Journal of Official Statistics
Vol. 6. No. 2, 1990, pp. 133–155

A Bayesian Loglinear Model Analysis of Categorical Data

Robert M. Leighty[1] and William J. Johnson[2]

Abstract: We illustrate a two-staged Bayesian strategy for making inferences about the loglinear model parameters that summarize the interaction structure in a multidimensional contingency table. In the first stage, we locate full and reduced uniformly ordered models whose parameter vectors enclose all important parameters. In the second-stage, posterior regions are used to identify important loglinear model terms. Parameters in these terms are estimated by Bayesian posterior means that compromise full and reduced model maximum likelihood estimates (MLE's). The likelihood is summarized by the approximate normal distribution of the sufficient full-model MLE. A hierarchical normal-Cauchy-tail prior, centered at the reduced-model MLE, is assumed. A relative precision hyperparameter measures our belief in the reduced model. We illustrate three methods of approximating the posterior moments: the Laplace method, an empirical Bayes method, and the diagonalized covariance method. Our Bayesian strategy is then used to reanalyze the Ries-Smith detergent preference data.

Key words: Hierarchical Bayes; Laplace approximation; credible regions.

1. Introduction

A principal objective in categorical data analysis is to summarize the interaction structure in a contingency table with a set of parameters. Assume that the contingency table possesses t cell frequencies, n_{ijk}, that have been generated by an independent Poisson, multinomial, or product-multinomial sampling model. Let m_{ijk} represent the corresponding expected cell means.

We choose to summarize interaction with the parameters of the loglinear model because of the desirable margin-free property of these parameters (Altham 1970a, 1970b). For government and social science studies that focus on interactions involving ethnicity, this margin-free property is especially important. In these studies, margin-dependent parameters, which are characteristic of most models, may be misleading. For instance, a margin-dependent parameter measuring asssociation for a minority will be smaller than an interaction parameter measuring association for a nonminority, merely because of the inequity in the size of minority and nonminority populations. To avoid this unappealing dependence on the marginal levels of ethnicity, we use the loglinear

[1] Assistant Professor of Mathematical Sciences, Denison University, Granville, Ohio 43023, U.S.A.
[2] Design Engineer for Indiana Bell, Crown Point, Indiana, 46307 U.S.A.

Acknowledgement: The authors wish to thank Professor Rob Kass for suggesting revisions to parts of the manuscript. Professor Leonard's tutorial was inspirational. We also gratefully acknowledge an associate editor and the referees for their constructive criticisms, which led to a much improved manuscript.

model

$$\boldsymbol{F}(\boldsymbol{m}) = X\boldsymbol{\beta} \tag{1.1}$$

Here, $\boldsymbol{F}(\boldsymbol{m})$ is the vector of the logarithm of expected means, with components $\log m_{ijk}$, and $\boldsymbol{\beta}$ is the p-dimensional ($p \leqslant t$) parameter vector and X is the full-rank design matrix associated with a full model. The full model will seldom coincide with the saturated model, but rather, shall be a hierarchical model with more terms than required to provide an adequate fit to the data.

We assume that the likelihood is adequately approximated by the asymptotic normal distribution of the maximum likelihood estimator $\boldsymbol{b}$, which is sufficient for $\boldsymbol{\beta}$ in model (1.1). The MLE $\boldsymbol{b}$ maximizes the likelihood $L(\boldsymbol{\beta})$ of the sampling model. The estimator's covariance matrix V_{β} equals the inverse of the Fisher information matrix, $-E(d^2 \log (L(\boldsymbol{\beta}))/d\boldsymbol{\beta}^2)$. We evaluate the covariance matrix V_{β} at $\boldsymbol{b}$ to obtain the consistent estimate V_b. Both the MLE $\boldsymbol{b}$ and the covariance estimate V_b are invariant under the conventional sampling models that we assume, and our approximate likelihood is proportional to

$$L(\boldsymbol{\beta}|\mathbf{b}) \propto \exp\{-\tfrac{1}{2}(\boldsymbol{b} - \boldsymbol{\beta})^T V_b^{-1}(\boldsymbol{b} - \boldsymbol{\beta})\}. \tag{1.2}$$

Our analysis is based on a hierarchical Bayesian model for the loglinear model parameters. The first stage of our prior model is conveniently chosen to be the conjugate multivariate normal distribution for $\boldsymbol{\beta}$ centered about a reduced-model MLE $\boldsymbol{\mu} = (\boldsymbol{b}_r, 0)$. In this data-dependent mean $\boldsymbol{\mu}$, the first r entries are the elements of the reduced-model MLE $\boldsymbol{b}_r$, while the remaining p-s entries are zeros. We measure the uncertainty of this reduced-model with a hyperparameter in the spirit of Leonard and Novick (1986). Inferences about $\boldsymbol{\beta}$ are based on credible regions centred about posterior means that compromise the prior mean $\boldsymbol{\mu}$ and the full-model MLE associated with our approximate likelihood (1.2). Since both the prior mean and the full-model MLE represent hierarchical loglinear models, the preliminary stage in our analysis identifies these reduced and full hierarchical models.

If abundant prior knowledge is available, identification of these hierarchical models may be accomplished through an elicitation of prior beliefs about the parameters and cell probabilities. This article addresses the problem when only vague information about the parameters is available. In this instance, we suggest the posterior distribution of our hyperparameter be used to help screen for the order of the parameter vector. All hierarchical models of uniform order, the constant probability, the main effects, and the kth order interaction models, are considered. Two models of uniform order are identified which have associated parameter vectors that enclose the nonnull parameters. The smaller model is designated the reduced model associated with the prior, while the larger model is referred to as the full model that determines the likelihood in (1.2). Use of this larger hierarchical model, rather than the saturated model, to represent the likelihood (1.2) both reduces dimension and improves the accuracy of the normal distribution approximation to the likelihood. Fitting a loglinear model as a preliminary smoothing device to increase the efficiency of the primary inferences is an integral part of functional asymptotic regression methodology (Imrey et al. 1981 and Koch, Imrey, Freeman, and Trolley 1976). We feel this preliminary screening should be distinguished from the rest of the analysis, and it will, therefore, be referred to as the first stage of our strategy.

Frequentist approaches employ multiple significance tests to determine the important

summarizing terms that are included in their models (Agresti 1984 and Fienberg 1980). In these tests, the significance probability provides an inflated assessment of the evidence for nonnull terms (Berger and Selke 1987), and hence the frequentist model fitting strategy identifies too many important terms. Our analysis, which employs an informative prior on the loglinear model parameters, identifies fewer nonnull terms than identified by frequentist methods. Leonard (1975) and Laird (1978) have previously used the normal distribution in the first stage of their prior when they analyzed two-way contingency tables. Recent Bayesian approaches typically model the cell means rather than directly modelling the parameter vector, (Albert 1988; Leonard and Novick 1986; and Kass and Steffey 1989). Our approach is intended for high-dimensional contingency tables where prior knowledge is absent. Our choice of hierarchical prior facilitates a preliminary screening for the order of the table's structure, which enables us to specify a reasonable prior mean.

The more detailed description of the Bayesian analysis begins in Section 2, where we provide the assumptions of our prior distribution. Section 3 describes the posterior analyses, upon which our Bayesian strategy is based. Section 4 discusses the stages of our strategy that permit inferences about the loglinear model parameters. Three methods of approximation are illustrated in Section 5: an empirical Bayes method, the diagonalized covariance method, and Laplace's method (Tierney and Kadane 1986; Tierney, Kass, and Kadane 1989; Leonard 1982; and Kass and Steffey 1989). In Section 6, we reanalyze the Ries-Smith detergent-preference data. This Bayesian analysis reveals fewer important parameters than typically identified by frequentist analyses. In particular, the water softness term is not identified as important, and hence the table is collapsed over softness resulting in a simpler structure.

2. The Prior Distribution

Our specification of the parameter vector's prior distribution assumes that we have located $\boldsymbol{\beta}$ between the parameter vectors of two hierarchical models of uniform order. If $\boldsymbol{\beta}_r$ and $\boldsymbol{\beta}_f$ represent the r-dimensional reduced-model and p-dimensional full-model vectors, then we believe the number of important parameters is between r and p. In Section 4, we describe a method of identifying these models, which relies on a relative precision hyperparameter.

A two-stage normal-Cauchy-tail prior distribution is assumed for the parameter vector $\boldsymbol{\beta}$. Conditional on the relative precision parameter, α, the first-stage distribution for $\boldsymbol{\beta}$ is assumed to be the informative conjugate normal distribution

$$\pi(\boldsymbol{\beta}|\alpha) \propto \alpha^{-p/2} \times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu})^T \frac{1}{\alpha} C^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})\right\}. \tag{2.1}$$

The prior mean, $\boldsymbol{\mu}$, is assumed to be equal to the MLE, b_r, associated with the reduced hierarchical model found in the first stage of our analysis. When possible, we advocate a subjective elicitation of the structure of the prior covariance matrix C. Here, we examine the situation where little or no information is available, and we assume C is proportional to the identity matrix, $C = cI$. We define c to equal $p/tr(V_b^{-1})$ so that the trace of C^{-1} agrees with the trace of V_b^{-1}, the precision matrix associated with the likelihood.

