

BAYESIAN TESTS AND MODEL DIAGNOSTICS IN CONDITIONALLY INDEPENDENT HIERARCHICAL MODELS

JIM ALBERT*

Bowling Green State University, Bowling Green, USA

SIDDHARTHA CHIB

Washington University, St. Louis, USA

March, 1996

Abstract

Consider the conditionally independent hierarchical model (CIHM) where observations y_i are independently distributed from $f(y_i|\theta_i)$, the parameters θ_i are independently distributed from distributions $g(\theta|\lambda)$, and the hyperparameters λ are distributed according to a distribution $h(\lambda)$. The posterior distribution of all parameters of the CIHM can be efficiently simulated by Monte Carlo Markov Chain (MCMC) algorithms. Although these simulation algorithms have facilitated the application of CIHM's, they generally have not addressed the problem of computing quantities useful in model selection. This paper explores how MCMC simulation algorithms and other related computational algorithms can be used to compute Bayes factors that are useful in criticizing a particular CIHM. In the case where the CIHM models a belief that the parameters are exchangeable or lie on a regression surface, the Bayes factor can measure the consistency of the data with the structural prior belief. Bayes factors can also be used to judge the suitability of particular assumptions in CIHM's including the choice of link function, the nonexistence or existence of outliers, and the prior belief in exchangeability. The methods are illustrated in the situation where a CIHM is used to model structural prior information about a set of binomial probabilities.

Keywords: Bayes factor; Binomial data; Exchangeability; Exponential family; Hierarchical model; Generalized linear model; Gibbs sampling; Link estimation; Markov chain Monte Carlo; Metropolis-Hastings algorithm; Outliers; Partial exchangeability; Random effects.

**Address for correspondence:* Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403, USA.

1 Introduction

1.1 The conditionally independent hierarchical model

This paper considers the general problem of model criticism and testing within the context of a conditionally independent hierarchical model (CIHM). Suppose independent observations $\mathbf{y} = (y_1, \dots, y_n)$ are observed, where y_i is distributed according to the density $f(y_i|\theta_i)$ with unknown parameter θ_i . Suppose that the set of n parameters $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_n\}$ are believed a priori to satisfy some structural relationship, such as exchangeability or, more generally, a belief that they lie on a particular lower dimensional regression surface. One can model this belief by assuming that $\{\theta_i\}$ are independently distributed, where θ_i is distributed from a prior density $p(\theta_i|\boldsymbol{\lambda})$ where $\boldsymbol{\lambda}$ are hyperparameters common to all of the n prior distributions. The model specification is completed by assigning a prior distribution $p(\boldsymbol{\lambda})$ to the unknown set of hyperparameters. General discussions of hierarchical models are contained in Berger (1985, Chapter 4), Lindley and Smith (1972) and Morris (1983a).

One particular class of CIHM's can be defined by taking the sampling density $f(\mathbf{y}|\boldsymbol{\theta})$ to be a member of an exponential family (Morris, 1983b, Albert, 1988, Kass and Steffey, 1989, Lu, 1992). In particular, suppose that the y_i are independently distributed from an exponential family density $f(y_i|\eta_i)$ with mean η_i . Suppose that the η_i are believed a priori to satisfy the generalized linear model $g(\eta_i) = \theta_i = \mathbf{x}_i^T \boldsymbol{\beta}$, where g is a known link function mapping the mean η_i to the real valued parameter θ_i and $\boldsymbol{\beta}$ is an unknown regression parameter vector. One can model a belief in this regression structure by taking θ_i independently distributed from $N(\mathbf{x}_i^T \boldsymbol{\beta}, \tau^2)$ distributions and then assigning the hyperparameters $(\boldsymbol{\beta}, \tau^2)$ a known distribution $p(\boldsymbol{\beta}, \tau^2)$.

This paper focuses on the following binomial exchangeable model. Suppose that the y_i are independently distributed from binomial(n_i, p_i) distributions with sample sizes n_i

and probabilities of success p_i . Reparameterize the probabilities p_i to the logits $\theta_i = \log(p_i/(1 - p_i))$. To model the belief that the p_i are exchangeable, the θ_i are assumed to be a random sample from a $N(\mu, \tau^2)$ distribution and (μ, τ^2) are assigned a prior at the second stage.

In this paper, this exchangeable model is used in two applications. In an example from Tsutakawa et al (1985), one is interested in simultaneously estimating cancer mortality rates from a number of cities in Missouri. The incidence of cancer is not believed to vary significantly between cities and the use of the exchangeable model appears plausible. In a second example, one wishes to estimate the proportions of correct answers on different items of a mathematics placement test. Since all of the questions cover basic algebra and trigonometry skills, one might expect that the probabilities of correct response on different items to be similar, and therefore the exchangeable assumption again seems reasonable.

In the twenty years since the introduction of the normal hierarchical model by Lindley and Smith (1972), CIHM's have been shown to be a useful model structure for combining information from related experiments (see Gaver et al, 1992, for a recent review of applications). In addition, a number of new computational techniques have been developed for summarizing posterior distributions in this class of models. For example, Albert (1988) and Kass and Steffey (1989) illustrate computations using the Laplace method and, more recently, Gelfand and Smith (1990) and Gelfand, et al (1990) illustrate the use of Monte Carlo Markov Chain (MCMC) algorithms in obtaining samples from the joint posterior distribution of CIHM's. Albert and Chib (1993) and Dellaportas and Smith (1993) have developed MCMC algorithms for particular exponential family distributions which can be generalized in a straightforward way to the CIHM.

1.2 Model criticism

Despite the attractiveness of the CIHM, it should be recognized that any posterior inference is made conditional on the collection of assumptions implicit in the modeling. These assumptions are typically made for convenience. For example, parametric forms may be selected on the basis of conjugacy or because the chosen forms simplify the fitting procedure. A link function, such as the logit, may be chosen, so that the transformed parameters are easy to interpret.

These modeling assumptions, however, should be made with caution. It is possible that the patterns of the observed data may be inconsistent with the model and one has to think of alternative assumptions that provide a better fit. Note that the parameters $\{\theta_i\}$ in the basic exchangeable CIHM are assumed to be conditionally independent and identically distributed. This assumption does not appear to be tenable from a preliminary analysis of the cancer mortality rates since a few of the observed mortality rates have unusually large values. As a result, it is of interest to examine an alternative model for the rates that permits the occurrence of a few outliers. Another concern relates the choice of the link function and whether the logit link is appropriate for this data. In the context of the placement data set, the observed proportions of correct answers appear to cluster into two groups, corresponding to the easy and difficult questions. In this case, it may be preferable to fit a “two-group” exchangeable model in which questions within a particular group are believed exchangeable. In both applications, it is difficult to specify the second-stage prior for (μ, τ^2) . Thus, any particular choice of prior is at best a rough match to one’s prior beliefs. Thus it is of interest to investigate the sensitivity of the inference with respect to the choice of functional form or hyperparameters for the second stage prior.

In situations where there are plausible alternative models, the Bayesian paradigm pro-

vides a useful methodology for assessing the sensitivity to model assumptions and for criticizing different assumptions. Suppose that there exist K possible models, M_1, \dots, M_K , for the observed data. Associated with M_k , let ϕ_k denote the parameter vector and let $f_k(\mathbf{y}|\phi_k)$ and $\pi_k(\phi_k)$ denote the respective sampling density and prior density. If the prior density is given by $\sum_{k=1}^K \gamma_k \pi_k(\phi_k)$, where γ_k is the prior probability of model M_k , then by Bayes rule, the posterior density is given by $\sum_{k=1}^K \gamma_k(\mathbf{y}) \pi_k(\phi_k|\mathbf{y})$, where $\pi_k(\phi_k|\mathbf{y}) = C_k \pi_k(\phi_k) f_k(\mathbf{y}|\phi_k)$ is the posterior density of the unknown parameter conditional on model M_k and $\gamma_k(\mathbf{y})$ is the posterior probability of model M_k . The collection of posterior densities $\{\pi_k(\phi_k|\mathbf{y}), k = 1, \dots, K\}$ is useful in investigating the sensitivity of inferences with respect to changes in the model assumptions, and the posterior model probabilities $\{\gamma_k(\mathbf{y})\}$ are helpful in criticizing particular sets of assumptions (see Kass and Raftery, 1995, for a recent review of the use of this method in comparing models, and Carlin and Chib, 1995, for a MCMC model indicator algorithm to estimate $\gamma_k(\mathbf{y})$).

