionally on λ. The compound sampling density of the data $\mathbf{y} = (y_1, \ldots, y_k)$ becomes

$$p(\mathbf{y} \mid \lambda) = \prod_{i=1}^{k} p(y_i \mid \lambda), \tag{2.1}$$

where

$$p(y_i \mid \lambda) = \int p(y_i \mid \theta_i, \lambda) p(\theta_i \mid \lambda)\, d\theta_i \tag{2.2}$$

and the likelihood function $L(\cdot)$ on Λ becomes

$$L(\lambda) = \prod_{i=1}^{k} L_i(\lambda), \tag{2.3}$$

where $L_i(\lambda) = p(y_i \mid \lambda)$. Similarly, we have the simplification

$$p(\theta_i \mid \mathbf{y}, \lambda) = p(\theta_i \mid y_i, \lambda). \tag{2.4}$$

Now suppose that g is a real-valued function on Θ. From (2.4), the conditional posterior expectation of $g(\theta_i)$ satisfies $E(g(\theta_i) \mid \mathbf{y}, \lambda) = E(g(\theta_i) \mid y_i, \lambda)$, so the posterior expectation of $g(\theta_i)$, $E(g(\theta_i) \mid \mathbf{y}) = \int \int g(\theta_i) p(\theta_i \mid \mathbf{y}, \lambda) p(\lambda \mid \mathbf{y})\, d\theta_i\, d\lambda$, satisfies

$$E(g(\theta_i) \mid \mathbf{y}) = E_\lambda\{E(g(\theta_i) \mid y_i, \lambda)\}, \tag{2.5}$$

where E_λ is the expectation with respect to the posterior distribution of λ. The corresponding expression for the variance is

$$V(g(\theta_i) \mid \mathbf{y}) = E_\lambda\{V(g(\theta_i) \mid y_i, \lambda)\} + V_\lambda\{E(g(\theta_i) \mid y_i, \lambda)\}, \tag{2.6}$$

where V_λ is the variance with respect to the posterior distribution of λ.

3. ASYMPTOTIC APPROXIMATIONS

3.1 First-Order Approximations

From (2.1), conditionally on λ, the Y_i's are iid observations from the compound parametric family of distributions with densities $p(y_i \mid \lambda)$. Under mild regularity conditions the posterior distribution on λ is asymptotically Normal with mean and variance given by the posterior mode $\tilde{\lambda}$ and the inverse of the negative Hessian of the log posterior evaluated at the mode $\tilde{\Sigma} = (-D^2 \log(L\pi)(\tilde{\lambda}))^{-1}$. In addition to limiting Normality, we have

$$E(\lambda \mid \mathbf{y}) = \tilde{\lambda} + O(k^{-1}) \tag{3.1}$$

and

$$V(\lambda \mid \mathbf{y}) = \tilde{\Sigma} + O(k^{-2}). \tag{3.2}$$

Now, if G is a smooth real-valued function on Λ having nonzero derivative (Jacobian) DG at $\tilde{\lambda}$, then a Taylor series expansion (the *delta method*) yields

$$E(G(\lambda) \mid \mathbf{y}) = G(\tilde{\lambda}) + O(k^{-1}) \tag{3.3}$$

and

$$V(G(\lambda) \mid \mathbf{y}) = (DG)^T\tilde{\Sigma}(DG) + O(k^{-2}), \tag{3.4}$$

where $DG = DG(\tilde{\lambda})$. Assuming that G is of constant order [i.e., $G(\lambda) = O(1)$ and $1/G(\lambda) = O(1)$] and both $G(\tilde{\lambda})$ and $DG(\tilde{\lambda})$ are nonzero, these approximations are equivalent to the multiplicative versions

$$E(G(\lambda) \mid \mathbf{y}) = G(\tilde{\lambda})\{1 + O(k^{-1})\} \tag{3.5}$$

and

$$V(G(\lambda) \mid \mathbf{y}) = (DG)^T\tilde{\Sigma}(DG)\{1 + O(k^{-1})\}. \tag{3.6}$$

For our application here it is somewhat preferable to use the multiplicative form because it allows G to decrease to 0 as k becomes infinite, as when $G(\lambda) = V(g(\theta_i) \mid y_i, \lambda)$ and n_i becomes infinite along with k.

These elementary Normality and moment results are proved by Laplace's method, which is described briefly in Section 3.2. The method is sufficiently simple that it is often re-invented and not often referenced by this name. Equations (3.5) and (3.6) require the assumption that $G(\tilde{\lambda})$ is nonzero and the first two derivatives of G at $\tilde{\lambda}$ are of the same order as, or smaller order than, $G(\tilde{\lambda})$. In addition, the posterior must be a probability distribution for some sample size k_0. This is a nontrivial assumption, because the likelihood function may not vanish at the boundary of Λ; in practice, however, a uniform prior on Λ may be considered to be uniform on some compact subset of Λ and 0 elsewhere. We now come to the results of interest.

Result 1: First-Order Expectation Approximation.

$$E(g(\theta_i) \mid \mathbf{y}) = E(g(\theta_i) \mid y_i, \tilde{\lambda})\{1 + O(k^{-1})\}. \tag{3.7}$$

Result 2: First-Order Variance Approximation.

$$V(g(\theta_i) \mid \mathbf{y}) = \left\{V(g(\theta_i) \mid y_i, \tilde{\lambda}) + \sum_{j,h} \tilde{\sigma}_{jh}\tilde{\delta}_j\tilde{\delta}_h\right\} \times \{1 + O(k^{-1})\}, \tag{3.8}$$

where $\tilde{\sigma}_{jh}$ is the (j, h)-component of $\tilde{\Sigma}$ and $\tilde{\delta}_j = (\partial/\partial\lambda_j) E(g(\theta_i) \mid y_i, \lambda)\,|_{\lambda=\tilde{\lambda}}$.

Proofs. Letting $G(\lambda) = E(g(\theta_i) \mid y_i, \lambda)$ we get (3.7) from (3.5) and (2.5). To obtain (3.8), we apply (3.5) to the first term on the right side of (2.6) with $G(\lambda) = V(g(\theta_i) \mid y_i, \lambda)$ and (3.6) to the second term with $G(\lambda) = E(g(\theta_i) \mid y_i, \lambda)$.

Remark 1. As $k \to \infty$, the effect of any given prior $\pi(\lambda)$ on the expectation and variance is of the same order as the terms neglected by the approximations (3.1)–(3.8). This implies that the values of $\tilde{\lambda}$ and $\tilde{\Sigma}$ may be computed using a prior different from $\pi(\lambda)$ without altering the validity of (3.1)–(3.8). When the uniform prior on λ is substituted for $\pi(\lambda)$, $\tilde{\lambda}$ and $\tilde{\Sigma}$ become the maximum likelihood estimator (MLE) and the inverse of observed information (which is the inverse of the negative Hessian of the log-likelihood evaluated at the MLE). In other words, the MLE and inverse of observed information could be substituted for $\tilde{\lambda}$ and $\tilde{\Sigma}$ without altering the order of the approximations in (3.1)–(3.8). By substituting the MLE $\hat{\lambda}$ for the mode $\tilde{\lambda}$ in (3.7), we have a statement of the well-known fact that PEB estimates are also approximate fully Bayesian posterior means.