The uncertainty of the first-stage normal

prior is measured by a second-stage prior on the overall precision parameter α, or equivalently, on the reparameterized version $\tau = 1/(1 + \alpha)$. If the precisions of the prior and the likelihood distributions are summarized by the trace of their precision matrices, then $\tau = \frac{1}{2}$ represents equal precision between the prior and the data. More generally, the relative precision parameter τ measures the proportion of the posterior precision that is attributable to the conjugate multivariate normal prior. Values of 1 and 0 for τ represent extreme belief and disbelief in the reduced prior model. We use Leonard's uninformative uniform prior for τ (Leonard 1977; Leonard and Novick 1986), which is centered at $\frac{1}{2}$ reflecting a neutral position concerning our belief in the reduced model. We then permit the data to help us update our knowledge of τ in the posterior analysis. The uniform prior on τ implies that the second-stage distribution for α is the Cauchy-tail prior with density

$$\pi(\alpha) = \frac{1}{(1 + \alpha)^2}. \tag{2.2}$$

We often find it computationally more convenient to express the prior distribution given in 2.1 and 2.2 in terms of τ rather than α

$$\pi(\boldsymbol{\beta}|\tau) \propto \left[\frac{\tau}{1 - \tau}\right]^{p/2} \times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu})^T \frac{\tau}{1 - \tau} C^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu})\right\} \tag{2.3}$$

$$\pi(\tau) = 1 \qquad 0 \leqslant \tau \leqslant 1. \tag{2.4}$$

Although we view the relative precision parameter τ as being analogous to Leonard's shrinkage proportion parameter, an important distinction in the parameter's interpretation should be mentioned. Leonard's shrinkage proportion parameter, in the Gamma or Dirichlet distributions, actually measures the proportion by which the MLE is shrunk towards the mean when one forms a Bayes estimate of the cell mean. In our Bayesian estimation of the loglinear model parameters, the components of the MLE are shrunk a variable amount so that τ cannot be interpreted as the constant proportion of shrinkage. This is the case either if C is specified to be proportional to the identity matrix, or if C is specified through an enlightened subjective elicitation procedure. Only if C were to be defined to equal the likelihood covariance V_b^{-1}, would τ represent the constant proportion of shrinkage. Fortunately τ does represent the relative precision and, therefore, may be interpreted as the typical shrinkage of the coordinate MLE's towards the prior mean.

We are indebted to Professor Leonard for the splendid tutorial (1987), in which we learned how the shrinkage proportion parameter could be employed in model-checking. Our choice of the Cauchy-tail second-stage prior was influenced by our need to identify a reasonable reduced-model MLE to serve as prior mean. The second-stage prior provides us with a sensible procedure for this initial screening (see Section 4.1). Other assignments for second-stage prior distributions have been used by researchers analyzing discrete data. These include the uniform prior (Leonard 1976; and Laird 1978), and the inverse chi-squared distribution (Leonard 1972, 1975). Lindley and Smith (1972), Box and Tiao (1973), Broemeling (1985), Berger (1980a, 1980b), and Press (1989) are among the authors who present Bayesian hierarchical methods for analyzing normal linear models.

We observe that a log uniform or uniform distribution would possess flatter tails than our Cauchy-tail distribution, and would, therefore, provide greater protection against

a misspecified prior model. This robustness is of vital concern to us, since our analysis is intended to be used when little prior information is available. In addition to providing a measure of robustness, the Cauchy-tail prior helps us to identify a reasonable prior mean. For this reason we believe it is preferable to other alternative distributions.

3. The Posterior Analysis

3.1. *The posterior analysis for* $\boldsymbol{\beta}$ *given* τ

A price is paid for using a uniform prior on the relative precision hyperparameter τ. The posterior analysis for the parameter vector $\boldsymbol{\beta}$ does not have a simple analytic representation. Our choice of a conjugate first-stage prior for $\boldsymbol{\beta}$, however, leads to the familiar multivariate normal posterior distribution when we condition on τ. That is, when the prior distribution in (2.3) and (2.4) is updated by the multivariate normal likelihood (1.2) of the full-model MLE, the posterior density of $\boldsymbol{\beta}$ given τ is

$$\pi(\boldsymbol{\beta}|\boldsymbol{b}_f, \tau) \propto |R_\tau|^{-1/2} \times \exp\{-\tfrac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_\tau)^T R_\tau^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_\tau)\} \tag{3.1}$$

where the conditional covariance matrix and conditional mean are

$$R_\tau = \left[V_b^{-1} + \frac{\tau}{(1-\tau)} C^{-1}\right]^{-1} \tag{3.2}$$

and

$$\boldsymbol{\beta}_\tau = R_\tau\left[V_b^{-1}\boldsymbol{b}_f + \frac{\tau}{(1-\tau)} C^{-1}\boldsymbol{\mu}\right] \tag{3.3}$$

respectively.

3.2. *The posterior density of* τ *and* $\boldsymbol{\beta}$

The posterior density of τ and $\boldsymbol{\beta}$ is given by

$$\pi(\tau, \boldsymbol{\beta}|\boldsymbol{b}_f) = \pi(\boldsymbol{\beta}|\boldsymbol{b}_f, \tau)\pi(\tau|\boldsymbol{b}_f) \tag{3.4}$$

where $\pi(\boldsymbol{\beta}|\boldsymbol{b}_f, \tau)$ is given in (3.1) and where $\pi(\tau|\boldsymbol{b}_f)$ is conveniently proportional to the multivariate normal distribution of $\boldsymbol{b}_f$ conditional on τ

$$\pi(\tau|\boldsymbol{b}_f) \propto \left|V_b + \frac{1-\tau}{\tau} C\right|^{-1/2} \times \exp\left\{-\tfrac{1}{2}(\boldsymbol{b}_f - \boldsymbol{\mu})^T \times \left(V_b + \frac{1-\tau}{\tau} C\right)^{-1} (\boldsymbol{b}_f - \boldsymbol{\mu})\right\}. \tag{3.5}$$

3.3. *The unconditional posterior analysis*

The unconditional posterior density of $\boldsymbol{\beta}$, a mixture of multivariate normal densities, is itself assumed to be approximately normal with mean $\tilde{\boldsymbol{\beta}}$ and covariance matrix $\tilde{R}$

$$\pi(\boldsymbol{\beta}|\boldsymbol{b}_f) \propto |\tilde{R}|^{-1/2} \times \exp\{-\tfrac{1}{2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^T \tilde{R}^{-1}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\}. \tag{3.6}$$

The two-staged prior distribution given in 2.3 and 2.4, is positive and continuous in a neighborhood of the limit of the maximum likelihood estimators. It follows that for a large number of observations, the posterior pdf is proportional to the approximately normal density of the likelihood (DeGroot 1970, ch. 10). Although this supports 3.6, when the likelihood does not dominate the posterior, this approximation is less accurate than the normal approximation for the conditional density displayed in 3.1. The mean $\tilde{\boldsymbol{\beta}}$ and covariance matrix $\tilde{R}$ in 3.6 must be approximated before posterior regions can be formed, and the important loglinear model terms identified. Similarly, the parameters of the posterior distribution of τ need approximating before an inference concerning a model's fit can be made. The posterior parameters, whose approximation are needed for Bayesian inferences, involve

posterior expectations of functions of $\boldsymbol{\beta}$ and τ, $G(\boldsymbol{\beta}, \tau)$. that may be represented as $(p + 1)$-dimensional integrals

$$E(G(\boldsymbol{\beta}, \tau)) = \int G(\boldsymbol{\beta}, \tau)\pi(\boldsymbol{\beta} | \boldsymbol{b}_f, \tau)\pi(\tau | \boldsymbol{b}_f)d\boldsymbol{\beta}d\tau. \quad (3.7)$$

Fortunately, when the posterior expectation of G given τ is known to be $g(\tau)$, the expectation (3.7) reduces to the one-dimensional integral

$$E(G(\boldsymbol{\beta}, \tau)) = \int_0^1 g(\tau)\pi(\tau | \boldsymbol{b}_f)d\tau. \quad (3.8)$$

Most of the posterior moments used in our analysis may be simplified in this way. The approximations of these moments is described in Section 5.

4. The Two-staged Analysis

Both stages of the analysis are based on the posterior hierarchical Bayesian distribution described in the preceding section. In the first stage, the posterior distribution of the relative precision parameter is used to identify the full and reduced models of uniform order, whose parameter vectors enclose the vector of nonnull parameters. In the second stage, the posterior distribution of the parameter vector $\boldsymbol{\beta}$, is used to identify important nonnull terms and to provide posterior regions for the parameters of these terms.

4.1. Stage 1: Identifying full and reduced models

Our belief in a model is reflected by the concentration of the parameter vector's distribution about the prior mean, the model-based MLE. A great deal of belief in the model is represented by a highly concentrated distribution for $\boldsymbol{\beta}$, which results when the overall precision parameter is very small ($\alpha \simeq 0$), or when the relative precision parameter τ is very large ($\tau \simeq 1$). Conversely, great disbelief in the model is represented by a disperse distribution for $\boldsymbol{\beta}$ occurring when α is large and when τ is small ($\tau \simeq 0$). The value $\tau = 1/2$ represents equal precision for the data and prior distributions, as measured by the trace of the precision matrices. We follow the hierarchical Bayesian approach advocated by Leonard and Novick (1986), and specify an uninformative uniform prior on τ. This prior is centered at 1/2, a neutral position for our belief in the reduced model. We then observe the data, and let the posterior distribution of τ indicate whether an increase or decrease has occurred in our belief in the model.

Under multinomial and Dirichlet distribution assumptions, Leonard (1977) used a posterior quadratic loss argument to suggest how τ should be used in model checking. He concluded that the decision concerning the model's fit should be based on the comparison for the posterior mean of τ with 1/2. If for mathematical convenience, we make the assignment of $C = V_b^{-1}$, then Leonard's argument can be replicated. This covariance structure alots the same precision as our assignment $C = cI$, $c = t/tr(V_b^{-1})$, so τ may still be interpreted as the proportion of posterior precision attributed to the prior distribution. Additionally, τ represents the constant proportion that the components of the MLE are shrunk towards the prior mean to form the posterior mean, when we assume the conjugate normal distribution that possesses this covariance specification. The argument for comparing the posterior expectation for τ with 1/2 is presented in the Appendix, Section A.4.

Our objective in the first stage, however, is not to make a dichotomous decision about the fit of a model. Instead, we wish to be confident that any reasonable model will be nested between a full and a reduced

model. A more stringent criterion is needed to ensure this confidence and we, therefore, base our decision on 95% posterior intervals for τ. An interval that is above or below 1/2 provides clear evidence that the model fits or does not fit the data adequately, whereas an interval containing 1/2 fails to provide a clear indication as to the model's fit.