1.3 Comparing fixed and random effects models

Given a particular CIHM, the focus above is to criticize the assumptions by means of mixture models. A perhaps more fundamental issue is whether the observed data is consistent with the prior belief in the regression structure. In the exchangeable case, one wishes to decide between the fixed-effects model, where θ_i are all equal and the random effects model where the θ_i are different. In the cancer mortality data set, one questions if the mortality rates differ between cities. To construct a Bayesian test of the fixed-effects model, note that as the hyperparameter τ^2 approaches 0, the prior distribution concentrates on the regression surface $g(\theta_i) = \mathbf{x}_i^T \boldsymbol{\beta}$. The fixed-effects hypothesis is equivalent to the hypothesis H that $\tau^2 = 0$. One can test H against the alternative hypothesis K that τ^2 takes some positive value τ_0^2 by means of the Bayes factor $B_{KH} = m(\mathbf{y}|\tau_0^2)/m(\mathbf{y}|0)$, where $m(\mathbf{y}|\tau^2)$ is the

marginal probability of the data for a fixed value of the hyperparameter τ^2 . One can assess the goodness of fit of the reduced model $\theta_i = \mathbf{x}_i^T \boldsymbol{\beta}$ by computing the Bayes factor for a sequence of values of τ_0^2 (see Kass and Raftery, 1995 for a review of the use of Bayes factors).

1.4 Outline of the paper

In Section 2, a MCMC scheme for comparing models is illustrated for the CIHM. Using the conditional independence structure of the model, a simple MCMC algorithm is proposed to estimate a regression CIHM with particular emphasis on the exchangeable situation. With the exchangeable model as the base, three types of model perturbations — scale-inflated mixture distributions, partial exchangeability, and mixture link models — are introduced and, in each case, a MCMC algorithm is outlined to perform model diagnostics with respect to the perturbation. A model indicator algorithm is also developed for the case where one is uncertain about the specification of the second stage prior distribution of the hierarchical exchangeable model.

Section 3 discusses the use of simulation methods in deriving a Bayesian test for a reduced regression model. A MCMC model indicator algorithm is developed that permits movement between the fixed and random effects models. The Bayes factor is obtained from the simulated posterior distribution of τ^2 . Section 4 contains illustrations and Section 5 some brief concluding remarks.

2 Model diagnostics using MCMC in an exchangeable model

2.1 A MCMC algorithm to fit a hierarchical regression model

Consider the following three-stage CIHM which is useful in modeling the belief that a set of n parameters from an exponential family distribution satisfy a regression structure:

1. $y_i | \eta_i$ independent from $f(y_i | \eta_i)$, where $\eta_i = E(y_i)$

2. $\theta_i = g(\eta_i)$ are independent from $\mathcal{N}(\mathbf{x}_i^T \boldsymbol{\beta}, \tau^2)$
3. $(\boldsymbol{\beta}, \tau^2)$ are independent with $\boldsymbol{\beta}$ distributed $\mathcal{N}(\boldsymbol{\beta}_0, B_0^{-1})$ and τ^2 distributed $\mathcal{IG}(a, b)$.

In the above model description, f is a particular member of the exponential family such as the binomial, Poisson, or gamma, η_i is the mean of the i th observation, and g is a known link function which maps the support of the mean parameter into the real line. To model the belief that the η_i satisfy a regression structure, the transformed parameters θ_i are taken to be independent from normal distributions with mean parameters $\mathbf{x}_i^T \boldsymbol{\beta}$ and common variance parameter τ^2 . In the final stage of the CIHM, the unknown hyperparameters are assigned independent multivariate normal and inverse gamma distributions with known hyperparameter values.

Due to the conditional independence structure of this model, it is straightforward to construct a MCMC simulation algorithm to generate variates from the joint posterior distribution of $(\{\theta_i\}, \boldsymbol{\beta}, \tau^2)$. Conditional on the parameters $\boldsymbol{\beta}$ and τ^2 , $\theta_1, \dots, \theta_n$ have independent posterior distributions with θ_i distributed according to the density proportional to

$$f(y_i | h(\theta_i)) \phi(\theta_i; \mathbf{x}_i^T \boldsymbol{\beta}, \tau^2),$$

where $\eta = h(\boldsymbol{\theta})$ is the inverse link function and $\phi(\cdot; \mu, \tau^2)$ is the normal density function. Let $\boldsymbol{\theta}$ and X denote the vector of real-valued parameters and regression matrix, respectively. Then by combining stages 2 and 3 of the model, one sees that the hyperparameters $\boldsymbol{\beta}$ and τ^2 have the following conditional distributions:

$$\begin{aligned} \boldsymbol{\beta} | \{\theta_i\}, \tau^2, \mathbf{y} &\text{ distributed } \mathcal{N}(\hat{\boldsymbol{\beta}}, B_1^{-1}) \\ \tau^2 | \{\theta_i\}, \boldsymbol{\beta}, \mathbf{y} &\text{ distributed } \mathcal{IG}(a + n/2, b + \frac{1}{2} \sum_{i=1}^n (\theta_i - \mathbf{x}_i^T \boldsymbol{\beta})^2), \end{aligned}$$

where $\hat{\boldsymbol{\beta}} = B_1^{-1}(\boldsymbol{\beta}_0 + X^T \boldsymbol{\theta} \tau^{-2})$ and $B_1 = B_0 + X^T X \tau^{-2}$. The structure of the above conditional posterior distributions suggests a simple MCMC scheme. Given an initial guess

at the hyperparameters β, τ^2 , use n independent Metropolis steps (Tierney, 1994, Chib and Greenberg, 1995) to simulate values of the transformed means θ_i . In the applications discussed here, a random walk chain was used with the increment density $\mathcal{N}(0, c)$, where the standard deviation c is approximately twice the standard deviation of the density being simulated. (From empirical work, if the density is approximately normally distributed, then it appears that this method will accept candidates with probability 0.5). Next, given the simulated values of the θ_i , simulate β, τ^2 from the above normal and inverse gamma distributions, where one is conditioning on the most recently simulated values. This cycle simulates a Markov Chain which converges to the joint posterior distribution of $\{\theta_i\}, \beta, \tau^2$. After discarding a set of burn-in iterations, the remaining simulated sample is taken as a sample from the joint posterior distribution. For the examples of this paper, the burn-in time appeared to relatively short, and the posterior calculations were based on a simulated sample of 5000 iterations.