Remark 2. The statements made in Remark 1 concern asymptotics when a prior is held fixed as $k \to \infty$. In practice, as Example 2 in Section 4 shows, even when approximations (3.7) and (3.8) are fairly accurate for both a uniform prior and an informative prior, the results, which are based on the MLE and the posterior mode, respectively, can be quite different. We comment on this point in Section 5.

Remark 3. When (3.8) is based on the MLE, it is invariant to reparameterization of λ. For the first term this is immediate; for the second term it follows from the transformation properties of observed information and is easily verified. When the mode is used, results for alternative parameterizations will be different, but will agree to the order specified by (3.8).

Formula (3.8) is not only useful computationally, it is also easy to interpret. Note first that the two terms in (2.6) [and thus (3.8)] are of possibly different orders: Assuming that $y_i = (y_{i1}, \ldots, y_{in_i})$ we have

$$E_\lambda\{V(g(\theta_i) \mid y_i, \lambda)\} = O(n_i^{-1}) \tag{3.9a}$$

and

$$V_\lambda\{E(g(\theta_i) \mid y_i, \lambda)\} = O(k^{-1}). \tag{3.9b}$$

The approximation to each term incurs a multiplicative error of order $O(k^{-1})$. From (3.9) we see that when n_i is small and k is large the first term in (3.8) will dominate and, applying Remark 1, the common practice of using the conditional posterior variance at the MLE $\hat{\lambda}$, as an approximation to the true posterior variance, will be appropriate. In this situation, the additional uncertainty due to the estimation of λ (i.e., due to the nonzero spread in the posterior on λ) will be negligible. From the first term of (3.8), knowledge of $g(\theta_i)$ will be more precise for some units than for others according to the conditional posterior precision, which will depend on the individual unit sample sizes n_i. When k is moderate in size or n_i is not small, however, the second term will become important. In this case, the second term of (3.8) is large when the conditional expectation is changing rapidly at $\tilde{\lambda}$ in directions of λ corresponding to substantial posterior uncertainty.

Further interpretation may be obtained when the first stage of the model is a one-parameter exponential family and the second stage is its conjugate prior. Suppose that the random variable Z has density $f(z \mid \theta) = a(z)\exp[z\theta - \psi(\theta)]$, where $\psi(\theta) = \log \int a(z)\exp[z\theta]\, dz$. The con-

jugate prior is a two-parameter exponential family with density $p(\theta \mid \xi, \nu) = \exp[\nu(\xi\theta - \psi(\theta)) - \chi(\xi, \nu)]$, where $\chi(\xi, \nu) = \log \int \exp[\nu(\xi\theta - \psi(\theta))]\, d\theta$. Using the mean value parameter $\mu = E(Z \mid \theta) = \psi'(\theta)$ and applying integration by parts [as in Deely and Lindley (1981)], the prior mean of μ given (ξ, ν) is found to be ξ; by conjugacy, the posterior mean is $E(\mu \mid z, \xi, \nu) = (z + \xi\nu)/(\nu + 1)$.

Now suppose that θ_i are iid according to $p(\theta_i \mid \xi, \nu)$ above, Z_i are independent according to $f(z \mid \theta_i)$, and Y_i is the sum of n_i iid observations from the distribution of Z_i, for $i = 1, \ldots, k$. Setting $\mu_i = E(Z_i \mid \theta_i)$, we obtain the posterior mean

$$E(\mu_i \mid y_i, \xi, \nu) = w_i\xi + (1 - w_i)(y_i/n_i), \quad (3.10)$$

where $w_i = \nu/(\nu + n_i) = n_i^{-1} \cdot (\nu^{-1} + n_i^{-1})^{-1}$. Transforming ν to $\rho = \log \nu$ and putting $g(\theta_i) = \mu_i$ and $\lambda = (\xi, \rho)$, according to the definition of $\tilde{\delta}_j$ immediately after (3.8) we get

$$\tilde{\delta}_1 = \tilde{w}_i \quad (3.11a)$$

and

$$\tilde{\delta}_2 = -\tilde{w}_i(1 - \tilde{w}_i)(y_i/n_i - \tilde{\xi}), \quad (3.11b)$$

where $\tilde{w}_i = \tilde{\nu}/(\tilde{\nu} + n_i)$. These expressions provide additional interpretation of (3.8). Here, the weights $\tilde{w}_i$ are small for the units having large sample sizes n_i. For such units there is little proportional shrinkage, and the contribution to the variance from the second term will remain small as long as the magnitude of the shrinkage $|\tilde{w}_i(y_i/n_i - \tilde{\xi})|$ remains small. On the other hand, for units having large n_i, the contribution of the variance from the second term will be large when y_i/n_i is far from $\tilde{\xi}$. This is quite intuitive: the units that shrink a lot also tend to have greater uncertainty attached to them; this may happen because either the sample sizes are relatively small (which increases the proportion of shrinkage and the first term of the variance) or the distance from the sample mean to the estimated second-stage mean is relatively large (which increases the magnitude of shrinkage and the second term of the variance). The examples in Section 4 illustrate the behavior described here.

3.2 Second-Order Approximations

In this section we briefly review available second-order approximations to the posterior expectation (and variance) of a function $G(\lambda)$, which have error of order $O(k^{-2})$. These may be used to check the accuracy of first-order approximations, as illustrated in Section 4. We also indicate how the first-order approximations of Section 3.1 are derived.

The posterior expectation of $G(\lambda)$ may be written in the form

$$E(G(\lambda) \mid \mathbf{y}) = \frac{\int G(\lambda)L(\lambda)\pi(\lambda)\, d\lambda}{\int L(\lambda)\pi(\lambda)\, d\lambda}, \quad (3.12)$$

and then the numerator and denominator integrals may be evaluated by asymptotic approximation using Laplace's method (e.g., see Erdelyi 1956). In general, if h is a smooth function of an m-dimensional vector λ having a minimum at $\hat{\lambda}$ and b is some other smooth function of λ, then, under suitable regularity conditions, we may expand h and b about $\hat{\lambda}$ to obtain

$$\int b(\lambda)\exp[-kh(\lambda)]\, d\lambda = (2\pi/k)^{m/2}\det(D^2h(\hat{\lambda}))^{-1/2}b(\hat{\lambda})\exp[-kh(\hat{\lambda})]\{1 + O(k^{-1})\} \quad (3.13)$$

as $k \to \infty$, where $D^2h(\hat{\lambda})$ is the Hessian of h at $\hat{\lambda}$, the order $O(k^{-1/2})$ terms having vanished on integration. Applying (3.13) with $h = h_k(\lambda) = -k^{-1} \cdot \log(L(\lambda)\pi(\lambda))$ to both the numerator and denominator of (3.12) yields (3.5). The first-moment approximation (3.1) follows as a special case and, although Laplace's method is often not referenced by name, this is the argument used to derive (3.1). Simple regularity conditions and further references are given by Kass, Tierney, and Kadane (in press, a).