An approximate $100(1 - \alpha)\%$ posterior interval for τ may be formed as

$$E(\tau|\boldsymbol{b}) \pm zSD(\tau|\boldsymbol{b}) \tag{4.1}$$

where z is the $100(1 - \alpha/2)$ percentile of the standard normal distribution, and where $\boldsymbol{b}$ is the parameter vector MLE summarizing the likelihood. When screening for full and reduced models, the parameter vector $\boldsymbol{b}$ is taken to be the unrestricted MLE, $\boldsymbol{b}_u$, associated with the saturated model. The conditional density for τ (3.5) must be modified by incorporating $\boldsymbol{b}_u$ and its associated estimated covariance matrix V_u. The mean and standard deviation of τ in (4.1) require the approximation of the first two moments of τ

$$E(\tau^k|\boldsymbol{b}) = \int_0^1 \tau^k \pi(\tau|\boldsymbol{b}_f)d\tau \qquad k = 1, 2. \tag{4.2}$$

Of course, the posterior distribution of τ is frequently skewed. Hence, when 1/2 is close to an endpoint of the interval, one should check the graph of τ's posterior distribution as well as posterior tail probabilities to evaluate the adequacy of the model.

4.2. *Stage 2: Inferences for the parameter vector* $\boldsymbol{\beta}$

Inferences about $\boldsymbol{\beta}$ are based on the unconditional posterior distribution of $\boldsymbol{\beta}$, which we have assumed to be multivariate normal with mean $\tilde{\boldsymbol{\beta}}$ and covariance matrix $\tilde{R}$. The parameters of the posterior distribution of $\boldsymbol{\beta}$, involve the evaluation of one dimensional integrals. Once the parameters of the posterior distribution are found, approximate highest posterior density (HPD) intervals and ellipsoids may be formed and used to identify important terms summarizing the interaction structure.

The approximate intervals for the one-parameter terms, $\boldsymbol{\beta}_i$, are given by

$$\tilde{\boldsymbol{\beta}}_i \pm z\tilde{R}_{ii}^{1/2} \tag{4.3}$$

where $\tilde{R}_{ii}$ is the ith diagonal element of $\tilde{R}$, the posterior covariance matrix of $\boldsymbol{\beta}$ given $\boldsymbol{b}_f$, and where z is the $100(1 - \alpha/2)$ percentile of the standard normal distribution.

If $\boldsymbol{\beta}^{(1)}$ is an s-dimensional vector consisting of the parameters making up a multi-parameter loglinear model term, then the posterior region is an ellipsoid. The approximate marginal posterior distribution of $\boldsymbol{\beta}^{(1)}$, is multivariate normal with parameters $\tilde{\boldsymbol{\beta}}^{(1)}$ and $\tilde{R}^{(1)}$, the respective subvector and submatrix of $\tilde{\boldsymbol{\beta}}$ and $\tilde{R}$. A $100(1 - \alpha)\%$ ellipsoid for $\boldsymbol{\beta}^{(1)}$ is given by

$$\{\boldsymbol{\beta}^{(1)}: (\boldsymbol{\beta}^{(1)} - \tilde{\boldsymbol{\beta}}^{(1)})^T \times \tilde{R}^{(1)-1}(\boldsymbol{\beta}^{(1)} - \tilde{\boldsymbol{\beta}}^{(1)}) \leqslant \chi^2_{s;1-\alpha}\} \tag{4.4}$$

where $\chi^2_{s;1-\alpha}$ is the $1 - \alpha$ percentile of the chisquare distribution with s degrees of freedom. A single parameter term is judged important when the $100(1 - \alpha)\%$ HPD interval fails to enclose 0, and a multi-parameter term is deemed important when $\boldsymbol{\beta}^{(1)} = 0$ is not contained within the ellipsoid. The ellipsoid fails to contain the null vector when the quadratic form Q exceeds the appropriate chisquare percentile

$$Q = \tilde{\boldsymbol{\beta}}^{(1)T}\tilde{R}^{(1)-1}\tilde{\boldsymbol{\beta}}^{(1)} \geqslant \chi_{s;1-\alpha}. \tag{4.5}$$

We note that many classical statisticians, who are accustomed to evaluating Wald statistics, will find the examination of quadratic forms to be a familiar procedure.

Formation of these posterior regions and

quadratic forms require the evaluation of the posterior parameters of the unconditional posterior distribution of $\boldsymbol{\beta}$. Components of the posterior mean $\tilde{\boldsymbol{\beta}}$, and covariance matrix $\tilde{R}$, may be expressed as one dimensional integrals. The ith component of the posterior mean, $\tilde{\boldsymbol{\beta}}$, is given by

$$\tilde{\boldsymbol{\beta}}_i = \int_0^1 \boldsymbol{\beta}_{ri}\pi(\tau|\boldsymbol{b}_f)d\tau \tag{4.6}$$

and the ijth element of the covariance matrix, $\tilde{R}$, may be calculated as

$$\begin{aligned}\tilde{R}_{ij} &= E(\mathrm{Cov}(\boldsymbol{\beta}_i, \boldsymbol{\beta}_j|\boldsymbol{b}_f, \tau)) \\ &\quad + \mathrm{Cov}(E(\boldsymbol{\beta}_i|\boldsymbol{b}_f, \tau), E(\boldsymbol{\beta}_j|\boldsymbol{b}_f, \tau)) \\ &= \int_0^1 (R_{\tau ij} + \boldsymbol{\beta}_{\tau i}\boldsymbol{\beta}_{\tau j})\pi(\tau|\boldsymbol{b}_f)d\tau - \tilde{\boldsymbol{\beta}}_i\tilde{\boldsymbol{\beta}}_j\end{aligned} \tag{4.7}$$

where $R_{\tau ij}$ is the ijth element of R_τ in (3.2).

4.3. *Inferences concerning other parameters*

Posterior tail probabilities, parametric residuals, cell means, and the logarithm of expected cell means are other parameters of interest to the researcher. All of these parameters may be expressed as functions of $\boldsymbol{\beta}$ and hence their marginal posterior distribution may be approximated by using the conditional (3.1) or unconditional (3.6) posterior distributions of $\boldsymbol{\beta}$.

4.3.1. *Posterior tail probabilities*

Our posterior probability regions are based on the additional distribution assumption (3.6) that the unconditional posterior distribution of $\boldsymbol{\beta}$ is approximately multivariate normal. In cases of questionable judgements, where the null parameter vector is near the boundary of the posterior region, it is advisable to calculate a posterior tail probability. For the contrast $c^T\boldsymbol{\beta}$, this tail probability is

$$\begin{aligned}&P(c^T\boldsymbol{\beta} \leqslant z|\boldsymbol{b}_f) \\ &= \int_0^1 P(c^T\boldsymbol{\beta} \leqslant z|\boldsymbol{b}_f, \tau)\,\pi(\tau|\boldsymbol{b}_f)d\tau\end{aligned} \tag{4.8}$$

where $P(c^T\boldsymbol{\beta} \leqslant z|\boldsymbol{b}_f, \tau) = \Phi((z - c^T\boldsymbol{\beta}_\tau/(c^T R_\tau c)^{1/2})$ and Φ is the cumulative normal distribution function.

4.3.2. *Parametric residuals, log contrasts, and parameter vector contrasts*

When the reduced and full models differ, it is useful to examine individual cells to see where the difference is most prominent, and to identify outliers. The parametric residual (Leonard and Novick 1986) is defined to be the difference between the logarithm of the full-model cell mean m_{rijk}

$$\rho_{ijk} = \log(m_{fijk}) - \log(m_{rijk}). \tag{4.9}$$

The parametric residual is a special case of the more general log contrast $l^T\boldsymbol{F}(\boldsymbol{m}) = \Sigma l_{ijk}\log(m_{ijk})$. Under either full or reduced models, however, the logarithm of the cell mean is a linear combination of the log-linear model parameters. The parametric residual, as well as the more general log contrast, may therefore be represented as a parameter vector contrast

$$\rho_{ijk} = z^T\boldsymbol{\beta}. \tag{4.10}$$

The posterior distribution of this parameter vector contrast $z^T\boldsymbol{\beta}$ is normal with mean $z^T\tilde{\boldsymbol{\beta}}$ and variance $z^T\tilde{R}z$.

4.3.3. *Cell means and other nonlinear differentiable functions of* $\boldsymbol{\beta}$

Sometimes, a goal of the analysis is to produce smoothed expected cell means, and at other times, hypotheses about the cell means are of importance (e.g., hypotheses of symmetry or marginal homogeneity). In these instances, the posterior distribution of

the cell mean is of interest. The vector of cell means, $\boldsymbol{m}$, may be expressed as a nonlinear differentiable transformation of $\boldsymbol{\beta}$. That is, $\boldsymbol{m} = \exp(X\boldsymbol{\beta}) = f(\boldsymbol{\beta})$, where $f_j(\boldsymbol{\beta})$, the jth component function of f, is the composition of the exponential operator with the linear function of $\boldsymbol{\beta}$ whose coefficients are the elements of the jth row of X. The delta method (see e.g., Bishop, Fienberg, and Holland (1975)) may be used to approximate the distribution of any differentiable transformation $f = f(\boldsymbol{\beta})$. Conditional on τ, the posterior distribution of f is the approximately multivariate normal

$$f|_{\boldsymbol{b}_f,\tau} \sim N(f(\boldsymbol{\beta}_\tau), H_\tau R_\tau H_\tau^T) \tag{4.11}$$

where the matrix, H_τ, has the derivative $df_i(\boldsymbol{\beta})/d\boldsymbol{\beta}_j$ evaluated at $\boldsymbol{\beta}_\tau$ for the ijth element. The unconditional posterior distribution of f has parameters, $\tilde{f}$ and $\tilde{R}_{fij}$, where components $\tilde{f}_i$ and $\tilde{R}_{fij}$ are obtained by integrating with respect to the posterior distribution of τ given in equation 3.5

$$\tilde{f}_i = \int_0^1 f_i(\boldsymbol{\beta}_\tau)\pi(\tau|\boldsymbol{b}_f)d\tau \tag{4.12}$$

$$\tilde{R}_{fij} = \int_0^1 [H_\tau R_\tau H_{\tau ij}^T + f_i(\boldsymbol{\beta}_\tau) f_j(\boldsymbol{\beta}_\tau)] \times \pi(\tau|\boldsymbol{b}_f)d\tau - \tilde{f}_i\tilde{f}_j. \tag{4.13}$$

Approximate HPD regions may be obtained by assuming f is normally distributed. For the special case when f represents the vector of cell means, $f(\boldsymbol{\beta}) = \exp(X\boldsymbol{\beta})$, the parameters of the conditional distribution (4.11) are $f(\boldsymbol{\beta}_\tau) = \exp(X\boldsymbol{\beta}_\tau)$ and $H_\tau R_\tau H_\tau^T = MXR_\tau X^T M$, where M is a diagonal matrix with ith diagonal element $\exp(X\boldsymbol{\beta}_\tau)_i$.