In the special case in which one is modeling a belief in exchangeability, stages 2 and 3 of the hierarchical model are replaced by the assumptions that, at the second stage, the θ_i are a random sample from a $\mathcal{N}(\mu, \tau^2)$ distribution and, at the third stage, μ and τ^2 have independent $\mathcal{N}(\mu_0, \sigma_\mu^2)$ and $\mathcal{IG}(a, b)$ distributions. In the following, this will be referred to as the “basic” exchangeable model. In this case, the hyperparameters μ and τ^2 have the following conditional distributions:

$$\mu | \{\theta_i\}, \tau^2, \mathbf{y} \text{ distributed } \mathcal{N} \left(\frac{\bar{\theta}n/\tau^2 + \mu_0/\sigma_\mu^2}{n/\tau^2 + 1/\sigma_\mu^2}, (n/\tau^2 + 1/\sigma_\mu^2)^{-1} \right)$$

$$\tau^2 | \{\theta_i\}, \mu, \mathbf{y} \text{ distributed } \mathcal{IG}(a + n/2, b + \frac{1}{2} \sum_{i=1}^n (\theta_i - \mu)^2).$$

2.2 Outlier and partial exchangeable models

As described in Section 1, suppose that one wishes to perturb the basic exchangeable model by replacing particular components by discrete mixtures which reflect alternative models or directions in which the model can fail. First, suppose that one or more of the θ_i are outlying in the sense that their locations are located far from the locations of the main group of parameters. A simple scale-multiple outlier model (Gastwirth and Cohen, 1970) is obtained by assuming that the θ_i are a priori independently distributed from the mixture density

$$\pi(\theta_i|\mu, \tau^2) = (1 - p_i)\phi(\theta_i; \mu, \tau^2) + p_i\phi(\theta_i; \mu, K\tau^2), \quad K > 1.$$

This density reflects the prior belief that, with probability p_i , the i th component θ_i is outlying. Typically, one chooses K to be equal to 2 or 3 and sets p_i equal to a small number, reflecting the prior belief that outliers are unlikely to occur.

The above MCMC scheme can be modified in a straightforward manner to incorporate this outlier model. Let A_i denote the scalar multiple of the variance corresponding to the component θ_i . A priori A_i is equal to the values 1 and K with known probabilities $1 - p_i$ and p_i , respectively. The posterior distributions of θ_i , μ , and τ^2 , conditional on all remaining parameters including A_i follow the expressions given above with the variance τ^2 of θ_i replaced by $A_i\tau^2$. Finally, the full conditional posterior distribution of A_i is discrete on the values 1 and K with probabilities proportional to $(1 - p_i)\phi(\theta_i; \mu, \tau^2)$ and $p_i\phi(\theta_i; \mu, K\tau^2)$, respectively (see also Guttman and Schollnik, 1994).

The above outlier model is one particular deviation from the assumption that the θ_i are exchangeable. Another deviation from exchangeability is a belief in *partial exchangeability*, which states that the θ_i can be partitioned into a number of groups such that, within each group, the parameters are exchangeable. With this assumption, one is interested in

shrinking the observations towards a central value in each group. (This behavior is termed multiple shrinkage by George, 1986 and O'Hagan, 1988. Consonni and Veronese, 1995, describe an alternative method of implementing partial exchangeability for binomial data.) Specifically, suppose that the partitioning is into a known number G of groups and let the random variable g_i denote the unknown group membership of θ_i ($g_i = 1, \dots, G$). Describe the partition as $\boldsymbol{\theta} = (\boldsymbol{\theta}_{(1)}, \dots, \boldsymbol{\theta}_{(G)})$, where $\boldsymbol{\theta}_{(j)} = \{\theta_i \text{ such that } g_i = j\}$. Suppose that a priori θ_i belongs to group j with probability p_{ij} , where $\sum_j p_{ij} = 1$.

Write the elements of the j th group as $\boldsymbol{\theta}_{(j)} = (\theta_{j1}, \dots, \theta_{jn_j})$. Assume that a priori the $\boldsymbol{\theta}_{(j)}$ are independent and, within the j th group, the θ_{jh} are exchangeable. This exchangeability assumption can be modeled as above by letting $\theta_{j1}, \dots, \theta_{jn_j}$ be a random sample from a $\mathcal{N}(\mu_j, \tau_j^2)$ and taking μ_j, τ_j^2 independent from known $\mathcal{N}(m_j, \gamma_j)$ and $\mathcal{IG}(a_j, b_j)$ distributions, respectively.

In this problem the unknown parameters consist of the θ_i , the group memberships g_i , and the second stage parameters (μ_j, τ_j^2) , which describe the locations of the G groups. A MCMC sampling scheme is constructed by alternately simulating from the θ_i and second stage parameters conditional on the group memberships, and the group memberships conditional on all remaining parameters. If the group memberships are known, then the $\boldsymbol{\theta}$ has been partitioned into the G groups. The θ_{jh} , μ_j , and τ_j^2 have the following conditional distributions:

1. θ_{jh} distributed from $f(y_{jh}|h(\theta_{jh}))\phi(\theta_{jh}; \mu_j, \tau_j^2)$
2. μ_j distributed $\mathcal{N}\left(\frac{\sum_{h=1}^{n_j} \theta_{jh}/\tau_j^2 + m_j/\gamma_j}{n_j/\tau_j^2 + 1/\gamma_j}, (n_j/\tau_j^2 + 1/\gamma_j)^{-1}\right)$
3. τ_j^2 distributed $\mathcal{IG}(a_j + n_j/2, b_j + \frac{1}{2} \sum_{h=1}^{n_j} (\theta_{jh} - \mu_j)^2)$.

To complete one cycle of the Gibbs sampler, one needs to simulate the group member-

ships. Conditional on all remaining parameters, the group membership g_i of θ_i is discrete on the integers $1, 2, \dots, G$ with probabilities proportional to $\{p_{ij}\phi(\theta_i; \mu_j, \tau_j^2), j = 1, \dots, G\}$.

2.3 Different link functions

Another assumption of the basic exchangeable model is that the mean parameters η_i are connected to the real-valued parameters θ_i by the known link function g . To investigate the appropriateness of the choice of g , one can imbed this choice of link within a family of links $\theta_i = g_\lambda(\eta_i)$, where λ is an element of a discrete set $\{\lambda_1, \dots, \lambda_L\}$. For example, in the binomial case where η_i is a proportion, one can consider the class of symmetric link functions (Aranda-Ordaz, 1981)

$$T_\lambda(\eta_i) = \frac{2\eta_i^\lambda - (1 - \eta_i)^\lambda}{\lambda\eta_i^\lambda + (1 - \eta_i)^\lambda},$$

where the choices $\lambda = 0, 1$ correspond to the use of logistic and linear links, respectively, and a probit link corresponds approximately to the value $\lambda = .39$. The choice of this family of links can not be done arbitrarily, since an exchangeable prior is placed on the θ_i , and so the family needs to be defined so that the function $g_\lambda(\eta_i)$ has approximately the same location for different values of λ . The choice of link family for the binomial problem has this characteristic. Values of η close to .5 are mapped to θ values about 0 and the different values of λ reflect differences in the function for values of η close to 0 or 1.

A MCMC scheme is derivable by a simple modification of the scheme for the basic exchangeable model (see also Carlin and Polson, 1991, for a similar method in regression models). The values of θ_i are generated independently from densities proportional to $f(y_i|h_\lambda(\theta_i))\phi(\theta_i; \mu, \tau^2)$, where $h_\lambda(\theta_i)$ is the inverse link transformation. Values of the second stage parameters μ and τ^2 are generated just in the basic exchangeable case. Lastly, one generates a value of the link function parameter λ by simulating from the discrete distribution $\{\lambda_j\}$ with probabilities proportional to $\{p_j \prod_{i=1}^n f(y_i|h_{\lambda_j}(\theta_i))\}$.

2.4 Alternative second-stage priors

In the basic exchangeable model, one is also concerned about the assessment of the priors for the second stage parameters μ and τ^2 . The assumptions of independence of the two parameters and assignments of normal and inverse gamma functional forms are made primarily for convenience. A user may wish to use alternative priors for (μ, τ^2) . One may wish to investigate the sensitivity of the posterior analysis for θ_i across changes in the prior for the second stage parameters. In addition, one is interested in model criticism. Are there particular choices of the prior distribution which are more consistent with the observed data?

