Title

Predictive Model Selection for Repeated Measures Random Effects Models Using Bayes Factors

Permalink

https://escholarship.org/uc/item/3dr3f20t

Authors

R. E. Weiss
Y. Wang
J. Ibrahim

Publication Date

2011-10-25

Predictive Model Selection for Repeated Measures Random Effects Models Using Bayes Factors

Robert E. Weiss
Yan Wang
Department of Biostatistics, UCLA School of Public Health,
Los Angeles CA 90095-1772, USA

Joseph G. Ibrahim
Department of Biostatistics, School of Public Health,
Harvard University, 677 Huntington Ave., Boston, MA 02115 USA

September 5, 1996

Abstract

The random effects model (REM) fit to repeated measures (RM) data is an extremely common model and data structure in current biostatistical practice. Modern data analysis often involves the selection of models within broad classes of pre-specified models, but for models beyond the generalized linear model, few model selection tools have been developed. In a Bayesian analysis, Bayes factors are the natural tool to use to explore these classes of models. In this paper we develop a predictive approach for specifying the priors of a RM REM with emphasis on selecting the fixed effects. The advantage of the predictive approach is that a single predictive specification is used to specify priors for all models considered. The methodology is applied to a pediatric pain data analysis.

Key Words: Bayesian Data Analysis, Elicitation, Hierarchical Model, Prediction, Variable Selection.

1 Introduction

Many longitudinal studies are designed to investigate changes over time in a characteristic which is measured repeatedly for each study participant. For example, in medical studies, measurements such as blood pressure, cholesterol level, or lung volume may be taken on each individual at different time points and possibly under changing experimental conditions. There has been some recent literature on the development of statistical models for analyzing such data. Perhaps the most common type of model for repeated measurements is the random effects linear model introduced by Laird and Ware (1982). An early Bayesian discussion on estimation for some random effects models is given by Broemeling (1985, chapter 4).

A major issue in Bayesian model selection is the method of quantification or elicitation of the prior input, that is, the specification of the prior distributions. In this article, we present a Bayesian approach for model selection in the random effects model. Recently, there have been several articles addressing model selection from a Bayesian viewpoint. These include Mitchell and Beauchamp (1988), George and McCulloch (1993), Gelfand and Dey (1993), Ibrahim and Laud (1994), Laud and Ibrahim (1995, 1996), George, McCulloch, and Tsay (1996), and Raftery (1993). For the logistic regression model with normal random effects, Karim and Zeger (1992) suggest a method similar to Aitken's (1991) posterior bayes factors. The Bayesian approach to model selection is straightforward in principle. One quantifies the prior uncertainties via probabilities for each model under consideration, specifies a prior distribution for the parameters in each model, and then uses Bayes theorem to calculate posterior model probabilities. In addition to the computational issues, there are other difficulties in carrying out such a plan. Specifying meaningful prior distributions for the parameters in each model is an arduous task requiring contextual interpretations of a large number of parameters. A need arises then to look for some useful, automated specifications. Reference priors can be used in many situations to address this. Often, however, they lead to ambiguous posterior probabilities, and require problem-specific modifications such as those in Smith and Spiegelhalter (1980). Recently, Berger and Pericchi (1996) have proposed a set of measures they call "intrinsic Bayes factors" that provide a generic solution to the ambiguity problem. However, reference priors exclude the use of any real prior information one may have.

To overcome such difficulties, Ibrahim and Laud (1994), and Laud and Ibrahim (1995, 1996) advocate priors based on observables for model selection in the linear model by adapting the philosophy in Geisser (1993). Specifying priors based on observables has been advocated by many including Geisser (1971), Kadane (1980), Oman (1985), and Winkler (1980). The prior specification proposed by Ibrahim and Laud (1994) and Laud and Ibrahim (1995, 1996) begins by using all prior knowledge to elicit a prior point prediction for the observable Y, denoted by μ_0, and a scalar c_0 which quantifies the fraction of information in this guess relative to the information to be collected in the rest of the experiment. Then, (μ_0, c_0), along with the covariate matrix X_m for model m, are used to specify an automated parametric informative prior for the regression coefficients. There are several ways of eliciting a prior prediction μ_0. For example, having previous experience with such studies, the investigator may elicit μ_0 by using expert opinion and/or case-specific information available for each of the n cases in the current study. Also, if a previous study was conducted with the same or similar covariates as the current study, the investigator may take μ_0 to be the raw data vector or the vector of fitted values from the previous study. The motivation behind these specifications is that the investigator often has prior information on the observables from expert opinion, case-specific information on the subjects in the current study, or from similar past studies measuring the same covariates and response variables. This prior information is often quantifiable in the form of a vector of prior predictions for the response variables in the current study. This prior prediction then yields an automated prior specification for the regression parameters arising from the various models. As mentioned by Geisser (1993), it is often easier to think of observables rather than parameters when specifying prior input since there are so many parameters arising from the different models, and all have different physical meaning. Generally, we call a prior a *predictive prior* if it is based on a (possibly point) prediction for the observables. In particular, the prior distributions of Ibrahim and Laud (1994) and Laud and Ibrahim (1995, 1996) are predictive prior distributions.

Here, we adapt the approach of Ibrahim and Laud (1994) and Laud and Ibrahim (1995, 1996) for specifying the predictive prior distributions for the random effects model. Using the marginal likelihood as specified by Laird and Ware (1982), we specify predictive prior distributions for the "fixed" effects. Moreover, we also specify informative prior distributions for the variance components arising in the marginal likelihood. The rest of the article is organized as follows. In the next section, we briefly review the random effects model, and describe the prior elicitation for it. In Section 3, we derive the posterior and predictive distributions of interest, and discuss computational techniques for these models. Section 4 gives a representative data analysis. We conclude the article with a brief discussion.