4.3.4. Belief ratios

In addition to identifying outlying cells, it is useful to measure a cell's effect on our belief in the model. This is done with the belief ratio

$$\nu_{ijk} = \frac{\tau_{ijk}}{\tau} \tag{4.14}$$

where τ_{ijk} is the hyperparameter measuring our belief in the reduced model after the outlying cell frequency, n_{ijk}, has been set aside. Removal of a cell frequency that was inconsistent with the model would increase the value of our hyperparameter, and consequently, would result in a large belief ratio. Values of ν_{ijk} greater than 1, therefore, indicate the cell lessens our belief in the reduced model.

5. Approximating Posterior Moments

The previous section illustrated how posterior regions are used to assess the importance of terms in the model. Formulation of these regions, however, depend upon the approximation of posterior parameters $\tilde{\boldsymbol{\beta}}$ and $\tilde{R}$. The one-dimensional integrals 4.6 and 4.7, are difficult to evaluate because the integrands involve the intractable inverses of matrices such as $(V_b + (1 - \tau)C/\tau)^{-1}$, whose elements involve both the variable τ and floating-point numbers. We illustrate three methods of approximation: an empirical-Bayes method, the diagonalized covariance method, and the Laplace method. Leonard (1972, 1975, 1976, 1982, 1987) has perhaps contributed most to the numerical approximation methods used in Bayesian categorical data analysis. Albert (1988) presents methods used in generalized linear models. General methods of approximation used in Bayesian statistics include analytical or numerical methods (Lindley 1980; Tierney and Kadane 1986), numerical integration methods (Naylor and Smith 1982), Monte Carlo methods (Kloek and Van Dijk 1978; Zellner and Rossi 1982; Shao 1989; Gelfand and Smith 1989). A summary of numerical methods and computer programs used in Bayesian analysis has been provided by Press (1989) and Smith, Skene, Shaw, Naylor, and Dransfield (1985).

5.1. *The empirical Bayes method*

We first consider empirical Bayes methods, which like hierarchical Bayes estimates, are robust against a misspecified prior reduced model. In the analysis of discrete data, the practice of replacing the unknown prior variance with an approximation in the formula for the Bayes estimate has been suggested by both empirical Bayesians and Bayesians (Leonard 1972; Fienberg and Holland 1973; Sutherland, Fienberg, and Holland 1974; Hudson 1974; Efron and Morris 1975; and Laird 1978). Typically, the prior parameters have been assumed independent. Our approximations are appropriate in the more general case involving an arbitrary symmetric prior covariance structure C. These approximations are computationally economical, and therefore are attractive when analyzing tables of especially large dimension. Although the estimates approximate hierarchical Bayes estimates, the connection with the uninformative second-stage uniform prior is lost. The distribution of $\boldsymbol{b}_f$ given τ

$$f(\boldsymbol{b}_f|\tau) \propto \left|V_b + \frac{1-\tau}{\tau} C\right|^{-1/2} \times \exp\left\{-\tfrac{1}{2}(\boldsymbol{b}_f - \boldsymbol{\mu})^T \times \left(V_b + \frac{1-\tau}{\tau} C\right)^{-1} (\boldsymbol{b}_f - \boldsymbol{\mu})\right\} \tag{5.1}$$

is used to estimate $\tilde{\tau}$. An estimate that generalizes the Efron and Morris estimator (1973), (Morris 1983), is the iterative solution to the equations

$$(\boldsymbol{b}_f - \boldsymbol{b}_r)^T \left(V_b + \frac{1-\tau}{\tau} C\right)^{-1} (\boldsymbol{b}_f - \boldsymbol{b}_r) = p - r - 1. \tag{5.2}$$

We use instead the positive part of an analytic estimate

$$\hat{\alpha} = 0 \wedge \left[\frac{Q_1 - (p-r)}{tr(V_b^{-1}C)}\right], \tag{5.3}$$

which generalizes an estimate that Oman (1982) proposed for use in regression analysis. Then $\tilde{\tau}$ is estimated as $\hat{\tau} = 1/(1 + \hat{\alpha})$. When the full hierarchical model is taken to be the saturated model, the quadratic form, $Q_1 = (\boldsymbol{b}_f - \boldsymbol{\mu})^T V_b^{-1} (\boldsymbol{b}_f - \boldsymbol{\mu})$, is the weighted-least-squares statistic used in frequentist analyses. Under certain sampling schemes, this statistic has been shown equal to the Wald statistic and to Neymann's minimum modified chi-square statistic (Bhapkar 1966; Bemis and Bhapkar 1983a, 1983b). The posterior parameters, $\tilde{\boldsymbol{\beta}}$ and $\tilde{R}$, may be approximated by substituting $\hat{\tau} = 1/(1 + \hat{\alpha})$ into the parameter expressions, $\boldsymbol{\beta}_\tau$ and R_τ. If these approximations are used in the intervals and ellipsoids given in 4.3 and 4.4, an approximate and computationally economical empirical Bayes analysis may be performed to identify the summarizing parameters. It should be noted that the approximate posterior mean, $\boldsymbol{\beta}_{\hat{\tau}}$, obtained by substituting $\hat{\tau}$ into $\boldsymbol{\beta}_\tau$, has attractive frequentist risk properties (Oman 1982). Simulations (Leighty 1985) suggest that the other related shrinkage estimators, such as Stein's Rule (Efron and Morris 1973) and Berger's Stein-effect Rule (Berger 1982), successfully improve on the frequentist risk of a vector of maximum likelihood estimators, over a significant region of the parameter space.

5.2. *The diagonalized covariance method*

The posterior parameters, $\tilde{\boldsymbol{\beta}}$ and $\tilde{R}$, may be found via one-dimensional numerical integration, if matrices such as $V_b + (1-\tau)C/\tau$ and $V_b^{-1} + \tau C^{-1}/(1-\tau)$ can be inverted quickly for each new value of the variable τ.

This inversion may be accomplished by simultaneously diagonalizing the Fisher information and prior covariance structure matrices, V_b and C. Diagonalization is a common approach used in the estimation of the multivariate normal mean (Strawderman 1978; Nebebe and Stroud 1986). Henceforth, we shall refer to this approximation as the diagonalized covariance method.

Both V_b and C are symmetric and C, at least, is positive definite. An iterative program may then be used to approximate the nonsingular matrix, P, that simultaneously diagonalizes V_b and C. P satisfies

$$P^T C P = I \quad \text{and} \quad P^T V_b P = D \tag{5.4}$$

where D is diagonal with ith diagonal element d_i, a root of the characteristic equation $det(V_b - xC) = 0$.

The matrices $V_b + (1 - \tau)C/\tau$ and $V_b^{-1} + \tau C^{-1}/(1 - \tau)$ can now be diagonalized and inverted

$$\left(V_b + \frac{1-\tau}{\tau} C\right)^{-1} = P\Omega_1 P^T$$

$$\left[V_b^{-1} + \frac{\tau}{1-\tau} C^{-1}\right]^{-1} = P^{T-1}\Omega_2 P^{-1} \tag{5.5}$$

where Ω_1 and Ω_2 are both diagonal with ith diagonal element equalling $\omega_{1i} = \tau/((d_i - 1)\tau + 1)$ and $\omega_{2i} = (1 - \tau)d_i/((d_i - 1)\tau + 1)$. These matrices and the formula $\Pi_{i=1}^{q} \omega_{1i}$, which is proportional to the determinant $det(V_b + (1 - \tau)C/\tau)^{-1}$, may be substituted into the expressions $\boldsymbol{\beta}_\tau$, R_τ, and $\pi(\tau|\boldsymbol{b}_f)$ in equations 3.2, 3.3, and 3.5. These one-dimensional integrals can then be numerically evaluated to approximate the posterior moments of τ and $\boldsymbol{\beta}$, which are used to make Bayesian inferences.

5.3. *The Laplace method*

In this section, we illustrate the Laplace method of approximating posterior moments of parameters with distributions concentrated above the origin (Tierney and Kadane 1986; Leonard 1982; Tierney, Kass, and Kadane 1989). Kass and Steffey (1989) show how this method may be used in conditionally independent hierarchical models. Our illustration uses the diagonalization of covariance to develop explicit expressions for the derivatives required in the Laplace method.

To use the Laplace approximation, one needs to maximize a modified likelihood function and evaluate the observed information at the maxima. Our convenient choice of the first-stage conjugate prior reduces our approximation task to a one-dimensional problem. Rather than approximating the $(p + 1)$-dimensional expectation of a function $G(\boldsymbol{\beta}, \tau)$, we typically need to approximate the one-dimensional integral

$$\begin{aligned} E(G) &= E(G(\boldsymbol{\beta}, \tau)|\boldsymbol{b}_f) \\ &= \int_0^1 g(\tau)\pi(\tau|\boldsymbol{b}_f)d\tau \end{aligned} \tag{5.6}$$

where the function $g(\tau)$ is the conditional posterior expectation of G given τ, $E(G(\boldsymbol{\beta}, \tau)|\boldsymbol{b}_f, \tau)$, and $\pi(\tau|\boldsymbol{b}_f)$ is given in expression (3.5). Expectation 5.6 may then be expressed as the ratio of two one-dimensional integrals of altered likelihoods, $\int e^{L^*(\tau)}d\tau / \int e^{L(\tau)}d\tau$. Upon approximating these likelihoods with second-order Taylor series expanded about their maxima, the Laplace approximation is obtained

$$E(G) = \frac{\sigma^*}{\sigma} \exp\{L^*(\hat{\tau}^*) - L(\hat{\tau})\}. \tag{5.7}$$

Here the likelihood $L(\tau)$ equals $\log \pi(\tau|\boldsymbol{b}_f)$, and the likelihood $L^*(\tau)$ equals $L(\tau) + \log g(\tau)$. The modes $\hat{\tau}$ and $\hat{\tau}^*$ maximize L and L^* respectively, and σ and σ^* represent the square roots of the observed information $-1/L''(\hat{\tau})$ and $-1/L^{*\prime\prime}(\hat{\tau}^*)$ respectively.