In general, suppose that the user has L alternative priors for (μ, τ^2) . The prior distribution can be represented as the discrete mixture $\pi(\mu, \tau^2) = \sum_{j=1}^L p_j \pi_j(\mu, \tau^2)$, where the $\pi_j(\cdot)$ represent the different priors and the p_j are the prior probabilities of the L models. If one was solely interested in comparing different choices for the prior, the probabilities $p_j = 1/L$ could be used.

To implement a MCMC algorithm, introduce the model indicator $I \in \{1, \dots, L\}$, which indicates which of the L priors is true. One wishes to simulate from the joint posterior distribution of $(\{\theta_i\}, \mu, \tau^2, I)$. As before, the θ_i are simulated independently from the densities $f(y_i|h(\theta_i))\phi(\theta_i; \mu, \tau^2)$. Given that the model indicator $I = j$, one simulates (μ, τ^2) from the bivariate density proportional to

$$\prod_{i=1}^n \phi(\theta_i; \mu, \tau^2) \pi_j(\mu, \tau^2).$$

One cycle of the MCMC iterations is completed by simulating I from $\{1, \dots, L\}$ with probabilities proportional to $\{p_j \pi_j(\mu, \tau^2)\}$.

3 A MCMC algorithm to obtain a Bayesian test of a reduced model

As in the previous section, consider the basic exchangeable model

1. $y_i|\eta_i$ independent from $f(y_i|\eta_i)$, where $\eta_i = E(y_i)$
2. $\theta_i = g(\eta_i)$ are independent from $\mathcal{N}(\mu, \tau^2)$
3. (μ, τ^2) are independent with μ distributed $\mathcal{N}(\mu_0, \sigma_\mu^2)$ and τ^2 distributed from $\pi(\tau^2)$,

where $\pi(\tau^2)$ is an arbitrary distribution placed on the variance hyperparameter. To construct a test of the hypotheses $H : \tau^2 = 0$, $K : \tau^2 > 0$, a discrete/continuous prior distribution needs to be constructed for τ^2 . Suppose that, with probability p_0 , $\tau^2 = 0$, and, with probability $1 - p_0$, τ^2 is assigned a density $\rho(\tau^2)$. Then the Bayes factor in support of the alternative hypothesis K over H is given by

$$BF = \frac{P(\mathbf{y}|K)}{P(\mathbf{y}|H)} = \frac{\int_0^\infty m(\mathbf{y}|\tau^2)\rho(\tau^2)d\tau^2}{m(\mathbf{y}|\tau^2 = 0)},$$

where $m(\mathbf{y}|\tau^2)$ is the marginal density of the observed data \mathbf{y} for a fixed value of τ^2 . The posterior probability of the alternative hypothesis is given by $P(K|\mathbf{y}) = (1 + \frac{p_0}{1-p_0}BF)^{-1}$. See Berger and Delampady (1987) for a general discussion of this method in testing point null hypotheses.

Values of the Bayes factor can be computed from a simulated sample of values of the marginal posterior density of τ^2 . A priori the hypotheses $\tau^2 = 0$ and $\tau^2 > 0$ have respective prior probabilities p_0 and $1 - p_0$. From the simulated sample, one can estimate the posterior probabilities of the hypotheses $p_1 = P(\tau^2 = 0|\mathbf{y})$ and $1 - p_1 = P(\tau^2 > 0|\mathbf{y})$. The Bayes factor is computed as the ratio of the posterior odds to the prior odds $BF = \frac{(1-p_1)/p_1}{(1-p_0)/p_0}$.

To simulate from the joint posterior distribution, the following MCMC scheme is used. First, rewrite the first stage of the exchangeable prior model as $\theta_i = \mu + \gamma_i$, where γ_i

is distributed $N(0, \tau^2)$. To simulate from the posterior density of $(\{\gamma_i\}, \mu, \tau)$, we block the parameters into the subsets $(\{\gamma_i\}, \tau)$ and μ , and simulate variates in turn from the distribution of $(\{\gamma_i\}, \tau)$ conditional on μ , and the distribution of μ conditional on $(\{\gamma_i\}, \tau)$.

To simulate from the conditional distribution of $(\{\gamma_i\}, \tau)$, a Metropolis/Hastings type algorithm is used. Suppose that $(\tau^{(g)}, \{\gamma_i^{(g)}\})$ represent the current value of the parameters. Generate a candidate value from the proposal distribution $q(\tau, \{\gamma_i\})$. This proposal distribution first generates a value of τ , $\tau^{(c)}$, and then generates a vector $\{\gamma_i^{(c)}\}$ conditional on the simulated value $\tau^{(c)}$. Compute the acceptance probability $PROB = \min\{1, q\}$, where

$$q = \frac{\prod_{i=1}^n [f(y_i | h(\theta_i^{(c)})) \phi(\gamma_i^{(c)}; 0, \tau^{(c)2}) \pi(\tau^{(c)2}) q(\tau^{(c)}, \{\gamma_i^{(c)}\})]}{\prod_{i=1}^n [f(y_i | h(\theta_i^{(g)})) \phi(\gamma_i^{(g)}; 0, \tau^{(g)2}) \pi(\tau^{(g)2}) q(\tau^{(c)}, \{\gamma_i^{(g)}\})}.$$

If a uniform random variate, U , is smaller than $PROB$, then the candidate vector $(\tau^{(c)}, \{\gamma_i^{(c)}\})$ is accepted; otherwise the algorithm remains at the current vector value $(\tau^{(g)}, \{\gamma_i^{(g)}\})$. A Metropolis algorithm is also used to simulate from the conditional distribution of μ .

The success of this algorithm depends on a suitable proposal density $q(\tau, \{\gamma_i\})$. This proposal density can be constructed using an approximation to the posterior density. This is described here for the binomial/logit model; a similar approximation can be developed for exponential family densities. Approximately, the empirical logit $z_i = \log((y_i + .5)/(n_i - y_i + .5))$ is approximately normal with mean $\mu + \gamma_i$ and variance $\sigma_i^2 = 1/(y_i + .5) + 1/(n_i - y_i + .5)$. Conditional on μ and τ , the posterior densities of the γ_i are approximately independent normal with mean and variance

$$\hat{\gamma}_i = \frac{(z_i - \mu)/\sigma_i^2}{1/\sigma_i^2 + 1/\tau^2}, \quad v_i = \frac{1}{1/\sigma_i^2 + 1/\tau^2}.$$

Using this approximation, the proposal density q is defined as follows.

1. Simulate a value of τ from the mixed prior density

$$p_0 \{\tau^2 = 0\} + (1 - p_0) \{\tau^2 > 0\} \rho(\tau^2).$$

Call the simulated value $\tau^{(c)}$.

2. If $\tau^{(c)} = 0$, then the proposed values of γ_i are all identically zero. If $\tau^{(c)} > 0$, then simulate γ_i from independent normal distributions with respective means and variances $\hat{\gamma}_i$ and v_i .

In the cancer mortality example discussed in the next section, the above algorithm is run using a discrete distribution $\rho(\tau^2)$ over a set of plausible τ^2 values, say $\tau_1^2, \dots, \tau_H^2$. A large prior probability p_0 is chosen for $\tau^2 = 0$, so that the algorithm will visit this value a sufficient number of times to obtain an accurate approximation to its posterior probability. From the empirical frequencies of the values $0, \{\tau_h^2\}$ from the simulation run, the posterior probabilities are calculated and these are converted to values of the Bayes factor. Specific comments about the choice of grid of τ^2 values and the accuracy of this simulation algorithm will be made in the example of section 4.1.2.