By carrying out higher-order expansions, the first neglected term in (3.13) may be calculated; it involves the first four derivatives of h and the first two of b. Applying the result to (3.12), there is cancellation of terms involving fourth derivatives of h and we obtain

$$E(G(\lambda) \mid \mathbf{y}) = G(\tilde{\lambda}) + \tfrac{1}{2}\sum_{a,b}\tilde{\sigma}_{ab}\left\{G_{ab} - G_a\sum_{c,d}\tilde{\sigma}_{cd}\tilde{l}_{bcd}\right\} + O(k^{-2}), \quad (3.14)$$

where $\tilde{l}(\lambda) = \log(L(\lambda)\pi(\lambda))$ is the log-posterior density, having $\tilde{\lambda}$ as its maximum, $\tilde{\sigma}_{ab} = \tilde{\Sigma}_{ab}$ with $\tilde{\Sigma} = (-D^2\tilde{l}(\tilde{\lambda}))^{-1}$, and the subscripts on G and $\tilde{l}$ indicate partial derivatives evaluated at $\tilde{\lambda}$. If (3.14) is applied to approximate $E(G(\lambda)^2 \mid \mathbf{y})$ and then the square of the approximation to $E(G(\lambda) \mid \mathbf{y})$ is subtracted, we arrive at (3.4) and (3.6). Again, this is the argument used to derive (3.2). If $\tilde{l}$ and the posterior mode $\tilde{\lambda}$ are replaced by the log-likelihood l and the MLE $\hat{\lambda}$ (and the derivatives are evaluated at $\hat{\lambda}$), the term $\Sigma_{a,b}\hat{\sigma}_{ab}G_a\rho_b$, where $\rho(\lambda) = \log(\pi(\lambda))$, must be added to the right side of (3.14), whereas (3.5) and (3.6) remain valid as written. These expansions have been given in this and other contexts by various authors [e.g., Lindley (1961, 1980) and Mosteller and Wallace (1964, sec. 4.6); we note that eq. (3) of Lindley (1980) omits a minus sign—see Tsutakawa (1985) for an application and Kass, Tierney, and Kadane (1988) for additional references].

In addition to its use in justifying the first-order variance approximation (3.2), Equation (3.14) is of direct interest as a second-order approximation to the posterior expectation of $G(\lambda)$. An alternative second-order approximation, due to Tierney and Kadane (1986), applies when G is a positive function. With $\tilde{l}$ and $\tilde{\Sigma}$ defined as in (3.14), letting $l^*(\lambda) = \tilde{l}(\lambda) + \log(G(\lambda))$, with maximum λ^*, and $\Sigma^* = (-D^2l^*(\lambda^*))^{-1}$, we have

$$E(G(\lambda) \mid \mathbf{y}) = \frac{\det(\Sigma^*)^{1/2}\exp(l^*(\lambda^*))}{\det(\tilde{\Sigma})^{1/2}\exp(\tilde{l}(\tilde{\lambda}))}\{1 + O(k^{-2})\}. \quad (3.15)$$

This results from an application of (3.13) to the numerator and denominator of (3.12) once the numerator integrand is put in the fully exponential form $\exp(l^*(\lambda))$; the error is of order $O(k^{-2})$ rather than $O(k^{-1})$ because of cancellation of the order $O(k^{-1})$ factor. A second-order variance approximation may be obtained by approximating $E(G(\lambda)^2)$ along with $E(G(\lambda))$, and then substituting the approximations in the formula $V(G(\lambda)) = E(G(\lambda)^2) - (E(G(\lambda)))^2$. Extensions to nonpositive functions G are discussed in detail by Tierney, Kass, and Kadane (1989). A numerical comparison of alternative expansions was made by Mazzuchi and Soyer (1987), and a PEB application of (3.15) was discussed by Gaver and O'Muircheartaigh (1987).

3.3 Tractable and Intractable Problems

The formulas of Sections 3.1 and 3.2 may be applied without difficulty when (a) the model is conjugate, so each likelihood factor L_i, as defined in (2.3), may be evaluated analytically and (b) the function $G(\lambda)$ may be obtained analytically. When either (a) or (b) fails, however, further techniques are needed, both for the approximations (3.7) and (3.8) and for those of Section 3.2. When (a) fails but the dimensionality of Θ is small, numerical quadrature (together with differentiation under the integral) may be used. These conjugate and nonconjugate situations are illustrated in Examples 1 and 2 of Section 4. In some problems, (a) holds but (b) does not. This is the case in Example 3, where Monte Carlo integration is used in conjunction with first- and second-order approximations.

Finally, when (a) fails and the dimensionality of Θ is not small, the problem becomes more difficult. Maximization of the posterior on λ itself can be problematic [but see Stiratelli, Laird, and Ware (1984) and Racine-Poon (1985) for approximate EM-like methods]. The special structure of CIHM's may still provide some simplification. For instance, in many models λ does not appear in the first stage; models in which a common parameter, such as a scale parameter, does appear in the first stage may often be altered so that the common parameter is eliminated from that stage, as by allowing the scale parameter itself to be unit-dependent according to a further second-stage distribution. In this case, writing $\tilde{l}_{\theta_i}(\lambda) = \log(p(y_i \mid \theta_i)p(\theta_i \mid \lambda))$ and $l_i(\lambda) = \log(L_i(\lambda))$, and differentiating under the integral, we have

$$Dl_i(\lambda) = E(D\tilde{l}_{\theta_i}(\lambda) \mid y_i, \lambda), \tag{3.16}$$

where the expectation is taken with respect to the posterior distribution of θ_i conditional on λ. Similarly,

$$D^2 l_i(\lambda) = E(D^2\tilde{l}_{\theta_i}(\lambda) \mid y_i, \lambda) + V(D\tilde{l}_{\theta_i}(\lambda) \mid y_i, \lambda). \tag{3.17}$$

Computation of the expectations and variance appearing in (3.16) and (3.17) are amenable to asymptotic and numerical methods; thus, although we do not pursue the matter further, we observe that these expressions may be helpful in applications involving intractable CIHM's.

4. EXAMPLES

We begin by introducing an especially simple CIHM, the Normal–Normal model with known first-stage variances. Although the application of this model is often somewhat artificial, it is the starting point of many discussions of "shrinkage" estimation (e.g., Lindley and Smith 1972).

Example 1. Suppose that conditionally on $(\theta_1, \ldots, \theta_k)$, $Y_1, \ldots, Y_k$ are independently Normal(θ_i, σ_i^2) with $\sigma_1, \ldots, \sigma_k$ known, and conditionally on (μ, τ), θ_i are iid Normal(μ, τ^2). This model is a CIHM with $\lambda = (\mu, \tau)$. Letting $g(\theta_i) = \theta_i$ and $w_i = \sigma_i^2(\sigma_i^2 + \tau^2)^{-1}$ we have the familiar expressions $E(\theta_i \mid y_i, \lambda) = w_i\mu + (1 - w_i)y_i = y_i - w_i(y_i - \mu)$ as a version of (3.10). Transforming τ to $\rho = -2 \log \tau$ and setting $\lambda = (\mu, \rho)$ gives $\tilde{\delta}_1 = \tilde{w}_i\tilde{\delta}_2 = -\tilde{w}_i(1 - \tilde{w}_i)(y_i - \tilde{\mu})$ as a version of (3.11), where $\tilde{w}_i = \sigma_i^2(\sigma_i^2 + \tilde{\tau}^2)^{-1}$. Thus we see that the second component of variance can be large for those units for which the weight $\tilde{w}_i$ on the mode $\tilde{\mu}$ of the location hyperparameter is large or the deviation $y_i - \tilde{\mu}$ is substantial. (Although some formulas are more naturally expressed in terms of τ, we have used ρ because it leads to a form of $\tilde{\delta}_2$ that is symmetrical in $\tilde{w}_i$; in addition, its appearance here may help emphasize that ρ is generally preferable to τ in numerical work.)