2 Predictive Priors for The General Random Effects Model

2.1 Model and Notation

Suppose we have n_i observations on case i for $i = 1, \ldots, n$; n is the number of cases or subjects, while $N = \sum_{i=1}^{n} n_i$ is the total number of observations. Let α denote a $p \times 1$ vector of unknown population parameters and X_i be a known $n_i \times p$ matrix of covariates linking the i^{th} response Y_i to α. We call Y_i a case and a single element of Y_i an observation. In addition, let β_i denote a $q \times 1$ vector of unknown random effects, and let Z_i be a known $n_i \times q$ covariate matrix linking β_i to Y_i. The general random effects model for repeated measures can now be written as

$$Y_i = X_i\alpha + Z_i\beta_i + \epsilon_i \tag{2.1}$$

where $\epsilon_i \sim N_{n_i}(0, R_i)$, and $N_p(\mu, \Sigma)$ denotes the p dimensional multivariate normal distribution with mean μ and covariance matrix Σ. Here, R_i is an $n_i \times n_i$ positive definite matrix, which we assume to be of the form $R_i = \sigma^2 I$ throughout. Also, we assume that the random effects β_i are a priori independent and identically distributed (iid) given the parameters (σ^2, D), and $\beta_i \sim N_q(0, \sigma^2 D)$, where D is a $q \times q$ positive definite matrix.

Following Laird and Ware (1982), we integrate out the random effect β_i from (2.1) and work with the marginal likelihood of (2.1) for making inferences. A straightforward calculation shows that

$$\left[Y_i \mid \alpha, \sigma^2, D\right] \sim N_{n_i}(X_i\alpha, \sigma^2 V_i)$$

where

$$V_i = (I + Z_i D Z_i^t) \ . \tag{2.2}$$

Thus, the marginal likelihood for subject i is given by

$$f(Y_i \mid \alpha, \sigma^2, D) = (2\pi)^{-n_i/2} |\sigma^2 V_i|^{-1/2} \exp\left\{-(Y_i - X_i\alpha)^t V_i^{-1}(Y_i - X_i\alpha)/2\sigma^2\right\} \ . \tag{2.3}$$

We assume that observation vectors for each case are independent. The likelihood based on all observations is given by the product of (2.3) over i running from 1 to n.

We specify a multivariate normal prior for $[\alpha \mid \sigma^2, D]$, which will have hyperparameters similar to those of Ibrahim and Laud (1994). We will specify independent priors for σ^2 and D. For σ^2, we will specify an inverse Gamma distribution, and for D, we will specify either a Wishart or other convenient priors as prior information requires. Letting $\pi(\cdot)$ be a generic prior density, with different densities distinguished by their arguments, we have

$$\pi(\alpha, \sigma^2, D) = \pi(\alpha \mid \sigma^2, D) \ \pi(\sigma^2) \ \pi(D) \ .$$

Let $\mathcal{M}$ denote the model space and let m be a specific member in $\mathcal{M}$. Define $Y = (Y_1^t, \ldots, Y_n^t)^t$, and let $X_m = (X_{1,m}^t, \ldots, X_{n,m}^t)^t$ denote the $N \times p_m$ matrix of covariates for the current experiment under model m. Define $Z_m = \text{diag}(Z_{1,m}, \ldots, Z_{n,m})$ as the $N \times nq_m$ block diagonal matrix of covariates for the random effects. Under model m, let D_m be the covariance matrix of the q_m random effects, $\sigma_m^2 I$ is the covariance matrix of ϵ_i, and $V_m = \text{diag}(V_{1,m}, \ldots, V_{n,m})$, where each $V_{i,m} = (I + Z_{i,m} D_m Z_{i,m}^t)$. Finally, $\alpha^{(m)}$ is the α vector under model m. Under model m, we have

$$\left[Y \mid m, \alpha^{(m)}, \sigma_m^2, D^{(m)}\right] \sim N_N(X_m \alpha^{(m)}, \sigma_m^2 V_m) \ .$$

2.2 Prior for α

The prior specification begins by defining $\tilde{n}$ design matrices and vectors of observables to obtain the hyperparameters for the prior for $\left[\alpha^{(m)} \mid \sigma_m^2, D_m\right]$. Each covariate matrix and vector of observables will have $\tilde{n}_j$ rows for $j = 1, \ldots, \tilde{n}$; each row corresponds to a single observation on a single, possibly hypothetical, case. The number of cases $\tilde{n}$ may be chosen for convenience. Later, we introduce a prior parameter to adjust the weight given to the constructed prior distribution. Let $\tilde{X}_{j,m}$, $j = 1, \ldots, \tilde{n}$, denote an $\tilde{n}_j \times p_m$ matrix with the same set of covariates as X_m; that is, the columns of X_m and $\tilde{X}_{j,m}$ represent the same information or measurements about subjects, but there is not necessarily any correspondence between rows. For convenience, for $m \neq m'$, $\tilde{X}_{j,m}$ and $\tilde{X}_{j,m'}$ will often have equal columns when the corresponding covariates are the same. Similarly, let $\tilde{Z}_{j,m}$ denote an $\tilde{n}_j \times q_m$ matrix with the same set of covariates as Z_m, and let $\tilde{V}_m = \text{diag}(\tilde{V}_{1,m}, \ldots, \tilde{V}_{\tilde{n},m})$, where each $\tilde{V}_{j,m} = (I + \tilde{Z}_{j,m} D_m \tilde{Z}_{j,m}^t)$. Define $\tilde{N} = \sum_{j=1}^{\tilde{n}} \tilde{n}_j$ as the number of prior observations, $\tilde{X}_m = (\tilde{X}_{1,m}^t, \ldots, \tilde{X}_{\tilde{n},m}^t)^t$ and $\tilde{Z}_m = \text{diag}(\tilde{Z}_{1,m}, \ldots, \tilde{Z}_{\tilde{n},m})$ and let μ_0 denote an $\tilde{N} \times 1$ vector of prior point predictions for a response $\tilde{Y}$. We have then specified $\tilde{n}$ prior cases which we combine to produce a prior for $\left[\alpha^{(m)} \mid \sigma_m^2, \tilde{V}_m\right]$ of the form