4 Examples

4.1 Cancer mortality data

4.1.1 Introduction

To illustrate the methods described in Sections 2 and 3, consider the cancer mortality data analyzed in Tsutakawa et al (1985). One is interested in simultaneously estimating the rates of death from stomach cancer for males “at risk” in the age bracket 45-64 for the 84 largest cities in Missouri. For the i th city ($i = 1, \dots, 84$), one observes the number n_i at risk and the number of cancer deaths y_i . Suppose that the $\{y_i\}$ are independent binomial with respective probabilities of death $\{p_i\}$. If the cancer death rates are not believed to vary significantly across cities (small or large) and across males in this age range, then it may be reasonable to believe a priori that the rates are exchangeable. This prior belief can

be modeled (as in Section 2.1) by letting the logits $\theta_i = \log(p_i/(1 - p_i))$ be independent $\mathcal{N}(\mu, \tau^2)$, and then assigning μ and τ^2 independent $\mathcal{N}(\mu_0, \sigma_\mu^2)$ and $\mathcal{IG}(a, b)$ distributions, respectively.

4.1.2 Fixed or random effects model?

First, one may question if a random effects model is appropriate for this data set. A simpler model would be a fixed effects model in which the probability of death is the same across cities. Is there sufficient evidence that the death rates $\{p_i\}$ do vary between cities? One can test the hypothesis that $p_1 = \dots = p_n$ by computing a Bayes factor for the equivalent hypothesis that $\tau^2 = 0$ against the alternative hypothesis that τ^2 equals some positive value τ_0^2 . Table 1 illustrates the Bayes factor computations using the MCMC algorithm described in Section 3. First, a grid of equally spaced values of $\log(\tau^2)$ was chosen and a prior was used which placed probability .5 on the value $\tau^2 = 0$ and the remaining .5 probability uniformly over the positive τ^2 values on the grid. In this example, a vague prior distribution was chosen for the hyperparameter μ . Next, the MCMC algorithm was run for $m = 100,000$ iterations and the posterior probabilities were estimated by the relative frequencies of the simulated values on the grid.

τ	0	.041	.061	.091	.135	.202	.301	.449	.670
Prior	.5	.0625	.0625	.0625	.0625	.0625	.0625	.0625	.0625
Posterior	.0689	.0151	.0278	.0697	.1762	.2823	.2614	.0932	.0053
Bayes Factor	1.00	1.75	3.23	8.09	20.4	32.4	30.2	10.7	.62
log(Bayes Factor)	0	.242	.509	.908	1.31	1.51	1.48	1.03	-.21

Table 1: Bayes factors to test fixed-effects model for cancer mortality data set.

To gauge the accuracy of these computed Bayes factors, an alternative computational method can be used. The marginal density of \mathbf{y} conditional on τ^2 can be written as the

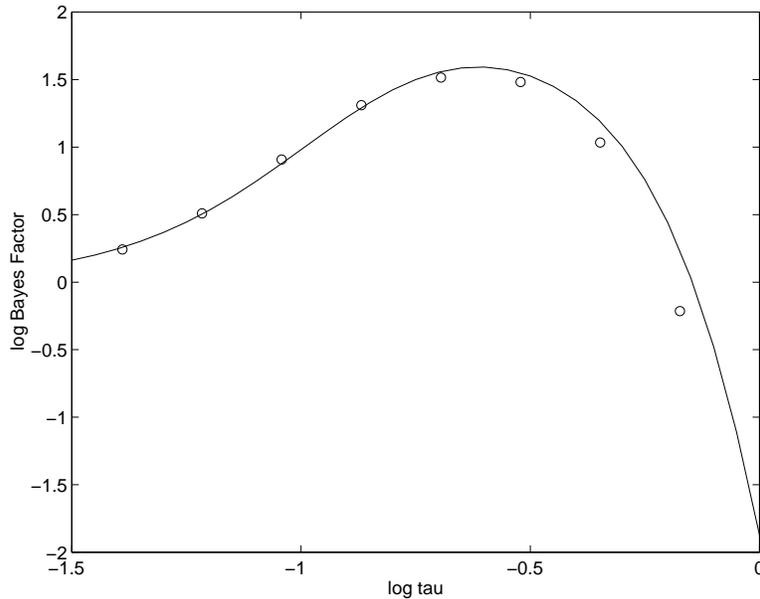


Figure 1: Values of the log Bayes factors computed using adaptive quadrature (smooth curve) and MCMC (plotted points) for cancer mortality data set.

$(n + 1)$ -dimensional integral

$$m(\mathbf{y}|\tau^2) = \int \left\{ \prod_{i=1}^n f(y_i|h(\theta_i))\phi(\theta_i; \mu, \tau^2) d\theta_i \right\} d\mu.$$

For a fixed value of μ , the integrand in $\{\theta_i\}$ is a product of n one-dimensional integrals, each of which can be computed by quadrature. Then the outer integral in μ is computed by quadrature. Each (univariate) integral is evaluated by the algorithm of Naylor and Smith (1982). Figure 1 contains results from these calculations. The “exact” values of the log Bayes factor for various values of $\log \tau$ (computed by the direct integration method) are plotted as a smooth curve. The corresponding values that appear in Table 1 (computed from the MCMC simulation method) are also shown in the figure. Note that all the simulation-based estimates lie close to the smooth curve, indicating that the MCMC algorithm has provided accurate estimates.

It should be also noted from Table 1 and Figure 1 that, as τ increases, the values of the

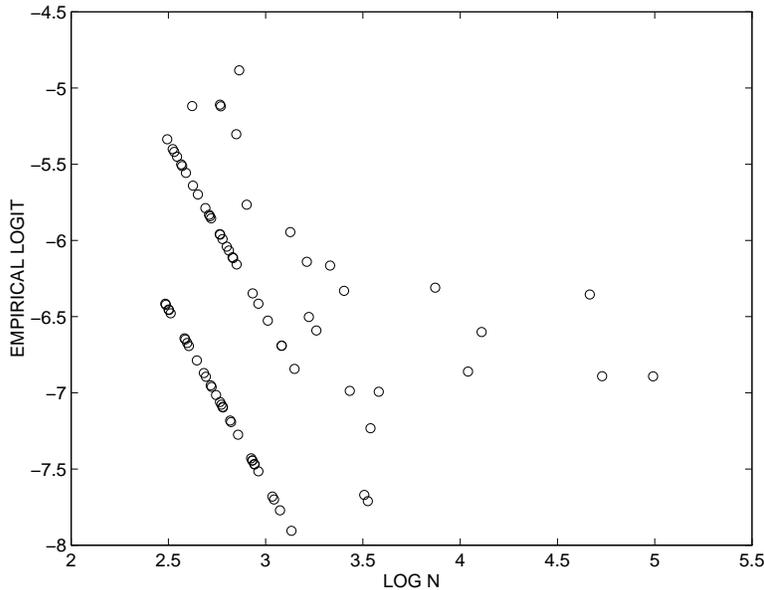


Figure 2: Plot of empirical logits against logarithms of the sample sizes for the cancer mortality data set.

log Bayes factor first increase and then sharply decrease when $\log \tau$ is approximately -0.6 . In addition, the log Bayes factors are in the $1 - 1.5$ range for an interval of τ values. (A grid of τ values should be chosen to cover the value at which the Bayes factor is maximized.) This analysis provides some support for a random effects model with a positive value of τ .

4.1.3 Fitting the exchangeable model

As a first look at the mortality data, Figure 2 plots the empirical logits $\{\log \frac{y_i + 1/2}{n_i - y_i + 1/2}\}$ against the logarithms of the sample sizes $\{n_i\}$ of each city. The two parallel lines of points in the plot correspond to observations that have death counts of 0 and 1. Generally, the variability of observed logits is relatively large for the smaller cities. More accurate estimates for these cities can be obtained by borrowing information across all cities.