As a numerical illustration, we consider data from a microbiology experiment in which 13 strains of *E. coli* bacteria were examined for association of two traits. The raw data for each strain were two pairs of sample sizes and corresponding proportions $(n_{i1}, \hat{p}_{i1})$ and $(n_{i2}, \hat{p}_{i2})$, the problem being to compare the underlying proportions p_{i1} and p_{i2} among the 13 strains. Of particular interest was the possibility that for some strains the proportions p_{i1} and p_{i2} might be nearly the same. We assume here that the data are distributed as binomial proportions, we transform to the logit scale according to $Y_i = \log[\hat{p}_{i1}(1 - \hat{p}_{i2})/(\hat{p}_{i2}(1 - \hat{p}_{i1}))]$, we assume Y_i to be Normally distributed, and we take σ_i^2 to be known and equal to the first-order approximate variance based on binomial sampling, $\sigma_i^2 = [n_{i1}\hat{p}_{i1}(1 - \hat{p}_{i1})]^{-1} + [n_{i2}\hat{p}_{i2}(1 - \hat{p}_{i2})]^{-1}$. This is, of course, somewhat crude, but by carrying out the analysis in this form we obtain an illustration of the Normal–Normal CIHM, with known first-stage variances. [The raw data appear in Sklar and Strauss (1980), except for those from strain 11, which were in a prepublication draft but were omitted from the published paper; a paired-binomial or paired-Poisson CIHM could be used instead, though we doubt that either would lead to substantially different conclusions.]

The transformed data and the first-order results are shown in Table 1. Here we use the mode $(\tilde{\mu}, \tilde{\rho})$ resulting from a uniform prior on (μ, ρ). The two components of the approximate variance are given in the last two columns of the table. We wish to emphasize here the importance of the second component of the approximate variance, which is often ignored; in many strains it is substantial compared with the first. In strain 10, the second component V2 is much larger than the first component V1. This may be

Table 1. Approximations for E. coli Data

Strain	y_i	σ_i	$\tilde{\theta}$	SD	V1	V2
1	1.36	.28	1.35	2.63	.07	.002
2	2.26	1.04	1.56	.621	.32	.07
3	2.23	.75	1.68	.568	.25	.08
4	1.32	.36	1.31	.319	.10	.003
5	1.21	.38	1.24	.339	.11	.004
6	1.27	.49	1.28	.403	.16	.007
7	1.43	.57	1.37	.441	.18	.01
8	1.85	.54	1.62	.453	.18	.03
9	1.34	.56	1.32	.444	.19	.01
10	3.44	.73	2.20	.738	.24	.30
11	−.42	.69	.53	.642	.23	.18
12	−.10	.31	.17	.354	.08	.05
13	1.25	.39	1.27	.342	.11	.004

NOTE: $\tilde{\theta}$ and SD are the approximate means and standard deviations given by (3.7) and (3.8). V1 and V2 are the two terms in (3.8). The posterior mode of (μ, ρ) based on the uniform prior is (1.30, −.47).

understood as coming from the large deviation of y_{10} away from the modal value $\tilde{\mu} = 1.30$; the shrinkage is substantial and so is the resulting uncertainty. In contrast, strain 2 has a much larger first component than strain 10, but a much smaller second component (the observation being less extreme), leading to a smaller standard deviation (SD) when the two components are combined. Similarly, the ninth and eleventh strains may be compared: the two first variance components are not drastically different, but the two second components are; there is much uncertainty associated with strain 11, which is greatly shrunk toward $\tilde{\mu}$. From a substantive point of view, strain 12 had a raw difference of logits greater than that of strain 11. A posteriori, however, the difference of logits is concentrated near 0 for strain 12, whereas that for strain 11 is less concentrated and its location has shrunk to a larger positive value.

Remark. In the homogeneous case of this Normal-Normal CIHM, in which $\sigma_1 = \cdots = \sigma_k = \sigma$, when a uniform prior is used on (μ, ρ), (3.8) produces $V(\theta_i \mid \mathbf{y}) = \{\hat{w}(1 - \hat{w})s^2 + \hat{w}\sigma^4/k + (2\sigma^4/k)(y_i - \bar{y})^2\}\{1 + O(k^{-1})\}$, where $s^2 = (k-1)^{-1}\sum_{i=1}^{k}(y_i - \bar{y})^2$ and $\hat{w} = \sigma^2(\sigma^2 + \hat{\tau}^2)^{-1}$, with $\hat{\tau}^2 = \exp(-\hat{\rho})$ being the modal value, that is, the second component of the MLE of (μ, τ^2). This approximate variance turns out to be the exact posterior variance based on a uniform prior on (μ, τ^2) (e.g., compare Morris 1986).

Example 2. Another simple case is that of dichotomous data, in which Y_i given θ_i is binomial$(n_i; \theta_i)$. As the second-stage distribution we consider first a conjugate beta$(v\xi, v(1-\xi))$ distribution and then, as in Leonard (1972), a Normal(μ, τ^2) distribution on $\eta_i = \log[\theta_i/(1-\theta_i)]$. In the latter case we consider both a flat prior on the parameters (μ, ω), where $\omega = \log\tau$, and an "informative" proper prior. In each case we take $g(\theta_i) = \theta_i$. For numerical illustration we use the last 10 cities of the toxoplasmosis data set analyzed by Efron (1986). (Although the qualitative features of the approximations are similar when the full data set is used, the results are more striking when the sample size is reduced by a factor of 3.) The data (y_i, n_i) are number of subjects tested to be positive for toxoplasmosis and number of subjects tested in each of many (here, 10) cities in El Salvador.

We begin with the conjugate beta second stage. With the (ξ, v) parameterization, (3.10) and (3.11) apply. The data and results based on a uniform prior on (ξ, ρ), where $\rho = \log(v)$, are given in Table 2. We again emphasize the importance of the second component of the approximate variance, and we also draw attention to comparisons between cities: the sixth has a larger sample size than the fifth, yet there is much more variability in its posterior because it has shrunk much farther. Similarly, the first variance components for the seventh and eighth cities are comparable, but the second component is much larger for the eighth city.

As a quick check on accuracy, (3.15) may be applied with $G(\lambda) = E(\theta_i \mid y_i, \lambda)$. The results are presented as $\hat{E}(\theta_i \mid \mathbf{y})$ in Table 2. In this case, first-order and second-order values are quite close.