$$\left[\alpha^{(m)} \mid \sigma_m^2, \tilde{V}_m\right] \sim N_{p_m}(\mu^{(m)}, c_0^{-1} \sigma_m^2 \Sigma_m) \tag{2.4}$$

where

$$\mu^{(m)} = (\tilde{X}_m^t \tilde{V}_m^{-1} \tilde{X}_m)^{-1} \tilde{X}_m^t \tilde{V}_m^{-1} \mu_0 \tag{2.5}$$

and

$$\Sigma_m = (\tilde{X}_m^t \tilde{V}_m^{-1} \tilde{X}_m)^{-1} \ . \tag{2.6}$$

The prior data $(\tilde{n}, \mu_0, \tilde{X}_m, \tilde{Z}_m)$ are potentially arbitrary and need not involve any of the covariate values of the current study or any other study. The vector μ_0 is fixed regardless of the model under consideration, and thus does not depend on m. The specification of $(\mu^{(m)}, \Sigma_m)$ depends entirely on μ_0, $\tilde{X}_m$, and $\tilde{Z}_m$. There are several semi-automatic ways of choosing μ_0. For example, if a previous experiment using the same covariates as the current study was conducted based on sample size n^* with response vector μ_0^* and covariate matrices X_m^* and Z_m^*, then we take $\tilde{n} = n^*$, $\tilde{X}_m = X_m^*$, $\tilde{Z}_m = Z_m^*$, and $\mu_0 = \mu_0^*$ and substitute these into (2.5) and (2.6) to obtain $(\mu^{(m)}, \Sigma_m)$. If the current study has a larger set of covariates than the previous study, then we can modify the prior as follows. Suppose we partition α as $\alpha = (\alpha_c, \alpha_n)$ where α_c represents the coefficients that are common to the previous study and α_n represents an s vector of new coefficients for covariates that were not included in the previous study. In this case, little prior information is available for α_n. Thus we can specify the prior in (2.4) – (2.6) for α_c and specify an independent $N_s(0, d_0^{-1}W)$ prior distribution for α_n, where W is a diagonal matrix and d_0 is a scalar parameter controlling the variance of α_n. Since we have no prior information regarding the new set of covariates, we typically take d_0 small. This prior for α_n thus reflects vague prior beliefs about α_n. This construction assumes an independence between the new covariates X_n and the old X_c, such that approximately $X_n^t X_c \approx 0$. More complicated constructions are the subject of ongoing research.

In (2.4), c_0^{-1} is a scalar quantifying the importance one wishes to attach to the prior guess μ_0. If c_0 is small, this reflects a lack of certainty about μ_0. In particular, if $c_0 = \tilde{N}^{-1}$, the prior is weighted similarly to a single observation; if $c_0 = \tilde{n}^{-1}$, the prior is weighted similarly to a single multivariate case. If c_0 is one, the prior is weighted as if it had $\tilde{n}$ multivariate cases; if $c_0 = \tilde{N}/N$ or $c_0 = \tilde{n}/n$, the prior and likelihood are weighted approximately equally.

If a previous study does not exist, then the investigator must rely on expert opinion and/or case specific information on the subjects in the current study in specifying $(\tilde{n}, \mu_0, \tilde{X}_m, \tilde{Z}_m)$. In these situations, convenient automatic specifications include taking $\tilde{n} = n$, $\tilde{X}_m = X_m$, and $\tilde{Z}_m = Z_m$. That is, we set $\tilde{n}$ to be the sample size of the current study, and $\tilde{X}_m$, $\tilde{Z}_m$ to be the covariate matrices of the current study. Such choices are natural in these settings. The investigator may elicit μ_0 directly or indirectly, depending on prior beliefs. A direct choice is $\mu_0 = 0$, which corresponds to taking a prior mean of 0 (i.e., no regression) for the fixed effects regardless of the choice of covariate matrices. Another direct choice for μ_0 is to take all components of μ_0 equal to the same value, such as the predicted mean response. Direct informative choices of μ_0 depend on the strength of the investigator's prior beliefs and the context of the study, and thus no general automated specification can be recommended here. However, in this setting, we recommend that the investigator examine several values of μ_0, including informative and noninformative, and conduct sensitivity analyses to study the behavior of the Bayes factors for each of these choices.

It sometimes occurs that the investigator does not have direct prior information on the response vector, but has indirect information from expert opinion on certain parameter values in the model. The prior information on the parameter values can be turned into a prior prediction for the response by projecting the elicited parameter vector into the response space and then taking μ_0 to be the resultant projection. Specifically, if α_0 is an elicited value for a specific parameter vector α and X_{μ_0} is some specified covariate matrix, then $X_{\mu_0}\alpha_0$ is a vector in the response space. Therefore, we take $\mu_0 = X_{\mu_0}\alpha_0$. This indirect specification of μ_0 via α_0 is quite useful since once μ_0 is obtained, we can immediately obtain prior parameters for $m \in \mathcal{M}$ via (2.5). The projection of α_0 onto the response space automates the prior specification and therefore is an important part of the elicitation procedure. On the other hand, if α_0 is not projected into the response space, then we would need to manually elicit parameter vectors for all $m \in \mathcal{M}$, which is precisely the task we are trying to avoid. In Section 4, we demonstrate an indirect elicitation scheme for the response by using expert opinion in an actual study. The method of elicitation of μ_0 depends on the context of the study and the prior information available to the investigator. We have proposed here some very general and flexible methods that are quite useful under a variety of situations.