Consider the use of the basic MCMC algorithm described in Section 2.1 with the hyperparameter values set at $\mu_0 = 0$, $\sigma_\mu^2 = 1000$, $a = 5$ and $b = 1$. These values reflect relatively

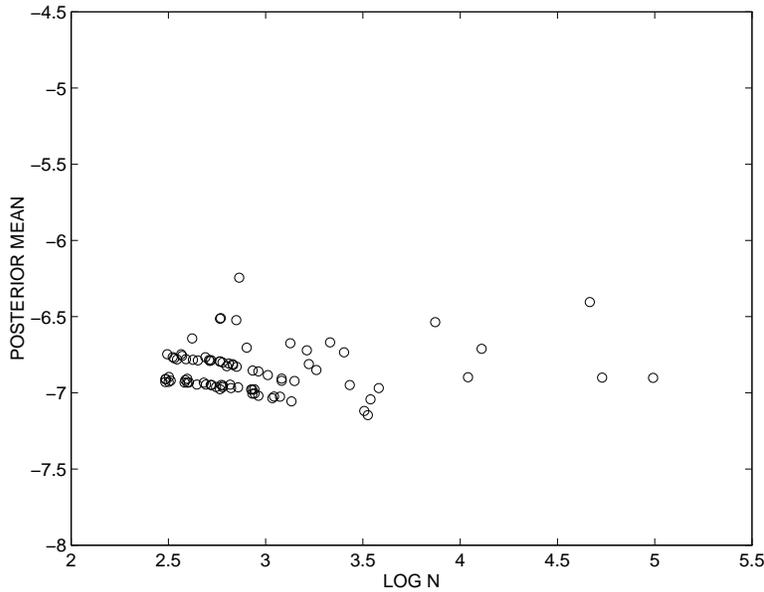


Figure 3: Posterior means of the logits for the cancer mortality data set.

vague prior beliefs about the location of the second stage parameters. Figure 3 contains the plot of the posterior means of the logits $\{\theta_i\}$. On comparing the plots in Figures 2 and 3, it can be noted that the observed logits of the smaller cities are shrunk substantively towards an average mortality value. This model appears to have a much smaller effect on the mortality rates for the larger cities, since their respective posterior means are close to the observed logits.

Next we examine the merits of the initial modeling assumptions. Note from Figure 3 that several of the posterior means appear to be somewhat distinct from the main group of observations. We thus consider an alternative outlier model which can accommodate unusual values of the θ_i . We also examine if an alternative link function can provide a better fit.

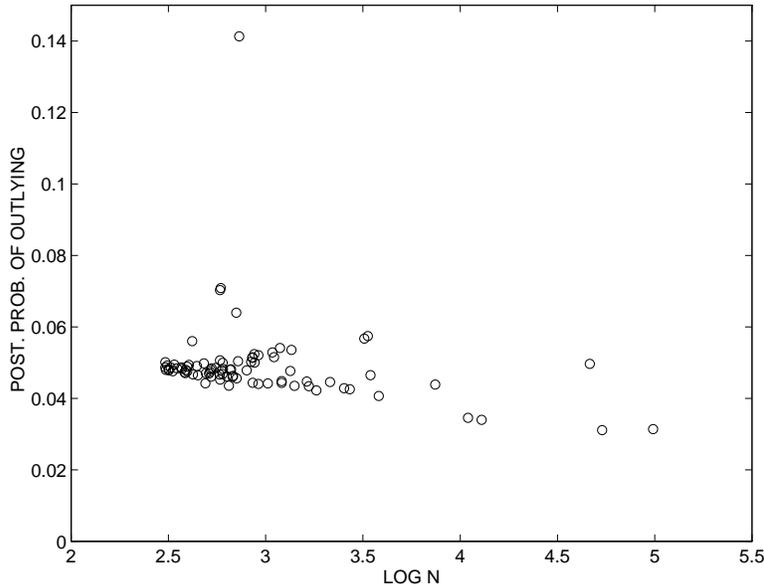


Figure 4: Posterior outlying probabilities using outlier exchangeable model for the cancer mortality data set.

4.1.4 Fitting an outlier model

Consider the scale-inflated mixture model described in Section 2.2 and suppose that, a priori, each logit θ_i is distributed $\mathcal{N}(0, \tau^2)$ with probability .95 and distributed $\mathcal{N}(0, 3\tau^2)$ with probability .05. Equivalently, the unknown scale multiple A_i is equal to 1 and 3 with prior probabilities .95 and .05, respectively. Suppose that the second stage hyperparameters take the same values as in the basic model. The posterior probability that $A_i = 3$, for each city, is obtained from the MCMC algorithm described above. The results are plotted in Figure 4. Note that one city has a large posterior outlying probability, suggesting that this city may have an usually large mortality rate.

Although this new model has identified a possible outlier, it is not clear that it is a significant improvement over the basic exchangeable model. To check this, a new model was run with an indicator variable which selects the basic model or the outlier model. With the prior probability of each model set at 0.5, the posterior probability of the basic model

was estimated to be .88. Thus, although the outlier model has identified one possible outlier, it has not led to a significantly better fit than the basic exchangeable model.

The MCMC algorithm with model indicators can also be used to investigate the suitability of the choice of a logistic link for this data set. The main problem is to define some plausible alternative link functions for this data set. In Section 2.3, a class of symmetric link functions mapping the real valued θ to the probability p was proposed. However, these functions are “matched” only for values of p in an interval about .5, and this will be unsuitable for this example, where the mortality rates are close to zero. A closely related family of transformations is the power family (Tukey, 1977)

$$g_v(p) = a_v \left(\frac{p}{1-p} \right)^v + b_v.$$

The choices $v = 0, a_v = 1, b_v = 0$ give the logistic link, and, for alternative power values v , the constants a_v and b_v can be chosen to match the logistic link for a particular interval of p values (Emerson and Stoto, 1983). For this data set, an average mortality rate is $p = .001$ and, for the choices $v = -.5$ and $v = .5$, values of the linear constants can be found so that the transformation $g_v(p)$ has the same value as the logistic link at $p = .001$ and the derivative at $p = .001$ has the same value as the logit. Figure 5 plots three inverse link functions $h_v(\theta)$: the logistic cdf function and the two alternative link functions for the range of θ values in this example. Using the terminology of Tukey (1977), the power value $v = .5$ is one step away from the logit towards a linear link function and the value $v = -.5$ has one step greater curvature than the logit.

For this example, the simulation was performed using a mixture model with equal prior probabilities for the three link functions and the second-stage hyperparameters set as above. After $m = 2000$ cycles of the algorithm, the posterior relative frequencies of the links $v = -.5, 0, .5$ were estimated to be .18, .52, .30, respectively. Thus, for this data set, the

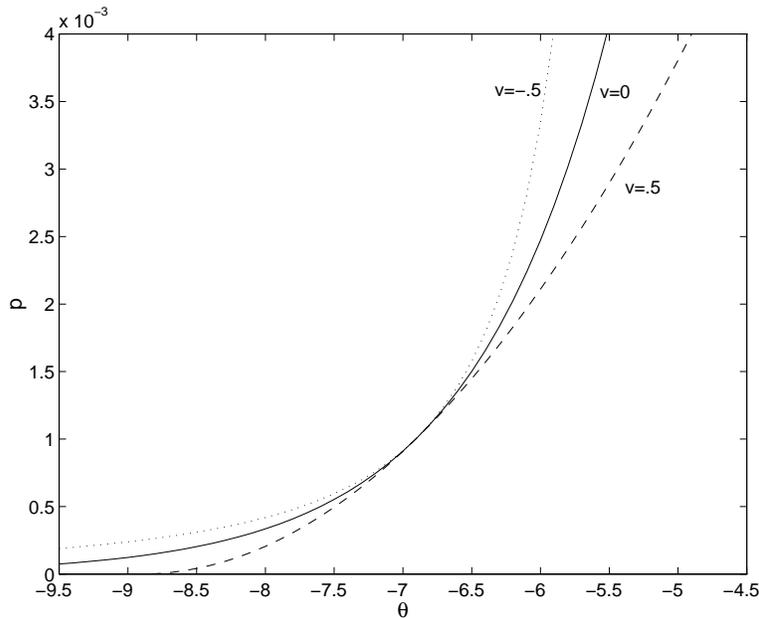


Figure 5: Three inverse link functions $h_v(\theta)$. Values of (v, a, b) are $(-.5, -.0633, -4.9068)$, $(0, 1, 0)$ and $(.5, 63.2139, -8.9068)$.

use of the logistic link is preferred, and the “more linear” link $v = .5$ is more preferable than the “more curved” link $v = -.5$. This analysis is not intended to be a thorough study. However, it gives support to the use of the logit link in combining proportions from related experiments.