In the nonconjugate logit-Normal model, the quantities entering (3.7) and (3.8) cannot be evaluated analytically. In this case, as discussed in Section 3.3, results may be obtained using one-dimensional numerical quadrature over Θ. When we use the nonconjugate Normal second stage together with a uniform prior on $\lambda = (\mu, \omega)$, the approximate posterior means given by (3.7) agree to three digits with those obtained from the conjugate model, shown in Table 2. The variances given by (3.8) agree with those of Table 2 to the accuracy given in that table except for three cities (the largest discrepancy being .0038 and .0055 for the variance terms for the eighth city, as opposed to .0037 and .0053 obtained for the conjugate model). In this example, with a tight second-stage distribution centered near $E(\theta_i \mid \lambda) = .5$, the conjugate and logit-Normal models are very similar. Thus the results and interpretation are essentially unchanged.

In contrast, if a more informative prior is specified for $\lambda = (\mu, \omega)$, the results can be dramatically different. To illustrate, we leave the prior on μ flat but take ω to be Normally distributed with mean .8035 and standard deviation .797. We chose this prior in a rather arbitrary fashion: we set $\mu = \mu_0 = 0$ and determined two values of ω_0 by setting $\Pr\{.05 < \theta_i < .95 \mid \mu = \mu_0, \omega = \omega_0\}$ equal to

Table 2. Approximations for Toxoplasmosis Data Using the Conjugate Model

City	y_i	n_i	$\hat{p}_i$	$\tilde{\theta}_i$	$\hat{E}(\theta_i \mid \mathbf{y})$	SD	V1	V2
1	24	51	.47	.515	.531	.067	23	22
2	7	16	.44	.530	.535	.078	34	27
3	46	82	.56	.559	.562	.045	18	2
4	9	13	.69	.582	.584	.073	35	18
5	23	43	.54	.547	.553	.056	25	6
6	53	75	.71	.642	.625	.076	17	41
7	8	13	.61	.567	.570	.067	35	10
8	3	10	.30	.517	.520	.095	37	53
9	1	6	.17	.518	.518	.098	39	57
10	23	37	.62	.582	.580	.060	26	10

NOTE: $\hat{p}_i$ is y_i/n_i, $\tilde{\theta}_i$ and $\hat{E}(\theta_i \mid \mathbf{y})$ are the first-order and second-order approximate means given by (3.7) and (3.15), SD is the first-order approximate standard deviation given by (3.8), and V1 and V2 are 10^4 times the first and second variance terms in (3.8). The posterior mode of $\lambda = (\xi, \rho)$ based on the uniform prior was found to be (.556, 4.03), with first-order asymptotic standard deviations (.038, .13) and correlation .31.

.5 and .99. We then interpreted the corresponding values of ω_0 (1.473 and .134, respectively) as the 80th and 20th percentiles of the Normal prior on ω.

Under this Normal prior, the posterior mode is $(\tilde{\mu}, \exp(\tilde{\omega})) = (.18, .51)$, compared with (.55, .27) obtained under the uniform prior; in the probability scale, $\exp(\tilde{\mu})/(1 + \exp(\tilde{\mu})) = .55$, compared with .63. The approximate (modal) standard deviations for $(\mu, \exp(\omega))$ are (.21, .40) with a correlation of $-.18$. The comparable quantities under the flat prior are (.16, .64) with a correlation of $-.28$.

First-stage approximations using this more informative prior are presented in Table 3. Because of the larger value of $\tilde{\omega}$ there is much less shrinkage among the posterior means [toward a smaller value, $\exp(\tilde{\mu}) = .55$]. Correspondingly, the second component of the variance diminishes while the first increases; thus the second component contributes less than before. It is still consequential in some cases: For city 9 the second component is 84% of the first (yielding an increase of 36% in the standard deviation). Second-order expectation approximations using (3.15) are given as $\hat{E}(\theta_i \mid \mathbf{y})$.

We checked the accuracy of these approximations in both the conjugate flat prior and the nonconjugate informative prior models using (respectively) Monte Carlo simulation and Gauss–Hermite quadrature (Naylor and Smith 1982) over Λ. In both the beta-binomial and logit-Normal models, the second-order approximations given by (3.15) and the exact posterior means were identical to three digits for 6 of the 10 cities; for the other 4 cities, the values differed by no more than one unit in the third digit. In the nonconjugate model, the first-order variance approximations given by (3.8) and exact posterior variances differed by at most four units in the second digit—for example, for city 6 the approximate variance was .0031 compared with an exact (quadrature) value of .0027. In terms of standard deviation, the approximate value for city 6 was 7% larger than the exact value.

Example 3. In hierarchical models involving multinomial data $Y_i = (Y_{i1}, \ldots, Y_{iq})$, the conjugate model entails specifying a Dirichlet(ν, ξ) prior for the first-stage parameter vector $\theta_i = E(Y_i)/n_i$, where $n_i = \sum_{j=1}^{q} Y_{ij}$. Here, the parameterization is such that $\xi = (\xi_1, \ldots, \xi_{q-1})$, $\xi_q = 1 - \xi_1 - \cdots - \xi_{q-1}$, and $E(\theta_{ij} \mid \nu, \xi) = \xi_j$. Under this model, $\lambda = (\nu, \xi)$ and if we take $G(\lambda)$ to be a component of $E(\theta_i \mid y_i, \lambda)$, the mean vector of a Dirichlet random vector, then an analytic result is available—namely, $E(\theta_{ij} \mid y_i, \lambda) = (\nu\xi_j + y_{ij})/(\nu + n_i)$ for $j = 1, \ldots, q$. If instead, as in the case of a 2×2 table with $q = 4$, we are interested in the log odds $g(\theta_i) = \log[(\theta_{i1}\theta_{i4})/(\theta_{i2}\theta_{i3})]$ and take $G(\lambda)$ to be some functional of its distribution, then $G(\lambda)$ may not have an analytic expression. For illustration, we consider the data from Beitler and Landis (1985), reproduced here in Table 4, which were originally obtained from a multicenter randomized clinical trial investigating the efficacy of two topical cream preparations (active drug, control) in curing an infection. As shown in Table 4, in seven of eight clinics the drug produced a higher proportion of favorable responses, but there exists substantial variation in the overall rate of favorable response (combining both drug and control groups), ranging from 4.8% to 76.9% across the clinics. We focus here on clinic 8, which is unusual in that it has the smallest sample size, and it is the only clinic for which the observed proportion of favorable responses is higher for the control group than for the treatment group.

Suppose that we are interested in the posterior probability that the treatment will be ineffective for clinic 8. Let $\gamma(\theta_8) = I\{g(\theta_8) < 0\}$, where $I(\cdot)$ is the indicator function, and assume a uniform prior on $(\log \nu, \xi)$. Then, the desired probability can be expressed as the posterior expectation of $G(\lambda)$, where $G(\lambda) = E(\gamma(\theta_8) \mid y_8, \lambda)$. Letting $\hat{\lambda}$ denote the MLE (the mode) of $\lambda = (\nu, \xi)$, a first-order approximation is given by

$$G(\hat{\lambda}) = \int \gamma(\theta_8) p(\theta_8 \mid y_8, \hat{\lambda})\, d\theta_8. \tag{4.1}$$

In conducting the data analysis, we obtained the following MLE's for the hyperparameters: $\hat{\nu} = 9.326$, $\hat{\xi}_1 = .205$, $\hat{\xi}_2 = .135$, and $\hat{\xi}_3 = .288$. The approximate density $p(\theta_i \mid y_i, \hat{\lambda})$ is Dirichlet with parameters $\nu^* = \hat{\nu} + n_i$ and $\xi_j^* = (\hat{\nu}\hat{\xi}_j + y_{ij})/(\hat{\nu} + n_i)$. Hence the integral in (4.1) can be computed by a Monte Carlo method in which Dirichlet observations are generated and the fraction yielding neg-