The posterior means, variances, and covariances of τ and $\boldsymbol{\beta}$ may be approximated in this manner. For instance, if we let $g(\tau)$ equal $R_{\tau ij} + \boldsymbol{\beta}_{\tau i}\boldsymbol{\beta}_{\tau j}$, expression 5.7 produces an approximation to $E(\boldsymbol{\beta}_i\boldsymbol{\beta}_j|\boldsymbol{b}_f)$. Using 5.7 to approximate the means of $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$, and subtracting their product from $E(\boldsymbol{\beta}_i\boldsymbol{\beta}_j|\boldsymbol{b}_f)$, would produce an approximate covariance of $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$.

The Newton–Raphson algorithm, used to maximize L and L^*, requires that we find expressions for the derivatives of L and L^*. When little or no prior information is available, we assume the prior covariance C is proportional to the identity matrix I. This simplifies diagonalization 5.4 so that P may be chosen to be orthogonal

$$P^TP = I \quad \textit{and} \quad P^TV_bP = D \tag{5.8}$$

where D is diagonal, now with ith diagonal element d_i, a root of the characteristic equation $det(V_b - xI) = 0$. This diagonalization permits us to evaluate expressions for the derivatives of L and L^*. L and L^* are functions of bilinear and quadratic forms of matrices H_1 and H_2 that may be diagonalized. H_1 and H_2, when expressed in terms of $\alpha = (1 - \tau)/\tau$, are

$$H_1(\alpha) = (V_b + \alpha cI)^{-1} = P\Lambda_1P^T$$

and

$$H_2(\alpha) = \left[V_b^{-1} + \frac{1}{\alpha c}I\right]^{-1}$$

$$= P^{T-1}\Lambda_2P^{-1} \tag{5.9}$$

where $\Lambda_1 = \text{diag}(\lambda_{11}, \lambda_{12}, \ldots, \lambda_{1p})$, $\lambda_{1i} = 1/(d_i + \alpha c)$, and $\Lambda_2 = \text{diag}(\lambda_{21}, \lambda_{22}, \ldots, \lambda_{2p})$, $\lambda_{2i} = \alpha cd_i/(d_i + \alpha c)$. When expressed in terms of τ, the log-likelihood L is proportional to

$$L(\tau) \propto \tfrac{1}{2}\log|H_1(\tau)| - \tfrac{1}{2}z^TH_1(\tau)z \tag{5.10}$$

where $z = \boldsymbol{b}_f - \boldsymbol{b}_r$ and where $H_1(\tau)$ and $H_2(\tau)$ are the expressions in 5.9 after $(1 - \tau)/\tau$ has been substituted for α.

The derivatives of L may be found by applying the chain rule, differentiating first with respect to $\alpha = (1 - \tau)/\tau$. When we express the first term, $f_1(\tau) = \frac{1}{2}\log|H_1(\tau)|$ in terms of α, we obtain $f_1(\alpha) = -\frac{1}{2}\sum_{i=1}^{p}\log(d_i + \alpha c)$. The derivatives of f with respect to τ are

$$f_1'(\tau) = f_1'(\alpha)\frac{d\alpha}{d\tau} \tag{5.11}$$

and

$$f_1''(\tau) = f_1''(\alpha)\left(\frac{d\alpha}{d\tau}\right)^2 + f_1'(\alpha)\frac{d^2\alpha}{d\tau^2}$$

where $d\alpha/d\tau = -1/\tau^2$, $d^2\alpha/d\tau^2 = 2/\tau^3$, $f_1'(\alpha) = -\frac{1}{2}\sum_{i=1}^{p}(c/(d_i + \alpha c))$, and $f_1''(\alpha) = \frac{1}{2}\sum_{i=1}^{p}c^2/(d_i + \alpha c)^2$. The same approach is taken with the second term, $f_2(\tau) = -\frac{1}{2}z^TH(\tau)z$. The function f_2 is expressed in terms of α, and the chain rule is applied as in 5.11 to obtain $f_2'(\tau)$. Here $f_2(\alpha) = \frac{1}{2}z^TP\Omega_1P^Tz = -\frac{1}{2}\sum_{i=1}^{p}(1/(d_i + \alpha c))\,y_i^2$, where $y = P^Tz$. Differentiating with respect to α, we obtain $f_2'(\alpha) = \frac{1}{2}\sum_{i=1}^{p}(c/(d_i + \alpha c)^2)\,y_i^2 = \frac{c}{2}z^T(H(\alpha))^2z$, and $f_2''(\alpha) = -\sum_{i=1}^{p}(c^2/(d_i + \alpha c)^3)\,y_i^2 = c^2z^T(H(\alpha))^3z$. Adding the derivatives of the two terms of L, we obtain $L'(\tau) = f_1'(\tau) + f_2'(\tau)$ and $L''(\tau) = f_1''(\tau) + f_2''(\tau)$. Although the matrix P, of computationally costly eigenvectors, is used to derive the derivative formulas, we observe these eigenvectors are not needed to evaluate the final expression. Other derivative formulas, used in the approximation of the moments of τ and $\boldsymbol{\beta}$, are provided in the Appendix.

In the Newton-Raphson iterations, we use an efficient subroutine for inverting symmetric matrices which uses the relationship between the inverse of a symmetric matrix with submatrices that form the matrix

(Searle 1971, pp. 27). It should be noted that the approximation is intended for positive functions, $g(\tau)$, with posterior distributions concentrated above the origin. The empirical Bayes approximations described in section 5.1 are used to suggest when $g(\tau)$ is negative or close to 0. In these special instances, one must approximate $E(-g(\tau))$ or $E(g(\tau) + e)$, e equalling some positive constant rather than $E(g(\tau))$. The empirical Bayes estimate of τ provides a starting value when maximizing $L(\tau)$.

6. The Ries-Smith Study

We chose the Ries-Smith study of laundry detergent preference to illustrate our treatment of a large dimensional cross-classification possessing scant prior information. The variables "softness", "use", "preference", and "temperature" are referred to as variables 1, 2, 3, and 4 respectively. To summarize the steps of our analysis:

1. We first use posterior probability intervals for the shrinkage proportion, 4.1, to situate the nonnull parameters between the parameter vectors associated with the constant probability reduced model, and the first-order interaction full model.
2. Bayesian estimates that compromise between the full-model and reduced-model maximum likelihood estimates are approximated in three ways: by empirical Bayes estimation methods, by the diagonalized covariance method, and by the Laplace method.
3. The associated approximate 95% HPD intervals and ellipsoids are then used to identify the important loglinear model terms.
4. The variable water softness is absent from the model, indicating that it is independent of the other variables. Hence, we collapse the table over water softness, refit the model with approximate Bayes estimates, and confirm the selection of terms obtained from the original data.

6.1. *Identifying full and reduced models*

The first stage of our analysis begins with the diagonalization of the estimated asymptotic covariance matrix, V_u. Then assuming the covariance structure, $C = cI$, where $c = t/trace(V_u^{-1})$, the posterior density of $\tau = 1/(1 + \alpha)$ in equation 3.5 is calculated for each model of uniform order. This is done for each model by setting the prior mean μ equal the the kth ordered model maximum likelihood estimates. The symbolic software package Maple was used to plot these densities. Hand drawn reproductions of these graphs are displayed in Figure 1. Estimates of $\tilde{\tau} = E(\tau|\boldsymbol{b})$ and 95% HPD intervals for $\tilde{\tau}$ are supplied in Table 1.

After examining Figure 1 and Table 1, it is unclear to what extent the main effects model fits the data. Although the posterior mean of τ exceeds 1/2, suggesting a good fit, both the sketch of the density and the 95% HPD interval suggest that values of τ less than 1/2 are also reasonable. We therefore

Table 1. Posterior means and HPD intervals for shrinkage proportion $\tilde{\tau}$

Model	$\tilde{\tau} = E(\tau/\boldsymbol{\beta})$	95% HPD Interval
Constant Probability (C)	.0025	(0, .1005)
Main Effects (0)	.647	(.387, .908)
First-Order Interaction (1)	.936	(.815, 1.00)
Second-Order Interaction (2)	.969	(.89, 1.00)