4.2 Placement test data

4.2.1 Introduction

As a second example, consider an item analysis from a mathematics placement test. At Bowling Green State University, freshmen are given a 32 multiple choice test to aid in mathematics class placement. For a sample of 200 students, the number of correct answers is recorded for each question. Figure 6 plots the observed logits $\log\left(\frac{\text{number correct}}{\text{number incorrect}}\right)$ against the question number.

Let p_i denote the probability a student answers question i correct, $i = 1, \dots, 32$. It is

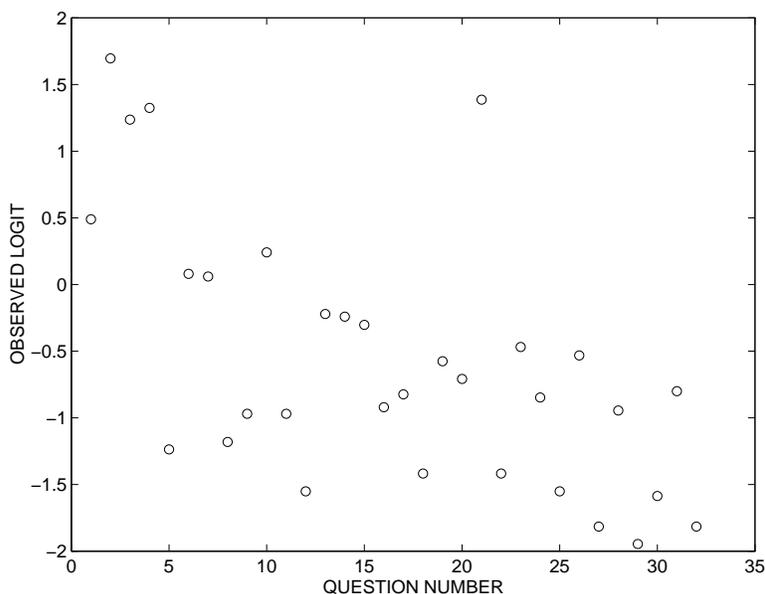


Figure 6: Plots of observed logits against question number for math placement data set.

clear from the figure that a fixed effects model in which the p_i are equal is not reasonable — the problems appear to differ in terms of the level of difficulty. However, it is not clear from this figure if the alternative random effects/exchangeable model is suitable. Many of the observed logits are in the -1.5 – $.5$ range. A few questions, however, appear to be relatively easy (logits values from 1 to 2) and there appears to be slight downward trend in the graph. These observations in the graph are consistent with knowledge of the test items. The questions that test basic algebra skills are relatively easy compared with questions on trigonometry or questions on simplifying algebraic expressions containing fractions.

4.2.2 Fitting a partial exchangeable model

The above comments motivate the consideration of a partial exchangeable model as described in Section 2.2. Consider the simplest model of this type, in which there are two classes of questions and, within a class, the questions are believed to be exchangeable. This

model requires the specification of second-stage hyperparameters which specify the locations of the two groupings of logits $\{\theta_i\}$. For this example, the values $(m_1, \gamma_1) = (-1, .01)$, $(m_2, \gamma_2) = (1.4, .01)$, $(a_i, b_i) = (10, 1)$ are used. These values of the hyperparameters reflect the prior belief that the two groups of logits will cluster about the values -1 and 1.4 . This prior belief is strong by giving small standard deviations to the parameters μ_1 and μ_2 . In addition, the relatively large shape parameter value $a_i = 10$ is chosen for τ_i^2 . This reflects the prior belief that the logits will cluster towards these two groups.

As described in Section 2.2, one introduces new parameters $\{g_i\}$, which indicate the group membership for each of the 32 questions and then uses the MCMC algorithm to simulate from the joint posterior distribution of the entire set of parameters. Figure 7(a) plots the posterior means of the logits $\{\theta_i\}$ against the observed logits for all of the questions. Figure 7(b) plots the ratios of the posterior standard deviations to the classical standard errors of the logits. With this selection of hyperparameter values, this model classifies most of the questions into one of the two categories. The posterior probabilities of the $\{g_i\}$ were close to 0 or 1 for 28 of the 32 questions. For a particular question, the posterior mean shrinks the observed logit towards the mean logit for the group for which that question has been classified. The most interesting aspect of this figure is the behavior of the posterior standard deviations in the bottom graph. The usual standard error of the observed logit is given by $\sqrt{y_i^{-1} + (n_i - y_i)^{-1}}$, where y_i and n_i are the number correct and sample size, respectively, for the i th question. Generally, the posterior standard deviations are smaller than these standard errors, reflecting the added precision from combining the data from related questions. However, the posterior standard deviations of the questions with group membership probabilities not close to 0 or 1 are significantly larger than the classical standard errors. These particular questions are not easily classified into one of the two groups

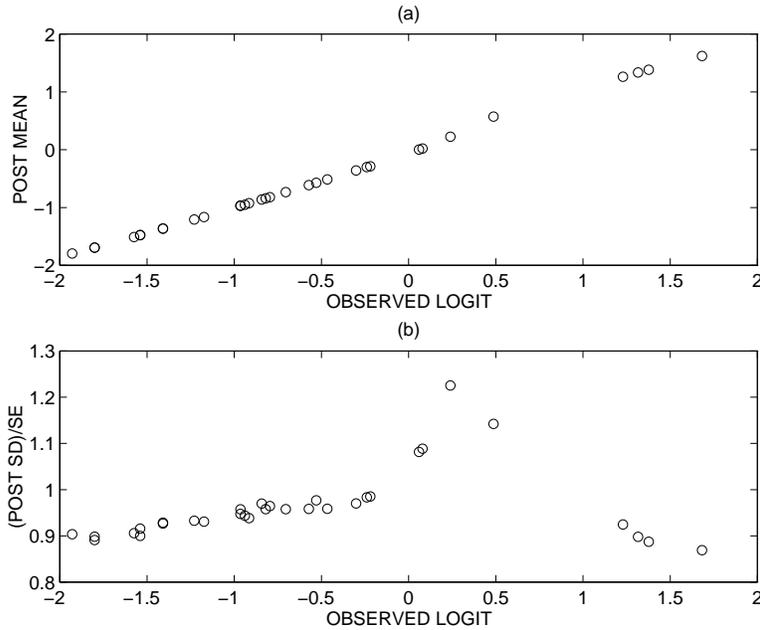


Figure 7: Plots of posterior means of the logits (a), and ratios of the posterior standard deviations to the classical standard errors (b) for math placement data set.

and the relatively large posterior standard deviations reflect the conflict in this classification.