Table 3. Approximations for Toxoplasmosis Data Using a Nonconjugate Model and an Informative Prior

City	$\hat{p}_i$	$\tilde{\theta}_i$	$\hat{E}(\theta_i \mid \mathbf{y})$	SD	V1	V2
1	.47	.488	.489	.064	37	4
2	.44	.490	.486	.097	76	18
3	.56	.558	.559	.051	25	.7
4	.69	.609	.616	.099	77	21
5	.54	.537	.538	.066	42	2
6	.71	.677	.677	.056	23	7
7	.61	.575	.579	.095	80	10
8	.30	.450	.437	.12	91	56
9	.17	.441	.421	.14	105	88
10	.62	.598	.599	.070	44	5

NOTE: Headings for the second through sixth columns are defined in the note to Table 2.

Table 4. Distribution of Favorable Response to Active Drug and Control Treatments From a Multicenter Randomized Clinical Trial

Clinic	Treatment	Response: Favorable	Response: Unfavorable	Total	Proportion favorable
1	Drug	11	25	36	.306
	Control	10	27	37	.270
2	Drug	16	4	20	.800
	Control	22	10	32	.688
3	Drug	14	5	19	.737
	Control	7	12	19	.368
4	Drug	2	14	16	.125
	Control	1	16	17	.059
5	Drug	6	11	17	.353
	Control	0	12	12	.000
6	Drug	1	10	11	.091
	Control	0	10	10	.000
7	Drug	1	4	5	.200
	Control	1	8	9	.111
8	Drug	4	2	6	.667
	Control	6	1	7	.857

ative log odds are tallied. Based on a simulation of 10,000 Dirichlet observations, we found $\Pr\{g(\theta_8) < 0 \mid y_8, \hat{\lambda}\} = .625(\pm.0048)$. The $\pm$ figure is the Monte Carlo standard deviation of the estimated probability computed as $.0048 = [(.625)(.375)/10{,}000]^{1/2}$.

The approximation (4.1) can be refined by using a second-order approximation $\hat{p}_2(\cdot \mid \mathbf{y})$ to the posterior density $p(\theta_8 \mid \mathbf{y})$. With the first-order density approximation serving as an importance function for Monte Carlo integration, the second-order approximate probability can be written as

$$\Pr\{g(\theta_8) < 0 \mid \mathbf{y}\} \doteq \int \gamma(\theta_8) \left[\frac{\hat{p}_2(\theta_8 \mid \mathbf{y})}{p(\theta_8 \mid y_8, \hat{\lambda})}\right] p(\theta_8 \mid y_8, \hat{\lambda})\, d\theta_8. \quad (4.2)$$

In this case, we have implemented the MLE version of the second-order approximation (3.14) with $G(\lambda) = p(\theta_8 \mid y_i, \lambda)$. Under the conjugate specification, analytic expressions can be derived for the second and third derivatives of the log-likelihood $l(\lambda)$, as well as for $G(\lambda)$ and its first derivatives. Numerical differentiation is used to obtain the second derivatives of $G(\lambda)$, which involve the second logarithmic derivative of the gamma function. The computation of (4.2) is then accomplished by Monte Carlo integration, the only difference from the computation of (4.1) being that the ratio of second-order to first-order approximate densities serves as a weight function. Using a sample size of 20,000 (so that the standard deviations of the estimated probabilities are comparable), we found the second-order approximate probability to be $.648(\pm.0053)$, reflecting a change in the second significant digit.

Example 4. We now consider the generalization of Example 1 to the class of general linear models in which there are k individual units and, for the ith unit, $Y_i = X_i\alpha + Z_i b_i + e_i$. In the notation and terminology of Harville (1977) and Laird and Ware (1982), Y_i is an $n_i \times 1$ vector of responses, α is a $p \times 1$ vector of unknown population parameters, and X_i is a known $n_i \times p$ matrix linking α to Y_i. In addition, b_i is a $q \times 1$ vector of unknown individual effects, Z_i is a known $n_i \times q$ matrix linking b_i to Y_i, and e_i is an $n_i \times 1$ vector of random errors. Models of this form are widely used in the analysis of longitudinal data. Analysis of panel data from economic surveys was reviewed by Johnson (1977, 1980) and Dielman (1983). Repeated-measures applications of the kind that occur frequently in prospective studies of human subjects were discussed by Laird and Ware (1982), following Dempster, Rubin, and Tsutakawa (1981). Because of the importance of these models, we present here the expressions needed in computing approximate posterior variances according to (3.8). Further details may be found in Kass and Steffey (1986).

In our terminology, α is a vector of common parameters and b_i is a vector of unit-specific parameters for the ith individual unit. This "mixed effects" model can be formulated as a conditionally independent hierarchical model. At the first stage, we regard α and b_i as fixed, and the e_i are assumed to be independent and normally distributed as $N(0, R_i)$, where R_i is an $n_i \times n_i$ positive-definite covariance matrix. Although R_i depends on i through its dimension n_i, the set of unknown parameters in R_i will not depend on i. At the second stage, we take the b_i to be distributed as $N(0, D)$, independently of each other and of the e_i, where D is a $q \times q$ positive-definite covariance matrix. Let ζ denote the vector of variance and covariance parameters found in R_i $(i = 1, \ldots, k)$ and D.

The posterior distribution of b_i given α and ζ is Normal with $E(b_i \mid y_i, \alpha, \zeta) = DZ_i^T V_i^{-1}(y_i - X_i\alpha)$ and $V(b_i \mid y_i, \alpha, \zeta) = (D^{-1} + Z_i^T R_i^{-1} Z_i)^{-1}$, where $V_i = R_i + Z_i D Z_i^T$. Substituting the posterior mode $(\tilde{\alpha}, \tilde{\zeta})$ for (α, ζ) in the foregoing expressions yields, respectively, the approximate posterior mean and the first term of the approximate posterior variance. The second term in the variance approximation, which accounts for the uncertainty in estimating α and ζ, requires the second-order partial derivatives of the log-likelihood function and the first-order partial derivatives of $G(\alpha, \zeta) = E(b_i \mid y_i, \alpha, \zeta)$, both evaluated at $(\tilde{\alpha}, \tilde{\zeta})$. First- and second-order derivatives of the log-likelihood may be found in Kass and Steffey (1986) and are analogous to expressions given by Harville (1977). We find

$$\partial G/\partial \alpha_j = (-DZ_i^T V_i^{-1} X_i)_j \quad (4.3)$$

and

$$\partial G/\partial \zeta_j = [(\partial D/\partial \zeta_j)Z_i^T - DZ_i^T V_i^{-1}(\partial V_i/\partial \zeta_j)]V_i^{-1} \times (y_i - X_i\alpha). \quad (4.4)$$

Computation of the negative Hessian of the log prior and substitution of $\tilde{\alpha}$ and $\tilde{\zeta}$ for α and ζ in (4.3) and (4.4) then provides the appropriate terms in (3.8).