2.3 Priors for the Variance Parameters

We now specify priors for (σ_m^2, D_m), which we assume independent a priori. We drop the subscripts m until the end of the section, where we discuss the case $q_m > 1$. It is important to specify a proper prior for D, since an improper prior for D may lead to an improper posterior for D (Hobert and Casella 1996), regardless of the prior for σ^2. We recommend an inverse gamma prior for σ^2

$$\sigma^2 \sim IG(\frac{b}{2}, \frac{c}{2}), \tag{2.7}$$

with density proportional to $(\sigma^2)^{-(b/2+1)} \exp(-c/(2\sigma^2))$. For D, we use a convenient prior, such as a Wishart prior when $q_m > 1$ and a gamma $G(\delta_0/2, \gamma_0/2)$ when $q_m = 1$, with density proportional to $D^{(\delta_0/2-1)} \exp(-\gamma_0 D/2)$. Because of the nature of model (2.1), there is no prior for D that permits a closed form expression of $f(Y|m)$. In contrast, choosing a conjugate prior for σ^2 is important and simplifies the computations.

Given the prior specification, there are many ways to elicit prior information to select (b, c) and (δ_0, γ_0). Most of the following assumes that $q_m = 1$, although they can be adapted to $q_m > 1$, which we discuss briefly below. Two pieces of prior information can be used to solve for δ_0 and γ_0 in the priors for σ^2 and D. We list several reasonable approaches to elicitation: (a) A point estimate for σ^2 or D can be elicited. For example, the posterior mean or mode or the maximum likelihood estimate from a previous experiment or an elicited point estimate can be used. For a purely elicited point estimate, it may be easier to do the elicitation for a standard deviation, σ or $D^{1/2}$. Another approach, which we used in our example, is to elicit a range r for the residuals ϵ_{ij}, and take $r/4$ as a point estimate for σ. A similar mechanism can be used for D when it is univariate. The point estimate can be plugged in as either the prior mean or mode. (b) We can elicit δ_0 directly, since it can be interpreted as a prior number of degrees of freedom, possibly some fraction of the number of degrees of freedom from the previous experiment or the prior sample size that we wish to give our prior information. (c) The investigator may specify a probability statement such as

$$P(\sigma^2 < a) = .95$$

where a is a specified constant and similarly for D. (d) A prior variance for σ^2 or σ may be specified. Choosing γ_0 to be of the form $\gamma_0 = \lambda(\tilde{n} - p_m)^{-1}$, where p_m is the rank of X_m, results in a decrease of the prior mean and precision of τ as the number of predictors p_m in model m increases.

When Z_m is a column of ones, model (2.1) is called a random intercept model. Assuming $D \gg \sigma^2$, then one quarter of the range r of the Y_i is a reasonable point estimate for $D^{1/2}$. If D and σ^2 are comparable, then $r/4 \approx (D + \sigma^2)^{1/2}$ and r can be reduced slightly either formally or informally to adjust for the variation due to σ^2.

For general D_m, with $q_m > 1$, we specify a q_m dimensional Wishart prior with ν_0 degrees of freedom and prior mean matrix $\nu_0^{-1}\Psi_{0m}$. We denote the prior distribution by $D_m \sim W_{q_m}(\nu_0, \Psi_{0m})$ with density $p(D_m) \propto |D_m|^{(\nu_0 - q_m + 1)/2} \exp\{-1/2 \, \mathrm{tr}(\Psi_{0m}^{-1} D_m)\}$. If a previous study exists with data $(n^*, \mu_0^*, X_m^*, Z_m^*)$ then one can take $\nu_0 = n^*$ and choose the prior mean matrix to be D_m^*, where D_m^* is the maximum likelihood estimate, posterior mode, or posterior mean of D_m using $(n^*, \mu_0^*, X_m^*, Z_m^*)$. If no previous experiment exists, one can choose Ψ_0 to be a diagonal matrix and choose ν_0 small to reflect vague prior beliefs. The prior for σ^2 would still be specified as before. Possibly the most common model when $q_m = 2$ would be a random intercept and slope model. The prior for D could be specified as Wishart with a diagonal Ψ_0. The prior mean for $D_{11}^{1/2}$ could come from an estimate of the range of Y_i at time $t = 0$ divided by 4 and $D_{22}^{1/2}$ could come from an estimate of the range of the slopes, again divided by 4.

3 Posteriors and Computations

In this section we first give the posteriors for α, σ^2, D given the likelihood (2.3) and priors (2.4) – (2.6), (2.7), and the prior for D of the previous section and indicate how we calculate

$$f(Y|m) = \int f(Y|\alpha, D, \sigma^2, m) p(\alpha|\sigma^2, D, m) p(\sigma^2|m) p(D|m) \, d\alpha \, d\sigma^2 \, dD \quad (3.8)$$

given the model specification. Formally, $D = D^{(m)}$, $X = X_m$, $Z = Z_m$, $\Sigma = \Sigma_m$, $\alpha = \alpha^{(m)}$, $\sigma^2 = \sigma^2_m$ and $\mu = \mu^{(m)}$ depend on the model specification m, but the dependence on m is omitted in this section to reduce the complexity of the notation.