4.2.3 Checking the second-stage prior

One way to investigate the suitability of the “two-groups” assumption in the partial exchangeable model is to consider a mixture prior for the values of the hyperparameters $\{m_i\}$. The values of m_1 and m_2 give the prior locations for the two groups and, as the distance $|m_1 - m_2|$ approaches zero, the two-groups model reduces to the basic exchangeable model. A prior distribution was chosen that placed equal probabilities on the seven pair of values $(m_1, m_2) = (-1, 1.4), (-.8, 1.2), (-.6, 1.0), (-.4, .8), (-.2, .6), (0, .4), (.2, .2)$. The values of the hyperparameters $(\gamma_1, \gamma_2, a_1, a_2, b_1, b_2)$ were set to the same values $(.01, .01, 10, 10, 1, 1)$ that were used in the initial partial exchangeable fit. The simulation algorithm was rerun—the pair $(m_1, m_2) = (-1, 1.4)$ received a posterior probability of .31 and the

pair $(-0.8, 1.2)$ a probability of .62 and no other pair had a significant posterior probability. Thus it appears that the data support the partial exchangeable model over the basic model. In other words, there is some evidence for the data clustering into two groups.

Another concern in the above fitting procedure is the specification of the second-stage prior on the variance parameters τ_1^2 and τ_2^2 . The hyperparameters a_1 and a_2 affect the general location of these priors. As the value of a_i increases, the prior for τ_i^2 places more mass towards zero. A large value of a_i reflects the prior belief that that the logits in that particular group will cluster. To assess the choice of different values for a_1 and a_2 , a prior was selected that placed equal probabilities on the hyperparameter values in the set $\{(a_1, a_2), a_i = 1, 2, 4, 8, 16\}$. The remaining hyperparameter values were set to $(m_1, m_2, \gamma_1, \gamma_2, b_1, b_2) = (-1, 1.4, .01, .01, 1, 1)$. The posterior probabilities obtained by a simulation run are displayed in Table 2. The marginal posterior probabilities are also shown. We see that there is support from the data for hyperparameter values of a_1 between 2–4 and values of a_2 between 1–2. Since there is a larger group of difficult questions, there is more support (a larger value of a_i) for shrinkage of the logits in this group towards an average value.

	a_2					
a_1	1	2	4	8	16	
1	.04	.04	.02	.01	.01	.11
2	.10	.09	.05	.02	.02	.28
4	.18	.16	.08	.03	.01	.46
8	.06	.05	.03	.01	.00	.15
16	.00	.00	.00	.00	.00	.00
	.38	.34	.17	.07	.04	1

Table 2: Posterior probabilities of variance hyperparameters a_1 and a_2 for partial exchangeable model.

5 Conclusions

This paper has presented a general method of robustifying a Bayesian hierarchical model. By the use of discrete mixtures, one can criticize the choice of prior and link function in a hierarchical generalized linear model. We have shown how the mixture approach can also be used to build outlier and partial exchangeable models. These models are particularly relevant when a large number of parameters are simultaneously being estimated and the prior belief of exchangeability may have to be questioned.

One attractive feature of the proposed mixture model perturbation approach is that it requires straightforward elaborations of standard MCMC methods for fitting exchangeable models. Indeed, this methodology can be applied to other models that share the same basic hierarchical structure and choice must be made amongst a set of contending models. In future research, we plan to explore the usefulness of this approach in the context of alternative classical and Bayesian approaches for selecting models.

References

- ALBERT, J. H. (1988), "Computational methods using a Bayesian hierarchical generalized linear model," *Journal of the American Statistical Association*, **83**, 1037-1045.
- ALBERT, J. H. and CHIB, S. (1993), "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, **88**, 669-679.
- ARANDA-ORDAZ, F. J. (1981), "On two families of transformations to additivity for binary response data," *Biometrika*, **68**, 357-363.
- BERGER, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, New York: Springer Verlag.
- BERGER, J. O. and DELAMPADY, M. (1987), "Testing precise hypotheses," *Statistical Science*, **3**, 317-352.
- CARLIN, B. P. and POLSON, N. G. (1991), "Inference for nonconjugate Bayesian methods using the Gibbs sampler," *Canadian Journal of Statistics* **4**, 399-405.
- CARLIN, B. P. and CHIB, S. (1995), "Bayesian model choice via Markov chain Monte Carlo," *Journal of the Royal Statistical Society, B*, **57**, 473-484.

- CHIB, S. and GREENBERG, E. (1995), "Understanding the Metropolis-Hastings algorithm," *American Statistician*, 49, 327-335.
- CONSONNI, G. and VERONESE, P. (1995), "A Bayesian method for combining results from several binomial experiments," *Journal of the American Statistical Association*, **90**, 935-944.
- DELLAPORTAS, P. and SMITH, A. F. M. (1993), "Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling," *Applied Statistics*, 42, 443-459.
- EMERSON, J. D. and STOTO, M. A. (1983), "Transforming data," in *Understanding Robust and Exploratory Data Analysis*, eds. D. C. Hoaglin, F. Mosteller and J. W. Tukey, New York: John Wiley.
- GASTWIRTH, J. L. and COHEN, M. L. (1970), "Small sample behavior of some robust linear estimators of location," *Journal of the American Statistical Association*, **65**, 946-973.
- GAVER, D. P., DRAPER, D., GOEL, P. K., GREENHOUSE, J. C., HEDGES, L. V., MORRIS, C. N. and WATERNAUX, C. (1992), *Combining Information: Statistical Issues and Opportunities for Research*, Washington: National Academy Press.
- GELFAND, A. E. and SMITH, A. F. M. (1990), "Sampling based approaches to calculating marginal densities," *Journal of the American Statistical Association*, **85**, 398-409.
- GELFAND, A. E., HILLS, S. E., RACINE-POON, A., SMITH, A. F. M. (1990), "Illustration of Bayesian inference in normal models using Gibbs sampling," *Journal of the American Statistical Association*, **85**, 972-985.
- GEORGE, E. I. (1986), "Minimax multiple shrinkage estimators," *Annals of Statistics*, **14**, 188-205.
- GUTTMAN, I. and SCOLLNIK, D. P. M. (1994), "An index sampling algorithm of a class of model selection problems," *Communications in Statistics* **23**, 323-339.
- KASS, R. E. and RAFTERY, A. E. (1995), "Bayes factors," *Journal of the American Statistical Association*, 90, 773-795.
- KASS, R. E., TIERNEY, L. and KADANE, J. B. (1988), "Asymptotics in Bayesian computation," in *Bayesian Statistics 3*, eds. J. Bernardo, M. DeGroot, D. V. Lindley and A. F. M. Smith, Oxford U. K.: Oxford University Press.
- KASS, R. E. 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.
- LINDLEY, D. V. and SMITH, A. F. M. (1972), "Bayes estimates for the linear model," *Journal of the Royal Statistical Society*, Series B, **34**, 1-41.

- LU, W. (1992), "Empirical and hierarchical estimation of several means in the natural exponential family," Technical report, Center for Statistical Sciences, The University of Texas at Austin.
- MORRIS, C. (1983a), "Parametric empirical Bayes inference: theory and methods," *Journal of the American Statistical Association*, **78**, 47-65.
- MORRIS, C. (1983b), "Natural exponential families with quadratic variance functions: statistical theory," *Annals of Statistics*, **11**, 515-529.
- O'HAGAN, A. (1988), "Modelling with heavy tails," *Bayesian Statistics 3*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, Oxford U. K.: Oxford University Press.
- RAFTERY, A. E. (1993), "Approximate Bayes factors and accounting for model uncertainty in generalised linear models," Technical Report 255, Department of Statistics, University of Washington.
- TIERNEY, L. (1994), "Markov chains for exploring posterior distributions," *Annals of Statistics*, **22**, 1701-1762.
- TIERNEY, L., and KADANE, J.B. (1986), "Accurate approximations for posterior moments and marginal densities," *Journal of the American Statistical Association*, **81**, 82-86.
- TSUTAKAWA, R. K., SHOOP, G. L., and MARIENFELD, C. J. (1985), "Empirical Bayes estimation of cancer mortality rates," *Statistics in Medicine*, **4**, 201-212.
- TUKEY, J. W. (1977), *Exploratory Data Analysis*, New York: Addison-Wesley.