An important simplification occurs when the prior on α is uniform and $\pi(\alpha, \zeta) = p(\zeta)$, so integration over α may be performed analytically. In this case, the integrated likelihood function $L_1(\zeta) = \int L(\alpha, \zeta)\, d\alpha$ plays a role analogous to the full likelihood $L(\alpha, \zeta)$ in the general case and, conditionally on ζ, the posterior expectation and variance of b_i may be computed in closed form. Putting $G(\zeta) = E(b_i \mid y_i, \zeta)$, we obtain

$$\partial G/\partial \zeta_j = [(\partial D/\partial \zeta_j)Z_i^T - DZ_i^T P_i(\partial V_i/\partial \zeta_j)]P_i y_i, \quad (4.5)$$

where $P_i = V_i^{-1} - V_i^{-1}X_i(\sum_{i=1}^k X_i^T V_i^{-1} X_i)^{-1} X_i^T V_i^{-1}$. Approximations to the posterior mean and variance based on (3.7) and (3.8) now follow, again, by computing the Hessian of $\log(L_1)$ and of the log prior and replacing ζ in (4.5) with the mode of $L_1 \cdot \pi$. Details are given in Kass and Steffey (1986). We note that the maximum of L_1 is the restricted maximum likelihood estimate introduced by Patterson and Thompson (1971) [see Dempster, Rubin, and Tsutakawa (1981) for discussion].

In the foregoing expressions we have not given explicit forms for the derivatives of D and V_i because these matrices might be assumed to have a special structure. For instance, when R_i is a multiple of the identity we obtain

$\partial V_i/\partial \zeta_{jm} = Z_{i(j)} Z_{i(m)}^T$, where ζ_{jm} is the (j, m)th element of the covariance matrix D and $Z_{i(j)}$ denotes the jth column of Z_i. Another important special case occurs when, in addition, D is assumed to be diagonal. An approximate Bayesian analysis of this case was given by Broemeling (1985, chap. 4).

5. DISCUSSION

Deely and Lindley (1981) pointed out that approximate Bayesian methods may be used with models employed in PEB technology. We have elaborated on that theme by introducing an alternative terminology, deriving and interpreting the first-order approximation to the posterior variance of the parameter estimated with the PEB methods, and discussing second-order approximations. The accuracy of the variance approximation depends on the number k of units, rather than the number n_i of observations on a particular unit. Although when k becomes sufficiently large, with n_i remaining small, the first term in (3.8) will suffice, we have found in several examples that it is advisable to compute both terms. We have proceeded heuristically. Rigorous justification for expansions based on Laplace's method may be found in Kass et al. (in press, a). We now add some comments on the role of approximate Bayesian methods in analysis with CIHM's.

Although we chose to illustrate the methodology we discussed with small data sets and simple models, part of our purpose has been to develop tools that may be used with large data sets and more elaborate models. In particular, we believe that approximate inference in CIHM's can be effective for analyzing data that are collected in longitudinal studies. The methodology we have described should complement existing technology, as used, for instance, by Waternaux, Laird, and Ware (1989), and allow generalizations beyond those already available. We refer here not only to the first-order variance approximations, but also to second-order asymptotic expressions: Longitudinal data usually include covariates and, even with only a few, we quickly reach a sufficiently large number of common parameters that a substantial computational effort may be required to produce "exact" posteriors.

Part of our emphasis on first-order asymptotics is due to the great simplification provided when the posterior is approximately Normal, with mean and variance given by first-order approximations. A priority in data analysis should be to check these approximations. Second-order asymptotics can be useful for this purpose, in that agreement of first- and second-order results at least provides some reassurance, and disagreement indicates inaccuracy. See Kass et al. (in press, b) for further development of methodology in the context of single-stage models. It remains advisable, whenever possible, to check at least some features of the posterior using nonasymptotic numerical techniques. For references and discussion, see Shaw (1988) and Kass et al. (1988) (and the discussion to these papers).

Asymptotic analysis provides motivation for approximations and helps explain accuracy. Order of accuracy, however, must be interpreted with care. For instance, one might consider the approximation to the posterior distribution of $g(\theta)$ given by the conditional posterior distribution, conditional on $\lambda = \tilde{\lambda}$. This would seem to furnish a valid first-order approximation, in the sense that $\Pr\{g(\theta_i) < c \mid \mathbf{y}\}$ may be approximated by $\Pr\{g(\theta_i) < c \mid y_i, \tilde{\lambda}\}$, with multiplicative error of order $O(k^{-1})$. Yet, use of the conditional posterior entails use of the first term of the variance approximation (3.8) without the second term and, as we have seen in the examples, this can seriously understate the posterior variance. Thus the seemingly appropriate order $O(k^{-1})$ approximation to the distribution function may not be of much inferential use; furthermore, this inaccuracy may occur even in cases for which the order $O(k^{-1})$ approximation to the variance given by (3.8) is quite adequate.

In addition, we have seen in Example 2 the important effects of the prior distribution of λ. On the one hand, the variance was shown to be well approximated by (3.8) in that case. Yet, on the other hand, in Remark 1 following (3.8) we noted that, to the order of the approximation, the prior is irrelevant. This indicates to us that the asymptotic argument we gave was not appropriate for treating the informative prior case. That argument assumes that the prior precision is of constant order and the observed information $-D^2 l(\hat{\lambda})$ is of order $O(k)$; when these precisions are instead of the same order of magnitude, that assumption will not be satisfied. An alternative analysis assumes instead that the prior precision is also of order $O(k)$ [formally, by assuming that the prior contains a scale parameter that decreases at the rate $O(k^{-1/2})$]. Laplace's method may be applied as in (3.13), and all results go through, but it is now no longer true that the prior may be omitted from the function being maximized in (3.13); therefore, the mode may no longer be replaced by the MLE. The practical observation, then, is that there is sometimes a serious discrepancy between results obtained using the mode and those obtained using the MLE and, when there is, the mode is likely to be preferable.

Example 2 also furnishes a reminder that the estimate of second-stage precision controls shrinkage, and that this may be strongly affected by the prior on Λ, especially when information about inter-unit variability is weak (see also Hill 1965, 1977). Here we proceeded by reparameterizing and putting a Normal prior on ρ, with the assumption that the posterior would be approximately Normal on (μ, ρ). An alternative would be to develop approximations that are appropriate for the variance components directly [proceeding, for instance, along the lines developed by Box and Tiao (1973, chap. 5)]. This is an important topic for further research.

Finally, a first-stage distribution is often selected for some fairly good reason, whereas the second stage is chosen purely for analytical convenience. In other situations there may be several plausible candidates for the second-stage model. With regard to Example 3 of this article, some authors (Fienberg and Holland 1973; Sutherland, Holland, and Fienberg 1974) elected to use the Dirichlet second stage, and others (Laird 1978; Leonard 1975; Leonard and Novick 1986) used Normal distributions on log-

linear model parameters. Sensitivity and robustness considerations need special attention in such instances.

[*Received December 1986. Revised February 1989.*]

REFERENCES

Beitler, P. J., and Landis, J. R. (1985), "A Mixed-Effects Model for Categorical Data," *Biometrics,* 41, 991–1000.

Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis,* Reading, MA: Addison-Wesley.