A posteriori, $[\alpha|\sigma^2, D, Y, m]$ is normal, and $[\sigma^2|D, Y, m]$ is inverse Gamma, but $p(D|Y, m)$ is not of standard form. Therefore, in (3.8), we integrate out α and σ^2 in closed form and use numerical integration to integrate out D. The posterior of $[\alpha|\sigma^2, D, Y, m]$ is

$$\left[\alpha \mid \sigma^2, D, Y, m\right] \sim N_{p_m}(\nu, \sigma^2 \Lambda) \quad (3.9)$$

where

$$\nu = (X^t Q_Z X + c_0 \Sigma^{-1})^{-1} (X^t Q_Z Y + c_0 \Sigma^{-1} \mu) \, ,$$

and

$$\Lambda = (c_0 \Sigma^{-1} + X^t Q_Z X)^{-1} \, ,$$

where

$$Q_Z = (Z(I \otimes D) Z^t + I)^{-1} \, .$$

The posterior of σ^2 is

$$\left[\sigma^2 \mid D, Y, m\right] \sim IG(\frac{N+b}{2}, \frac{Q(Y)}{2}) \, , \quad (3.10)$$

where

$$Q(Y) = (Y - X\alpha^*)^t Q_Z (Y - X\alpha^*) + (\alpha^* - \mu)^t c_0 \Sigma^{-1} \Lambda X^t Q_Z X (\alpha^* - \mu) + c$$

and $\alpha^* = (X^t Q_Z X)^{-1} X^t Q_Z Y$ is the estimate of α based on the likelihood alone. After integrating out α and σ^2 algebraically, the remaining integral is

$$f(Y|m) = \int f(Y, D|m)\, dD \tag{3.11}$$

and

$$f(Y, D) = \frac{(c)^{b/2} |\Lambda|^{1/2} |Q_Z|^{1/2} \Gamma((N+b)/2) p(D)}{\pi^{N/2} |c_0^{-1} \Sigma|^{1/2} (Q(Y))^{(N+b)/2} \Gamma(b/2)} .$$

The case of $q_m > 1$ is briefly discussed in section 5. When $q_m = 1$, the integral in (3.11) is one dimensional, and we use standard one dimensional numerical integration tools to calculate $f(Y|m)$ after plotting $f(Y, D|m)$ and finding the value $D_{\max}$ that maximizes $f(Y, D|m)$ as a function of D. Since the values are quite small, we divide $f(Y, D|m)$ by a suitable constant such as $f(Y, D_{\max}|m)$ before executing further calculations. To further simplify calculations, we often first calculated $f(Y|m = 0)$, where the model $m = 0$ is the model with $\alpha \equiv 0$. Depending on context, we may report any of $f(Y|m)$; $f(Y|m)/f(Y|m^*)$ and $f(Y|m^*)$, where m^* is some default, possibly null model; or $f(Y|m = k+1)/f(Y|m = k)$, for appropriate choices of k. We also perform a sensitivity analysis with respect to the specifications μ_0 and c_0. The advantage of reporting $f(Y|m)$ or $f(Y|m)/f(Y|m^*)$ is that Bayes factors between different prior specifications may also be calculated; since this is not of great interest in our example, we report Bayes factors of the form $f(Y|m = k+1)/f(Y|m = k)$.

4 Data Analysis

We illustrate our methodology on a repeated measures data set from Pediatric Pain. The response for each of the $n = 58$ children with complete data is a 4-vector of the log times that the children were able to immerse an arm in cold water. The covariance structure is a random intercept model for all models with $q_m \equiv 1$ and Z_m a 4-vector of ones. The children are classified into one of two groups, attenders or distracters (A or D), depending on their style of coping (CS) with the pain of the cold. Attenders pay attention to their arm in the water and the experimental apparatus during the trial; distracters think about topics unrelated to the trial. Prior to the fourth trial, an intervention occurs. The intervention (treatment or TMT) is a short counseling session. Three types of counseling are given: counseling to attend (A); counseling to distract (D); or a null counseling without instructions (N). If TMT has any affect, then a priori, CS and TMT were presumed to interact. Interest lies in whether the two CS groups have different baseline response times, and given that CS has an effect, whether treatment effects the response time.

We consider 3 major models for the data, M_1, M_2, M_8. Model M_1 has an intercept; M_2 also includes a CS baseline effect; and M_8 also includes both the TMT main effects and CS by TMT interaction effects. We are particularly interested in comparing M_2 to M_1 and in comparing M_8 to M_2. The subscript m on M_m indicates the number of columns in the X_i predictors. In M_8, X_i is 4×8. The four rows correspond to the four trials for each child. The first column of X_i is a vector of ones for the intercept. The second column is either a vector of zeros or ones for the coping style (CS) baseline effect. Columns 3-8 are all zeros, except for a single 1 in the last row, indicating which of the $6 = (2 * 3)$ CS*TMT combinations AA, AD, AN, DA, DD, or DN the child belongs to. Treatment was randomized, but came out nearly balanced. The CS is split nearly 50-50 in the sample. The matrix $V_m \equiv V$ is the $58 * 4$ by $58 * 4$ block diagonal covariance matrix of V_i's and is the same for all three models. The X_m are the $58 * 4$ by m matrix of stacked X_i matrices for each model.