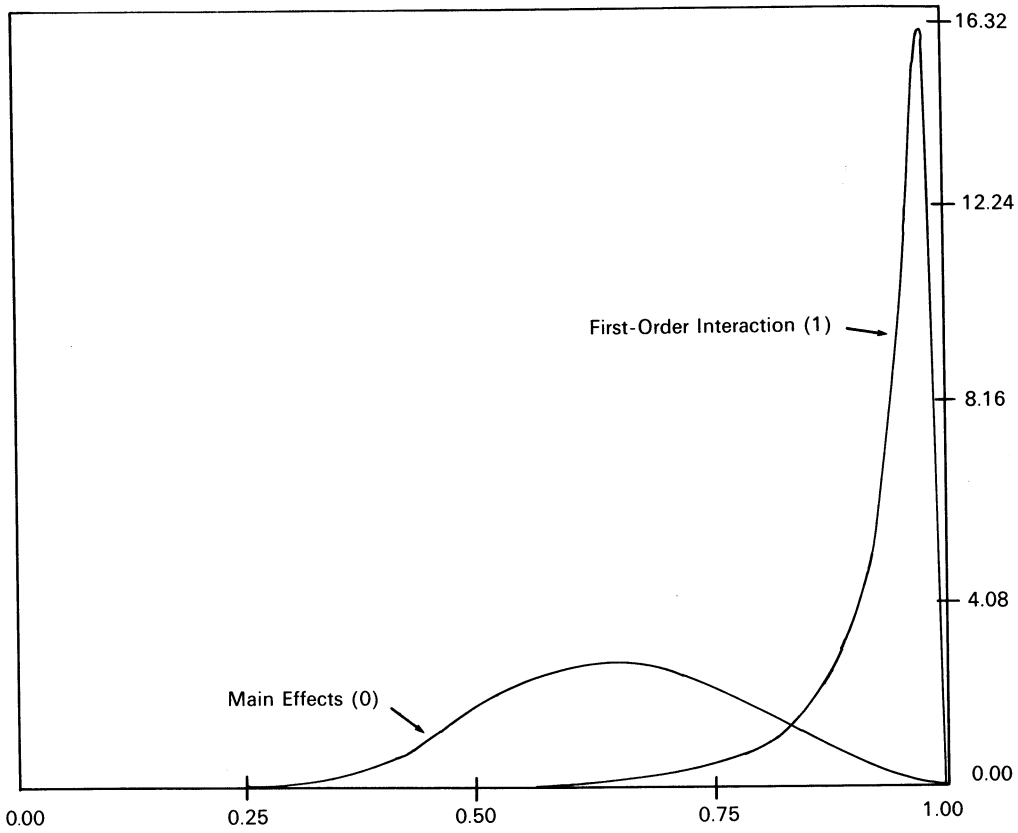

Fig. 1. The posterior densities of τ for models of uniform order

recommend that the hierarchical Bayesian estimation, performed in the second stage of the analysis, compromise the constant probability and first-order interaction model maximum likelihood estimates. The full-model maximum likelihood estimates of the loglinear model parameters are functions of the model-based estimates of cell counts that are more efficient than the observed frequencies. The assumed normal likelihood, therefore, is better approximated when these full-model estimates are used in place of the unrestricted maximum likelihood estimates. The use of loglinear models as a smoothing device in the first stage of the analysis has been referred to as functional asymptotic regression methodology (Koch et al. 1976; Landis, Heyman, and Koch 1978; and Imrey et al. 1981).

6.2. *Bayesian inferences for loglinear-model parameters*

All three approximations to the Bayesian model fitting strategy fit the same hierarchical model. The terms included in the Bayesian model are contrasted with the frequentist model obtained by using 95% confidence intervals and ellipsoids instead of HPD intervals and ellipsoids.

These models are suggested by the Bayesian and frequentist model fitting strategies

Table 2. Important terms identified by Bayesian and frequentist analyses

Parameter	u	u_1	u_2	u_3	u_4	u_{12}	u_{13}	u_{14}	u_{23}	u_{24}	u_{34}
Bayesian model	+	−	−	−	+	−	−	−	+	−	−
Frequentist model	+	−	−	−	+	−	−	+	+	?	−

(Bishop, Fienberg, and Holland 1975; Fienberg 1980; and Goodman 1971).

(a) [4][23]
(b) [4][23][14]
(c) [4][23][14][24]

Model (a), suggested by Bayesian 95% probability regions, includes only the use x preference interaction term, u_{23}. All these models appear to fit adequately since the posterior densities of τ associated with these models are all shifted to the right of 1/2.

The models differ, however, with respect to the decision to include the softness × temperature, u_{14}, and the use x temperature, u_{24}, interactions. The u_{24} parameter, having a frequentist significance probability of $P = .055$, is usually included in the frequentist model arising from other model fitting strategies (Bishop, Fienberg, and Holland 1975; Fienberg 1980; and Goodman 1971).

The more important conclusion is the absence of the softness × temperature

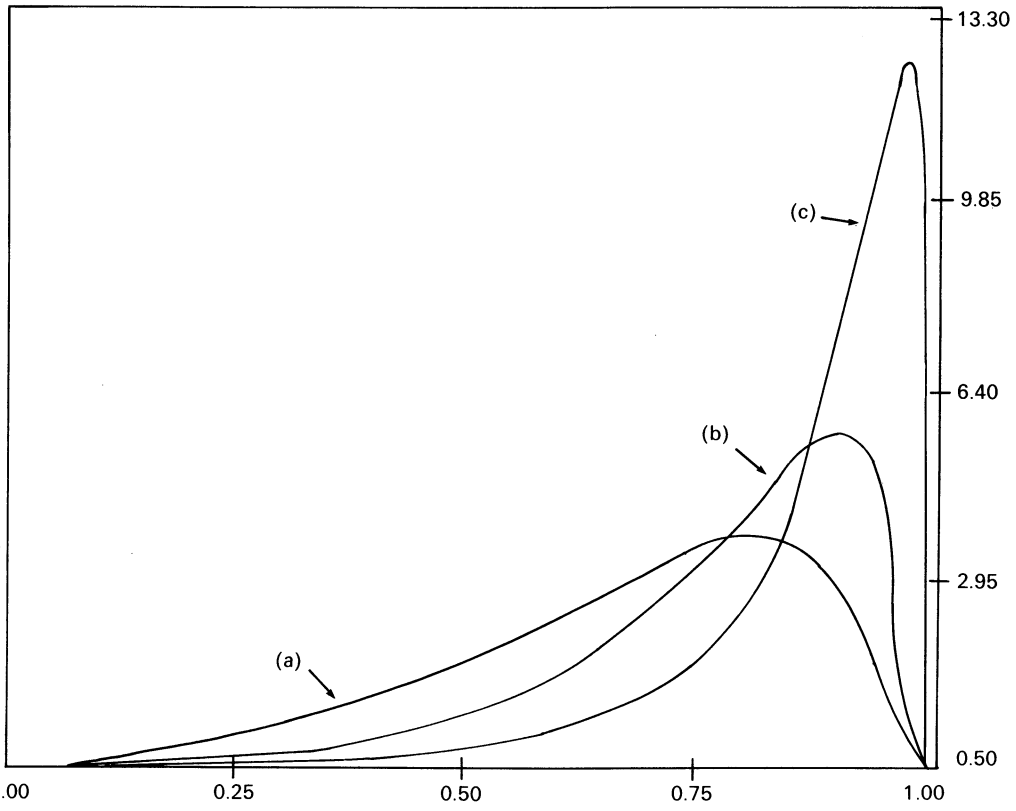

Fig. 2. The posterior densities of τ for candidate models

Table 3. Approximate means and HPD intervals for important parameters

Posterior Means, $\tilde{\beta}_i$
Original Data Estimates (Collapsed Data Estimates)

Method	Term:	u_4	u_{23}	u_{24}
Empirical Bayes		−.2366(−.2529)	−.1223(−.1317)	−.0544(−.0589)
Diagonalized Covariance		−.2303(−.2474)	−.1192(−.1290)	−.0530(−.0575)
Laplace		−.2304(−.2476)	−.1192(−.1291)	−.0530(−.0576)

95% HPD Intervals
Original Data Intervals
Collapsed Data Intervals

Method	Term:	u_4	u_{23}	u_{24}
Empirical Bayes		(−.2965, −.1767)	(−.1809, −.0637)	(−.1148, .0061)
		(−.3146, −.1912)	(−.1920, −.0715)	(−.1211, .0033)
Diagonalized Covariance		(−.2972, −.1635)	(−.1791, −.0592)	(−.1130, .0071)
		(−.3131, −.1817)	(−.1898, −.0682)	(−.1193, .0042)
Laplace		(−.2972, −.1636)	(−.1791, −.0593)	(−.1131, .0071)
		(−.3132, −.1820)	(−.1899, −.0683)	(−.1192, .0042)

Term	Method	Wald-like Quadratic Form
u_{14}	Empirical Bayes	Q = 4.429
	Diagonalized Covariance	Q = 4.085
	Laplace	Q = 4.080

term, u_{14}, in the Bayesian model: the 90% ellipsoid encloses the null value. The absence of this term and every other term involving variable 1 suggests collapsing over this variable, (Bishop, Fienberg, and Holland 1975, theorem 2.5.1). After collapsing and redoing the estimation the same model is implied, confirming that collapsing was the correct step in the analysis. The three approximations of the posterior means and intervals are shown in Table 3 for both the original and collapsed data.

The computationally economical empirical Bayes estimates agree with the hierarchical Bayes approximations to the second decimal place in the example, while the two hierarchical Bayes approximations typically agree to the third place. There is agreement concerning which of the four parameters are to be included in the model by the 95% posterior region criterion. Only the intervals and ellipsoids associated with u_{24} and u_{14} contain the null value. The largest hierarchical Bayes posterior intervals that fail to contain 0 have probabilities .916 for the original data and .932 for the collapsed data suggesting that u_{24} might be included if the 90% posterior region criterion were employed. The largest ellipsoid for u_{14}, which fails to include 0, has probability .870 suggesting u_{14} be excluded from the model.

Our assumption, that the unconditional posterior distribution of β_{13} is normal, suggests the posterior tail probability be

Table 4. Parametric residuals for the collapsed Ries-Smith data

		High Temperature		Low Temperature	
Brand Preference		X	M	X	M
Previous	M	−.450* (−)	−.056	.105	.268* (+)
Use	X	−.181	−.304* (−)	.485* (+)	.132

Starred residuals are more than three standard deviations from zero.

used to see if this normality assumption effects our conclusion concerning u_{24}. Using the Gaussian assumption, the tail probability is $P = P(\boldsymbol{\beta}_{13} \geqslant 0 | \boldsymbol{b}_f) = (1 - .916)/2 = .042$. When we use equation 4.8 we obtained $P = .0412$, which agrees closely with the Gaussian percentile.