Broemeling, L. D. (1985), *Bayesian Analysis of Linear Models,* New York: Marcel Dekker.

Deely, J. J., and Lindley, D. V. (1981), "Bayes Empirical Bayes," *Journal of the American Statistical Association,* 76, 833–841.

Dempster, A. P., Rubin, D. B., and Tsutakawa, R. K. (1981), "Estimation in Covariance Components Models," *Journal of the American Statistical Association,* 76, 341–353.

Dielman, T. E. (1983), "Pooled Cross-Sectional and Time Series Data: A Survey of Current Statistical Methodology," *The American Statistician,* 37, 111–123.

Efron, B. (1986), "Double Exponential Families and Their Use in Generalized Linear Regression," *Journal of the American Statistical Association,* 81, 709–721.

Efron, B., and Morris, C. (1973), "Stein's Estimation Rule and Its Competitors—An Empirical Bayes Approach," *Journal of the American Statistical Association,* 68, 117–130.

——— (1975), "Data Analysis Using Stein's Estimator and Its Generalizations," *Journal of the American Statistical Association,* 70, 311–319.

Erdelyi, A. (1956), *Asymptotic Expansions,* New York: Dover Publications.

Fienberg, S. E., and Holland, P. W. (1973), "Simultaneous Estimation of Multinomial Cell Probabilities," *Journal of the American Statistical Association,* 68, 683–691.

Gaver, D. P., and O'Muircheartaigh, I. G. (1987), "Robust Empirical Bayes Analysis of Event Rates," *Technometrics,* 29, 1–15.

Good, I. J. (1980), "Some History of the Hierarchical Bayesian Methodology," in *Bayesian Statistics,* eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Valencia, Spain: University Press, pp. 489–519.

Harville, D. A. (1977), "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," *Journal of the American Statistical Association,* 72, 320–340.

Hill, B. (1965), "Inferences About Variance Components in the One-Way Model," *Journal of the American Statistical Association,* 60, 806–825.

——— (1977), "Exact and Approximate Bayesian Solutions for Inference About Variance Components and Multivariate Inadmissibility," in *New Developments in the Application of Bayesian Methods,* eds. A. Aykac and C. Brumat, Amsterdam: North-Holland, pp. 129–152.

Johnson, L. W. (1977), "Stochastic Parameter Regression: An Annotated Bibliography," *International Statistical Review,* 45, 257–272.

——— (1980), "Stochastic Parameter Regression: An Additional Annotated Bibliography," *International Statistical Review,* 45, 95–102.

Kass, R. E., and Steffey, D. (1986), "Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models)," Technical Report 386, Carnegie-Mellon University, Dept. of Statistics.

Kass, R. E., Tierney, L., and Kadane, J. B. (1988), "Asymptotics in Bayesian Computation," in *Bayesian Statistics 3,* eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 261–278.

——— (in press, a), "The Validity of Posterior Expansions Based on Laplace's Method," in *Essays in Honor of George Barnard,* eds. S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner, Amsterdam: North-Holland.

——— (in press, b), "Approximate Methods for Assessing Influence and Sensitivity in Bayesian Analysis," *Biometrika,* 76.

Laird, N. M. (1978), "Empirical Bayes Methods for Two-Way Contingency Tables," *Biometrika,* 65, 581–590.

Laird, N. M., and Louis, T. A. (1987), "Empirical Bayes Intervals Based on Bootstrap Samples," *Journal of the American Statistical Association,* 82, 739–750.

Laird, N. M., and Ware, J. H. (1982), "Random-Effects Models for Longitudinal Data," *Biometrics,* 38, 963–974.

Leonard, T. (1972), "Bayesian Methods for Binomial Data," *Biometrika,* 59, 581–589.

——— (1975), "Bayesian Estimation Methods for Two-Way Contingency Tables," *Journal of the Royal Statistical Society,* Ser. B, 37, 23–37.

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

Lindley, D. V. (1961), "The Use of Prior Probability Distributions in Statistical Inference and Decisions," in *Proceedings of the Fourth Berkeley Symposium* (Vol. 1), Berkeley: University of California Press, pp. 452–468.

——— (1971), "The Estimation of Many Parameters" (with discussion), in *Foundations of Statistical Inference,* eds. V. P. Godambe and D. A. Sprott, Toronto: Holt, Rinehart & Winston, pp. 435–455.

——— (1980), "Approximate Bayesian Methods," in *Bayesian Statistics,* eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Valencia, Spain: University Press, pp. 223–237.

Lindley, D. V., and Smith, A. F. M. (1972), "Bayes Estimates for the Linear Model" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 34, 1–41.

Mazzuchi, T. A., and Soyer, R. (1987), "Software Reliability Assessment Using Posterior Approximations," in *Computer Science and Statistics: Proceedings of the Nineteenth Symposium on the Interface,* Washington, DC: American Statistical Association, pp. 400–405.

Morris, C. (1983), "Parametric Empirical Bayes Inference: Theory and Applications" (with discussion), *Journal of the American Statistical Association,* 78, 47–65.

——— (1986), Comment on "Why Isn't Everyone a Bayesian?" by B. Efron, *The American Statistician,* 40, 7–8.

Mosteller, F., and Wallace, D. L. (1964), *Inference and Disputed Authorship: The Federalist,* Reading, MA: Addison-Wesley.

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.

Patterson, H. D., and Thompson, R. (1971), "Recovery of Inter-Block Information When Block Sizes Are Unequal," *Biometrika,* 58, 545–554.

Racine-Poon, A. (1985), "A Bayesian Approach to Nonlinear Random Effects Models," *Biometrics,* 41, 1015–1023.

Robbins, H. (1964), "The Empirical Bayes Approach to Statistical Decision Problems," *The Annals of Mathematical Statistics,* 35, 1–20.

Shaw, J. E. H. (1988), "Aspects of Numerical Integration and Summarisation," in *Bayesian Statistics 3,* eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 411–428.

Sklar, R., and Strauss, B. (1980), "Role of the *uvr E* Gene Product and of Inducible 0*-Methylguanine Removal in the Induction of Mutations by *N*-Methyl-*N'*-Nitro-*N*-Nitrosoguanidine in *Escherichia coli,*" *Journal of Molecular Biology,* 143, 345–363.

Stiratelli, R., Laird, N., and Ware, J. H. (1984), "Random-Effects Models for Serial Observations With Binary Response," *Biometrics,* 40, 961–971.

Sutherland, M., Holland, P. W., and Fienberg, S. E. (1974), "Combining Bayes and Frequency Approaches to Estimate a Multinomial Parameter," in *Studies in Bayesian Econometrics and Statistics,* eds. S. E. Fienberg and A. Zellner, Amsterdam: North-Holland, pp. 585–617.

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. (1989), "Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions," *Journal of the American Statistical Association,* 84, 710–716.

Tsutakawa, R. K. (1985), "Estimation of Cancer Mortality Rates: A Bayesian Analysis of Small Frequencies," *Biometrics,* 41, 69–80.

Waternaux, C., Laird, N. M., and Ware, J. H. (1989), "Methods for Analysis of Longitudinal Data: Blood-Lead Concentrations and Cognitive Development," *Journal of the American Statistical Association,* 84, 33–41.