Table 1 contains $B(8, 2) = f(Y|M_8)/f(Y|M_2)$ and $B(2, 1) = f(Y|M_2)/f(Y|M_1)$ for several choices of priors for $[\alpha|\sigma^2, D]$. For our priors, we took $\tilde{X}_m = (\tilde{X}_{1,m}^t, \ldots, \tilde{X}_{N,m}^t)^t = X_m$. The columns correspond to five choices for c_0 in the prior (2.4) for $[\alpha^{(m)}|\sigma_m^2, \tilde{V}_m]$. Rows are discussed beginning with the next paragraph. Moving from left to right in Table 1, the prior contains decreasing information, with $c_0 = N/8$ producing a very informative prior with mean $\mu^{(m)}$; this choice gives the prior a weight comparable to a likelihood with $(N/8) * \tilde{n} = 29 * 58$ data cases, since $N = 4n$ and $n = \tilde{n} = 58$. In comparison, the data always represents $n = 58$ cases with 4 observations each for a total of $N = 232$ observations. This informative prior was chosen to illustrate how strong the prior would need to be to produce Bayes factors that strongly favored a treatment effect. The setting $c_0 = 1$ represents equal weight between data and prior, 58 cases each. This still represents substantially stronger prior information than actually was held at the beginning of the study. The setting $c_0 = 8/N$ gives the prior a weight corresponding to two complete data cases or 8 observations. This setting is approximately what was believed was the actual strength of our initial prior information. Two diffuse but still informative settings for c_0 are also considered. Setting $c_0 = 1/N$ gives the prior the weight of a single observation; and $c_0 = (8N)^{-1}$ gives the prior the weight of $1/8^{\text{th}}$ of an observation.

The outer rows in Table 1 correspond to different choices for the prior prediction μ_0 in (2.5). For all rows, the prior prediction was in the form $\mu_0 = X_{\mu_0} w$ with $X_{\mu_0} = X_8$, with various choices for w, labeled w_1 up to w_7 given in Table 2. Note that μ_0 is our prior point prediction and it stays the same for all models M_k and that w_k has 8 elements regardless of the model – by including additional variables such as age in months, we could have had X_{μ_0} with 9 or more columns. The choice $w_1 = (w_{1j})^t$ is an actual elicitation specified by the first author based on information supplied by the original investigator. Along with the choice $c_0 = 8/N$, this represents the actual single choice of priors for this data analysis. The choice of w_1 represents a prior with a prior mean of $w_{11} = \log(16 \text{ seconds})$ for attenders at baseline and since $w_{12} = \log(1.2)$, a 20% longer baseline pain tolerance for distracters over attenders. There is a prior belief that attenders taught to attend (AA group) and distracters taught to distract (DD) will also have an improvement of 20%; this corresponds to values $w_{13} = w_{17} = \log(1.20)$. There is a prior belief that non-matching treatments, attenders taught to distract (AD) and distracters taught to attend (DA) would reduce tolerance by 10% corresponding to coefficients of $\log(.9)$. The prior belief for the null counseling session is that they would not change the pain tolerance, so the coefficients in columns AN and DN are zero.

Our other choices for w can be considered both as sensitivity analysis and as illustration of other reasonable choices for the prior specification. The vector w_2 is a more extreme version of w_1. The choice w_3 illustrates use of the estimated posterior mean from a flat prior based on a Gibbs sample of size 1000 (see Table 3). Rows w_4 and w_5 are variants of w_1 and w_2 that support M_2, they have the treatment parameter means $w_{k3}, \ldots, w_{k8}$ set to zero. Row w_6 supports model M_1, with only the intercept having a non-zero value, and w_7 illustrates the use of the zero vector. Prior inputs w_6 and w_7 represent what we expect to be common prior specifications under our priors, a vector μ_0 of constant predictions (w_6), as might happen in the absence of any knowledge about subjects or covariates and the important special case where μ_0 is the vector of zeros (w_7), especially useful for data that is somehow centered. The parameters of the priors for D and σ^2 are $\delta_0 = 4$, $\gamma_0 = 1.6$, $b = 4$, $c = .68$ corresponding to point estimates $D = 2.5$ and $\sigma^2 = .17$, the posterior means based on flat priors and model M_8. The values of the different w_k are tabulated in Table 2. Two other priors for $p(\sigma^2)$ and $p(D)$ were tried without substantially different results and are discussed briefly later.

Several conclusions follow from the analysis. Generally, for the very diffuse priors with $c_0 = 1/(8N)$ or $c_0 = 1/N$, the choice of w does not matter. These values of c_0 generally favor model M_1, modestly over M_2 and strongly over model M_8. For both of these choices of c_0, we would conclude that there is no treatment effect, and that the data were equivocal on whether there was even a baseline effect or not. For $c_0 = 8/N$, the data generally favor model M_2 over M_8 with Bayes factors $B(8, 2)$ of around .01 to .05; also for $c_0 = 8/N$, the data slightly favor M_2 over M_1 with Bayes factors of around 2.5. Again we conclude no treatment effect, and equivocal results on whether there is a baseline coping style effect, this time slightly favoring the possibility of an effect. For the informative priors with $c_0 = 1$ we do get Bayes factors favoring a treatment effect; for example, for $\mu_0 = X_8 w_1$ we have $B(8, 2) = 18.71$, lending support for a non-zero treatment effect. For $c_0 = 1$ and $c_0 = N/8$, the choice of w matters. The choices of w_2 and

	Bayes	c_0				
prior	Factor	$N/8$	1	$8/N$	$1/N$	$1/(8N)$
w_1	B(8,2)	20.3	18.7	.0438	1.07×10^{-4}	2.16×10^{-7}
	B(2,1)	3.92	4.73	2.50	.948	.338
w_2	B(8,2)	1.87×10^{-13}	1.83×10^{-7}	.011	8.97×10^{-5}	2.11×10^{-7}
	B(2,1)	.509	1.74	2.36	.941	.338
w_3	B(8,2)	1250	173.	.0513	1.09×10^{-4}	2.16×10^{-7}
	B(2,1)	117.	31.0	2.87	.965	.339
w_4	B(8,2)	1.13	3.88	.0392	1.06×10^{-4}	2.15×10^{-7}
	B(2,1)	3.85	4.66	2.50	.948	.338
w_5	B(8,2)	1.14	4.21	.0398	1.06×10^{-4}	2.15×10^{-7}
	B(2,1)	.564	1.83	2.36	.941	.338
w_6	B(8,2)	1.12	3.68	.0388	1.06×10^{-4}	2.15×10^{-7}
	B(2,1)	1.04	1.99	2.31	.938	.338
w_7	B(8,2)	.956	.487	.0207	9.36×10^{-5}	2.12×10^{-7}
	B(2,1)	.995	.943	.896	.752	.328