Some frequentist models include the three-factor softness × use × preference interaction, u_{123}. In consideration of this finding, we also performed the hierarchical Bayes estimation, equating the full model with the second-order interaction model. This alteration had little effect. The same terms were identified as important, and u_{123} was not identified as significant by either the 95% or 90% posterior probability region criteria. The largest HPD ellipsoid for u_{123} that failed to contain 0 had probability .855.

The pattern of the parametric residuals is displayed in Table 4. Four residuals are especially important because their means are at least three standard deviations away from 0.

These residuals reflect the difference between the full and reduced models that is largely summarized by the important terms, temperature and use × preference. The low temperature residuals are the only positive residuals reflecting the comparatively small number of low temperature observations. Within the strata defined by temperature, the sign of the important residuals form a pattern consistent with the use × preference interaction: those who have previously used brand M are more apt to prefer brand M.

We summarize the use × preference interaction, u_{23}, by focusing on the ratio of the odds of preferring brand X at the two levels of use. This margin-free association is peculiar to the loglinear model, and is unaffected by the relative sizes of the variable levels. The inequity in the number of low temperature versus high temperature observations, for instance, has no effect on the odds ratio, or the loglinear model parameter u_{23}. We observe that $\psi = 4\boldsymbol{\beta}_5$ is the average of the logarithms of our summarizing odds ratios, or cross-product ratios, evaluated on the six strata defined by levels of water softness × temperature. Furthermore, since none of the higher-order relative terms of u_{12} were identified as important, the odds ratios may be assumed constant over the strata, and hence, ψ also represents the common log odds ratio. An estimate of ψ is $4\tilde{\boldsymbol{\beta}}_5 = -.516$. Exponentiating ψ provides the common odds ratio. Using the delta method (4.11), followed by one-dimensional integration with respect to τ, we may approximate the moments of this parameter. The mean of exp (ψ) is .596. Previous exposure to brand M is, therefore, associated with a 60% reduction in the odds of preferring

brand X over brand M. Ninety-five percent HPD intervals for the common log odds ratio and the typical odds ratio are $(-.759. -.273)$ and $(.448, .737)$, respectively.

7. Conclusions, Enhancements, and Generalizations

We have presented a general Bayesian analysis of higher-dimensional contingency tables. The analysis is appropriate when cell sizes are sufficiently large to warrant the assumption of normally distributed parameter maximum likelihood estimates. Inferences are based on regions and quadratic forms that are reminiscent of those regions and Wald statistics used in classical analyses. These familiar procedures may appeal to the researcher who is concerned about the inflated assessment of parameter importance that results when classical significance tests are used to fit a model (Berger and Selke 1987).

Although the regions may be familiar to many researchers, one enhancement is to use Bayes factors, or lower bounds for Bayes factors, to evaluate loglinear model terms. Another improvement in the precision of our inferences would result from a more careful elicitation of our prior distribution. The incorporation of a prior covariance matrix results in striking gains in the precision of estimating the multivariate normal mean (Berger 1980a, 1982). Eliciting a prior covariance matrix, then, should produce similar gains in the precision of estimating loglinear model parameters.

Our analysis easily generalizes to complex sample survey data. Bayesian or classical methods may first be used to obtain the parameters of the approximately normal parameter estimates. Then, apart from the slight alteration in the likelihood assumption, our Bayesian analysis remains the same.

A. Appendix

Here we provide more detail on the Laplace method of approximation. The common posterior moments that need approximating are $\tilde{\tau}$, $\tilde{\boldsymbol{\beta}}_i$, and $\tilde{R}_{ij}$. For the proper choice of conditional posterior expectation $g(\tau) = E(G(\boldsymbol{\beta}, \tau)|\boldsymbol{b}_f, \tau))$, each moment may be approximated by expression (5.6): $E(G) = \sigma^*/\sigma \exp\{L^*(\hat{\tau}^*) - L(\hat{\tau})\}$. First and second derivatives of L and L^* are incorporated into the Newton-Raphson algorithm that maximizes these altered likelihoods. In Section 5.3, the diagonalization (5.9) was used to develop the derivatives of L. To obtain the derivatives of $L^*(\tau) = L(\tau) + \log g(\tau)$, it is sufficient to find the first and second derivatives of g. With these derivatives, $L^{*\prime}$ and $L^{*\prime\prime}$ are

$$L^{*\prime}(\tau) = L'(\tau) + \frac{g'(\tau)}{g(\tau)}$$

and (A.1)

$$L^{*\prime\prime}(\tau) = L''(\tau) + \frac{g(\tau)g''(\tau) - g'(\tau)^2}{g(\tau)^2}.$$

A.1. The relative precision hyperparameter τ

Letting $g(\tau) = \tau$, the derivatives of L^* are $L^{*\prime}(\tau) = L'(\tau) + 1/\tau$ and $L^{*\prime\prime}(\tau) = L''(\tau) - 1/\tau^2$.

A.2. The posterior mean $\tilde{\boldsymbol{\beta}}$

To approximate $\tilde{\boldsymbol{\beta}}_i$, we assign $g(\tau)$ equal to the ith component of the conditional posterior mean $\boldsymbol{\beta}_\tau$. To find the derivatives, we first express $\boldsymbol{\beta}_\tau$ in terms of α as $\boldsymbol{\beta}(\alpha) = H_2(\alpha)z(\alpha)$ where $H_2(\alpha) = [V_b^{-1} + (1/\alpha c)\, I]^{-1} = P\Lambda_2 P^T$, and where $z(\alpha) = V_b^{-1}\boldsymbol{b}_f + (1/\alpha c)\boldsymbol{\mu}$. The chain rule may then be applied by initially differentiating with

respect to α

$$\frac{d}{d\tau}\boldsymbol{\beta}_\tau = \boldsymbol{\beta}'(\alpha)\frac{d\alpha}{d\tau} \tag{A.2}$$

$$\frac{d^2}{d\tau^2}\boldsymbol{\beta}_t = \boldsymbol{\beta}''(\alpha)\left(\frac{d\alpha}{d\tau}\right)^2 + \boldsymbol{\beta}'(\alpha)\frac{d^2\alpha}{d\tau^2}$$

where $d\alpha/d\tau = -1/\tau$ and $d^2\alpha/d\tau^2 = 2/\tau^2$. The vector of derivatives of $\boldsymbol{\beta}(\alpha)$ with respect to α is given by

$$\boldsymbol{\beta}'(\alpha) = H_2'(\alpha)z(\alpha) + H_2(\alpha)z'(\alpha)$$

and (A.3)

$$\boldsymbol{\beta}''(\alpha) = H_2''(\alpha)z(\alpha) + 2H_2'(\alpha)z'(\alpha) + H_2(\alpha)z''(\alpha)$$

where $z'(\alpha) = -(1/\alpha^2 c)\boldsymbol{\mu}$ and $z''(\alpha) = (2/\alpha^3 c)\boldsymbol{\mu}$ and where the matrices of derivatives with respect to α are expressed in terms of V_b and $H_1(\alpha) = (V_b + \alpha cI)^{-1}$

$$H_2'(\alpha) = cV_b^2 H_1(\alpha)^2$$

and (A.4)

$$H_2''(\alpha) = -2c^2 V_b^2 H_1(\alpha)^3.$$

The diagonalization (5.9) may be used to verify these last expressions. For instance, $H_2'(\alpha) = (d/d\alpha)P\Lambda_2 P^T = P\,\mathrm{Diag}((d/d\alpha)\lambda_{21}, (d/d\alpha)\lambda_{22}, \ldots, (d/d\alpha)\lambda_{2p})P^T$, where $(d/d\alpha)\lambda_{2i} = (d/d\alpha)\alpha c d_i/(\alpha c + d_i)) = -cd_i^2(1/(\alpha c + d_i)^2)$. Hence $H_2'(\alpha) = -cPD^2\Lambda_1^2 P^T = cV_b^2 H_1(\alpha)^2$.

A.3. The posterior covariance $\tilde{R}$

The ijth element $\tilde{R}_{ij}$ equals $E(\boldsymbol{\beta}_i\boldsymbol{\beta}_j|\boldsymbol{b}_f) - \tilde{\boldsymbol{\beta}}_i\tilde{\boldsymbol{\beta}}_j$. Choosing $g(\tau) = R_{\tau ij} + \boldsymbol{\beta}_{\tau i}\boldsymbol{\beta}_{\tau j}$, permits us to approximate the first term of $\tilde{R}_{ij}$, $E(\boldsymbol{\beta}_i\boldsymbol{\beta}_j|\boldsymbol{b}_f)$. The derivatives of g are the sums of the derivatives of $R_\tau = H_2(\tau)$ and of $\boldsymbol{\beta}_{\tau i}\boldsymbol{\beta}_{\tau j}$. Applying the chain rules once again

$$H_2'(\tau) = H_2'(\alpha)\frac{d\alpha}{d\tau}$$

and (A.5)

$$H_2''(\tau) = H_2''(\alpha)\left(\frac{d\alpha}{d\tau}\right)^2 + H_2'(\alpha)\frac{d^2\alpha}{d\tau^2}$$

where $H_2'(\alpha)$ and $H_2''(\alpha)$ are given in (A.4). To obtain the derivatives of $\boldsymbol{\beta}_{\tau i}\boldsymbol{\beta}_{\tau j}$, the product rule is applied

$$\frac{d}{d\tau}\boldsymbol{\beta}_{\tau i}\boldsymbol{\beta}_{\tau j} = \boldsymbol{\beta}_{\tau i}\frac{d}{d\tau}\boldsymbol{\beta}_{\tau j} + \boldsymbol{\beta}_{\tau j}\frac{d}{d\tau}\boldsymbol{\beta}_{\tau i}$$

and (A.6)

$$\frac{d^2}{d\tau^2}\boldsymbol{\beta}_{\tau i}\boldsymbol{\beta}_{\tau j} = \boldsymbol{\beta}_{\tau i}\frac{d^2}{d\tau^2}\boldsymbol{\beta}_{\tau j} + 2\frac{d}{d\tau}\boldsymbol{\beta}_{\tau i}\frac{d}{d\tau}\boldsymbol{\beta}_{\tau i} + \boldsymbol{\beta}_{\tau j}\frac{d^2}{d\tau^2}\boldsymbol{\beta}_{\tau i}.$$

Although the matrix P of eigenvectors is used to derive these derivatives, it is not required to evaluate them.