Table 1: Bayes factors $B(8,2)$ in favor of M_8 against M_2 and $B(2,1)$, the Bayes factors in favor of M_2 against M_1, for different choices of $\mu_0 = X_8 w_k$ and c_0.

coefficient	intercept	CS	AA	AD	AN	DA	DD	DN
w_1	$\log 16$	$\log 1.2$	$\log 1.2$	$\log 0.9$	0	$\log 0.9$	$\log 1.2$	0
w_2	$\log 16$	$\log 2.4$	$\log 2.4$	$\log 0.45$	0	$\log 0.45$	$\log 2.4$	0
w_3	3.12	0.42	0.091	0.036	-0.10	-0.25	0.39	-0.29
w_4	$\log 16$	$\log 1.2$	0	0	0	0	0	0
w_5	$\log 16$	$\log 2.4$	0	0	0	0	0	0
w_6	$\log 16$	0	0	0	0	0	0	0
w_7	0	0	0	0	0	0	0	0

Table 2: Vectors of coefficients w_k for input to prior for α. Coefficients for w_4 are rounded to 2 significant digits. Actual calculations used all digits output from the estimated posterior mean based on a Gibbs sample of size 2001 using a flat prior.

w_3, in particular, behave differently from the other choices, with w_2 strongly against M_8 and w_3 strongly favoring M_8. To obtain a large Bayes factor in favor of a non-zero treatment effect, we must take $\mu_0 = X_8 w_3$, where we have used the data more than once.

Two other choices for the priors for σ^2 and D were considered. The substantive conclusions regarding the relative preference of the data for model M_2 were the same as in Table 1, unless as in Table 1, the prior was overly diffuse or strongly favored M_8. One prior included an ad hoc choice of parameters, $b = c = \delta_0 = \gamma_0 = 2$. The other prior was chosen with $b = 4$, and $c = .4356$, corresponding to an elicited point estimate for σ^2 of $.33^2$, where $.4356/4 \approx .33^2$ and 4 degrees of freedom in the prior. For the prior for D the choices were $\delta_0 = 4$ degrees of freedom in the prior, and $\gamma_0 = .512$, corresponding to a point estimate of $D\sigma^2 = .85$. The choice of .85 comes from $(.85)^{1/2} \approx (\log 240 - \log 6)/4$, where $\log 240 - \log 6$ is roughly the range of the observed data on the log scale, and so $D^{-1} \approx .33^2/.85 \approx .512/4$. The data generally preferred the reported prior to the two alternate priors for σ^2

and D. All three priors for the variance parameters are quite weak.

Our take on the results is based primarily on $c_0 = N/8$ and w_1. The data seem to have a mild preference for model M_2 over M_1, and that the data strongly prefer M_2 or M_1 over M_8. In terms of the problem, this means that we find no effect due to treatment, and that we are unsure, a posteriori, whether there is a baseline CS effect.

We can compare this result with the results from other traditional analyses. Posterior summaries under model M_8 using flat priors are presented in Table 3. Calculations in Table 3 are based on a posterior Gibbs sample of size 1000 after ignoring the first 100 samples. Suppose we define Bayesian significance as the posterior probability that the coefficient is less than some fixed value. This corresponds approximately to a one sided classical p-value. From Table 3, we see that the coefficient of the CS baseline is significant with a tail area of .01. Among the 6 treatment parameters, the three treatments AA, AD, and AN applied to attenders are not particularly different from zero; the posterior means are all less than one posterior standard deviation away from zero. Of the treatments DA, DD, DN in the distractor population, DA and DN have a one-sided p-value approximately .05 (respectively .047 and .055) and DD is quite significant with a p-value of .004. In addition to the expected results that the A treatment decreased distractors' pain tolerance and the D treatment increased pain tolerance, this includes the surprising result that the N treatment produced a decrease in pain tolerance.

We can also compare our results to results based on likelihood ratio tests and AIC and BIC. Table 4 gives for each model the maximized log likelihood

source	mean	sd	$P(\text{parameter} > 0 \vert Y)$
intercept	3.12	.12	1.00
CS	.42	.18	.99
AA	.072	.15	.69
AD	.036	.15	.60
AN	-.10	.16	.26
DA	-.25	.15	.047
DD	.39	.14	.996
DN	-.29	.18	.055

Table 3: Posterior means, standard deviations, and $P(\text{parameter} > 0|Y)$ based on M_8 with a flat prior.

value (max lik), the number of parameters in the model (#parms), basically $m + 2$; the number of observations, #obs, which is $N = 232$ for all models; AIC, which is the maximized log likelihood minus the number of parameters; BIC which is the maximized log likelihood minus $\log(N) * \#\text{parms}/2$. The likelihood ratio test between consecutive nested models is given in the column LRT, along with the degrees of freedom for the test, df; and the p-value associated with the LRT based on the χ^2 distribution with df degrees of freedom. Our #parms included fixed effects and variance parameters in the count. Using a tail area of .05 for a cutoff, stepwise procedures would lead to forward and backward selection of model M_8. Reducing the tail area cutoff to .01 leads to selection of model M_1. In contrast, AIC picks M_1 first, then M_8 and then M_2, while BIC does not distinguish between M_1 or M_2 as a preferred model. Model M_0 in Table 4 is the model with no fixed effect at all, not even an intercept. Model M_0 is never selected by any procedure as the best model.

The Bayes factor $B(8, 1)$ in favor of M_8 against M_1 can be obtained by multiplying $B(8, 2) * B(2, 1)$. Model M_0 could have also been used as a

Model	max lik	#parms	#obs	AIC	BIC	LRT	df	p-value
M_8	-186.727	10	232	-196.727	-213.961	14.464	6	.025
M_2	-193.959	4	232	-197.959	-204.852	5.328	1	.021
M_1	-196.623	3	232	-193.623	-204.793	183.642	1	.000
M_0	-288.444	2	232	-290.444	-293.890			

Table 4: Comparison of models based on AIC, BIC, and the likelihood ratio test, LRT. In the column headed LRT, the tabulated figure is the difference between the maximized likelihood for the model of the current row and the row below. The p-values are tail areas based on the χ^2 distribution with the stated degrees of freedom.

reference point for our results. Reporting the Bayes factors in favor of model M_k against M_0 would also permit comparison of different priors using Bayes factors. We omit this table of numbers since it is not of great interest in our example. Generally, the data supported priors based on w_1 through w_6 equally, for each choice of $c_0 \in (8/N, 1/N, 1/(8N))$, and the data supported the choice $c_0 = 8/N$ slightly more than the other two. The data preferred the choices $c_0 \in (N/8, 1)$ with w_2, w_5, and to a lesser extent w_3, but as discussed before, prior information could not be used to justify these choices for c_0.

5 Discussion

The case of $q_m = 1$ covers a great many important situations as illustrated by our example in the previous section. When q_m is greater than 1, we would consider some sort of importance sampling procedure. One method starts with a sample D_k, $k = 1, \ldots, K$ from the posterior of $p(D)$. We might use

$$[f(Y|m)]^{-1} \approx K^{-1} \sum_{k=1}^{K} [f(Y, D^k|m)]^{-1}$$

which is adapted from Newton and Raftery (1994) and Gelfand and Dey (1994). Alternatively, if we have a sample C_k of $q_m \times q_m$ matrices distributed with density function $g(C)$ which approximates $f(D|Y)$, we could use

$$f(Y|m) \approx K^{-1} \sum_{k=1}^{K} \frac{f(Y, C_k|m)}{g(C_k)}.$$

The obvious candidate for $g(C)$ is the Wishart distribution with mean estimated by an estimate of $E[D|Y]$ from a Gibbs sample D_k. The degrees of freedom parameter would have to be selected to give good properties for the approximation. Other potentially promising methods for computing the marginal density of the data include those proposed by Chen (1994), Chen and Shao (1995), or Chib (1995). We mention that all of these are preliminary suggestions which need much further investigation.

Acknowledgements

Weiss and Wang were supported by NIH grant #GM 50011; Ibrahim was supported by NIH grant #CA 70101-01.

References

Aitken, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society, Series B*, **53**, 111-142.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91**, 109-122.

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

Chen, M.-H. (1994). Importance-weighted Marginal Bayesian Posterior Density Estimation. *Journal of the American Statistical Association* **89**, 818-824.

Chen, M.-H., and Shao, Q.-M. (1995). Estimating Ratios of Normalizing Constants for Densities with Different Dimensions. Research Report No. 653, Department of Mathematics, National University of Singapore.

Chib, S. (1995). Marginal Likelihood From the Gibbs Output. *Journal of the American Statistical Association*, 90, 1313-1321.

Geisser, S. (1971). The Inferential Use of Predictive Distributions. In *Foundations of Statistical Inference*, V. P. Godambe and D. A. Sprott (eds.). Toronto: Holt, Rinehart & Winston, 456-469.

Geisser, S. (1993). *Predictive Inference: An Introduction.* London: Chapman & Hall.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society*, Ser. B **56**, 501-514.

George, E. I. and McCulloch, R. E. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association* **88**, 881-889.

George, E. I., McCulloch, R. E. and Tsay R. (1996). Two Approaches to Bayesian Model Selection with Applications. In *Bayesian Analysis in*

Statistics and Econometrics, eds D. A. Berry, K. M. Chaloner, and John K. Geweke, 339-348.

Hobert, J. P. and Casella, G. (1996). The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models. JASA, 91, to appear.

Ibrahim J. G. and Laud P. W. (1994). A Predictive Approach to the Analysis of Designed Experiments. *Journal of the American Statistical Association* **89**, 309-319.

Kadane, J. B., (1980). Predictive and Structural Methods for Eliciting Prior Distributions. In *Bayesian Analysis in Econometrics and Statistics - Essays in Honor of Harold Jeffreys*, ed. A. Zellner, Amsterdam: North Holland. pp. 89-93.

Karim, M. R. and Zeger, S. L. (1992). Generalized linear models with random effects; Salamander mating revisited. *Biometrics* **48**, 631-644.

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

Laud P. W. and Ibrahim J. G. (1995). Predictive Model Selection. *Journal of the Royal Statistical Society*, Ser. B **57**, 247-262.

Laud P. W. and Ibrahim J. G. (1996). Predictive Specification of Prior Model Probabilities in Variable Selection. *Biometrika* **83**, 267-274.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression (C/R: p1033-1036). *Journal of the American Statistical Association* **83**, 1023-1032.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society*, Ser. B **56**, 3-48.

Oman, S. D. (1985). Specifying a prior distribution in structured regression problems. *Journal of the American Statistical Association* **80**, 190-195.

Raftery, A. E. (1993). Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear models. Technical report # 255, Department of Statistics, University of Washington.

Smith, A. F. M. and Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society*, Ser./ B **42**, 213-220.

Winkler, R. L. (1980). Prior Information, Predictive Distributions, and Bayesian Model-Building. In *Bayesian Analysis in Econometrics and Statistics - Essays in Honor of Harold Jeffreys*, ed. A. Zellner, Amsterdam: North Holland. pp. 95-109.