A.4. Leonard's preliminary test of significance

Leonard (1977) proposed a sensible alternative to the significance test for examining the null hypothesis that the cell probabilities of a contingency table are equal to a set of hypothesized values. The argument for this procedure, which is based on minimizing expected squared error loss, applies to hypotheses about the loglinear model parameters if the prior covariance structure is assumed to equal V_b, the covariance matrix associated with the likelihood. Suppose that we must decide whether to accept or reject the null hypothesis H_0 that the loglinear parameters $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_t$ equal the hypothesized values $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_t$ that form the components of our prior mean. Typically this procedure is used in classical statistics to make the simplified choice between estimators: choosing either the hypothesized values $\boldsymbol{\mu}$, or the unrestricted

MLE $\boldsymbol{b}_u$. That is, we estimate $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$, which has ith coordinate

$$\hat{\boldsymbol{\beta}}_i = \boldsymbol{b}_{ui} \quad \text{for } \boldsymbol{b}_u \in C \tag{A.7}$$

and

$$\hat{\boldsymbol{\beta}}_i = u_i \quad \text{for } \boldsymbol{b}_u \notin C$$

where C is the critical region. The quadratic loss

$$L(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \sum_{j=1}^{t} (\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)^2 \tag{A.8}$$

gives equal importance to the type I and type II errors. We find the optimal critical region C by maximizing the posterior expectation of the loss. When the prior covariance structure is assigned equal to V_b, the posterior mean conditional on τ given in equation (3.3) reduces to

$$\boldsymbol{\beta}_\tau = (1 - \tau)\boldsymbol{b}_u + \tau u. \tag{A.9}$$

The posterior expected loss given τ is

$$\sum_{j=1}^{t} (\hat{\boldsymbol{\beta}}_j - E(\boldsymbol{\beta}_j|\boldsymbol{b}_u, \tau))^2 + \sum_{j=1}^{t} \text{var}\,(\boldsymbol{\beta}_j|\boldsymbol{b}_u, \tau). \tag{A.10}$$

When $\boldsymbol{b}_u \notin C$ and $\hat{\boldsymbol{\beta}}_i = \boldsymbol{\mu}_i$, this reduces to

$$(1 - \tau)^2 \sum_{j=2}^{t} (\boldsymbol{b}_{uj} - \boldsymbol{\mu}_j)^2 + \sum_{j=1}^{t} \text{var}\,(\boldsymbol{\beta}_j|\boldsymbol{b}_u, \tau). \tag{A.11}$$

When $\boldsymbol{b}_u \in C$ and $\hat{\boldsymbol{\beta}}_i = \boldsymbol{b}_{ui}$, (A.10) reduces to

$$\tau^2 \sum_{j=1}^{t} (\boldsymbol{b}_{uj} - \boldsymbol{\mu}_j)^2 + \sum_{j=1}^{t} \text{var}\,(\boldsymbol{\beta}_j|\boldsymbol{b}_u, \tau). \tag{A.12}$$

After taking the expectation with respect to the posterior density of τ, we find that the critical region C consists of the values of $\boldsymbol{b}_u$ for which the expectation of (A.11) is less than the expectation of (A.12). Hence, $\boldsymbol{b}_u \in C$ when

$$E(\tau|\boldsymbol{b}_u) > 1/2. \tag{A.13}$$

8. References

Agresti, A. (1984). Analysis of Ordinal Categorical Data. New York: John Wiley.

Albert, J.A. (1988). Computational Methods Using a Bayesian Generalized Linear Model. Journal of the American Statistical Association, 83, 1037–1044.

Altham, P.M.E. (1970a). The Measurement of Association in a Contingency Table: Three Extensions of the Cross-Ratios and Metric Methods. Journal of the Royal Statistical Society, ser. B, 32, 395–407.

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

Bemis, K.G. and Bhapkar, V.P. (1983a). On BAN Estimators for Chi-Square Test Criteria. Annals of Statistics, 11, 183–196.

Bemis, K. G. and Bhapkar, V.P. (1983b). BAN Estimation for Chi-Square Test Criteria in Categorical Data. Communications in Statistics, Theory and Methods, 11, 1211–1223.

Berger, J. (1980a). A Robust Generalized Bayes Estimator and Confidence Region for a Multivariate Normal Mean. Annals of Statistics, 8, 716–761.

Berger, J. (1980b). Statistical Decision Theory. New York: Springer-Verlag.

Berger, J. (1982). Bayesian Robustness and the Stein Effect. Journal of the American Statistical Association, 77, 358–368.

Berger, J. and Selke, T. (1987). Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence. Journal of the American Statistical Association, 82, 112–122.

Bhapkar, V.P. (1966). A Note on the Equivalence of Two Test Criteria for Hypotheses in Categorical Data. Journal of the

American Statistical Association, 61, 228–235.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). Discrete Multivariate Analysis. Cambridge: M.I.T. Press.

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.

DeGroot, M.H. (1970). Optimal Statistical Decisions. New York: McGraw-Hill.

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.

Efron, B. and Morris, C. (1975). Data Analysis Using Stein's Estimator and Its Generalizations. Journal of the American Statistical Association, 70, 311–319.

Fienberg, S.E. (1980). The Analysis of Cross-Classified Categorical Data. Cambridge, MA: M.I.T. Press.

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

Froberg, C. (1969). Introduction to Numerical Analysis. Reading, MA: Addison-Wesley Publishing Company.

Gelfand, A.E. and Smith, A.F.M. (1989). Sampling Based Approaches to Calculating Marginal Densities. Technical Report, Department of Mathematics, University of Nottingham.

Goodman, L.A. (1971). The Analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimation Methods for Building Models for Multiple Classifications. Technometrics, 13, 33–61.

Grizzle, J.E., Starmer, C.F., and Koch, G.G. (1969). Analysis of Categorical Data by Linear Models. Biometrics, 25, 489–504.

Hudson, H.M. (1974). Empirical Bayes Estimation. Technical report no. 58, Stanford University, Stanford, Ca.

Imrey, P.B., Koch, G.G., Stokes, M.E., Darroch, J.N., Freeman, D.H., and Tolley, D. (1981). Categorical Data Analysis: Some Reflections on Loglinear Model and Logistic Regression: Part I: Historical and Methodological Overview. International Statistical Review, 49, 265–283.

Kass, R. and Steffey, D. (1989). Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models). Journal of the American Statistical Association, 84, 717–726.

Kloek, T. and Van Dijk, H.K. (1978). Bayesian Estimation of Equation System Parameters: An Application of Integration by Monte Carlo, Econometrica, 46, 1–19.

Koch, G.G., Imrey, P.B., Freeman, H.F., Jr., and Tolley, H.D. (1976). The Asymptotic Covariance Structure of Estimated Parameters From Contingency Table Loglinear Models. Proceedings of the Ninth International Biometric Conference, Biometrics Society, Vol. I, 317–336.

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

Landis, J.R., Heyman, E.R., and Koch, G.G. (1978). Average Partial Association in Three-Way Contingency Tables: A Review and Discussion of Alternative Tests. International Statistical Review, 46, 237–254.

Leighty, R.M. (1985). Interaction Terms for Categorical Data. Unpublished Ph.D. thesis, University of Connecticut.

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

Leonard, T. (1975). Bayesian Estimation Methods for Two-Way Contingency Tables. Journal of the Royal Statistical Society, ser. B, 37, 23–37.

Leonard, T. (1976). Some Alternative Approaches to Multi-Parameter Estimation. Biometrika, 63, 69–75.

Leonard, T. (1977). A Bayesian Approach to Some Multinomial Estimation and Pretesting Problems. Journal of the American Statistical Association, 72, 869–874.

Leonard, T. (1982). Comment on A Simple Predictive Density Function. Journal of the American Statistical Association, 77, 657–658.

Leonard, T. (1987). Applied Bayesian Categorical Analysis. American Statistical Association continuing education tutorial.

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

Lindley, D.V. and Smith, A.F.M. (1972). Bayesian Estimation for Linear Models (with discussion). Journal of the Royal Statistical Society, ser. B, 34, 1–41.

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

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.

Nebebe, F. and Stroud, T.W.F. (1986). Bayes and Empirical Bayes Shrinkage Estimation of Regression Coefficients. The Canadian Journal of Statistics, 14, 267–280.

Oman, S.D. (1982). Shrinking Towards Subspaces in Multiple Linear Regression. Technometrics, 24, 307–313.

Press, S.J. (1989). Bayesian Statistics: Principles, Models, and Applications. New York: John Wiley.

Sclove, S.L., Morris, C, and Radhakrishnan, R. (1972). Non-Optimality of Preliminary Test Estimators for the Multinormal Mean. Annals of Mathematical Statistics, 43, 1481–1490.

Searle, S.R. (1971). Linear Models. New York: John Wiley.

Shao, J. (1989). Monte Carlo Approximations in Bayesian Decision Theory. Journal of the American Statistical Society, 84, 727–732.

Smith, A.F.M., Skene, A.M., Shaw, J.E.H., Naylor, J.C., and Dransfield, M. (1985). The Implementation of the Bayesian Paradigm. Communications in Statistics, Theory and Methods, 14, 1079–1102.

Strawderman, W.E. (1978). Minimax Adaptive Generalized Ridge Regression Estimators. Journal of the American Statistical Association, 73, 623–627.

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, 585–617.

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

Tierney, L., Kass, R.E., and Kadane, J. (1989). Fully Exponential Laplace Approximations to Expectations and

Variances of Nonpositive Functions. Journal of the American Statistical Association, 84, 710–716.

Zellner, A. (1971). An Introduction to Bayesian Inference in Econometrics. New York: John Wiley.

Zellner, A. and Rossi, P.E. (1982). Bayesian Analysis of Dichotomous Quantal Response Models. Technical report, University of Chicago, Graduate School of Business.

Received July 1989
Revised February 1990