BAYESIAN STATISTICS 6, pp. 157–185
J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (Eds.)

Bayesian Model Averaging and Model Search Strategies

MERLISE A. CLYDE
Duke University, USA

SUMMARY

In regression models, such as generalized linear models, there is often substantial prior uncertainty about the choice of covariates to include. Conceptually, the Bayesian paradigm can easily incorporate this form of model uncertainty by building an expanded model that includes all possible subsets of covariates. In Bayesian model averaging, predictive distributions or posterior distributions of quantities of interest are obtained as mixtures of the model-specific distributions weighted by the posterior model probabilities. A major difficulty in implementing this approach is that the number of models in the mixture is often so large that enumeration of all models is impossible and some type of search strategy is required to determine a subset of models to use. In the case of an orthonormal design, some computationally simple approximations to the posterior model probabilities are introduced. These are used to develop efficient methods for deterministic or stochastic sampling from high-dimensional model spaces.

Keywords: GENERALIZED LINEAR MODELS; REVERSIBLE JUMP MARKOV CHAIN MONTE CARLO.

1. INTRODUCTION

Linear regression and its generalizations are some of the most commonly used methods in the sciences for finding relationships between explanatory variables and dependent variables. Scientists are collecting larger data sets, and, as the number of observations increases, so does the number of possible explanatory variables. To find relationships and develop models, researchers often turn to various methods of automated model selection and data-mining to select covariates. A major problem is that, by selecting a single model, most analyses ignore model uncertainty, which often significantly outweighs other sources of uncertainty (see Raftery *et al*. 1996, for examples and further discussion). Bayesian model averaging (BMA) provides a coherent and effective approach for incorporating model uncertainty: predictions and inferences are based on a set of models, rather than a single model, and each model in the mixture distribution contributes proportionally to the support it receives from the observed data. While BMA has a long history (de Finetti 1937, Leamer 1978, Mitchell and Beauchamp 1988), it is only the major advance in modern computing environments that has led to the substantial increase in use of BMA and Bayesian variable selection in large problems (Chipman *et al*. 1998, Clyde and DeSimone 1997, Clyde and Parmigiani 1998, Clyde, DeSimone, and Parmigiani 1996, Dellaportas and Forster 1994, Denison 1997, Denison, Mallick, and Smith 1998, George and McCulloch 1993, 1997, Geweke 1996, Madigan and Raftery 1994, Raftery *et al*. 1997, Raftery *et al*. 1996, Smith and Kohn 1996, 1998).

While BMA provides inferences that incorporate model uncertainty, there are a number of difficulties in implementing BMA in high-dimensional problems with correlated variables. BMA and current approaches and issues are reviewed in Section 2 of the paper. In Section 3, we discuss implementing BMA in the context of a normal linear regression model with an

orthogonal design matrix in order to handle larger problems efficiently. Under orthogonality, one can find analytic expressions for posterior model probabilities, means and variances in extremely high-dimensional problems, bypassing the usual problems with convergence of MCMC methods. In situations where a subset of models is desirable, orthogonality can be used to create novel deterministic and stochastic sampling schemes that are more efficient than Gibbs sampling or current MCMC methods. In Section 4, the normal linear model results are used to derive approximations to posterior model probabilities for generalized linear models. These approximations and others based on a model of posterior independence are used either directly to approximate BMA estimates or as proposal distributions for sampling models with high posterior probability. Section 5 illustrates the different approximations in several examples.

2. IMPLEMENTING BAYESIAN MODEL AVERAGING

Consider a regression problem with n observations, with a response variable, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ and a $n \times p$ matrix of covariates, $\boldsymbol{X}$. Typically, there is a great deal of prior uncertainty about which subset of covariates should be included in the regression model. This can be modelled by introducing the vector $\boldsymbol{\gamma}$ of indicator variables that reflect which covariates are included in a model; the model space, Γ, will refer to the set of all possible vectors $\boldsymbol{\gamma}$. Under model $\boldsymbol{\gamma}$, the distribution of $\boldsymbol{Y}$ depends on the covariates through a linear predictor $\boldsymbol{\eta}_\gamma = \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma$, where $\boldsymbol{X}_\gamma$ corresponds to the columns of $\boldsymbol{X}$ where $\boldsymbol{\gamma}$ equals one. Standard Bayesian updating of $f(\boldsymbol{\gamma})$, the prior probability of model $\boldsymbol{\gamma}$, leads to the posterior probability of $\boldsymbol{\gamma}$ given the data $\boldsymbol{Y}$,

$$f(\boldsymbol{\gamma}|\boldsymbol{Y}) = \frac{f(\boldsymbol{Y}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}' \in \Gamma} f(\boldsymbol{Y}|\boldsymbol{\gamma}')\pi(\boldsymbol{\gamma}')}, \qquad f(\boldsymbol{Y}|\boldsymbol{\gamma}) = \int f(\boldsymbol{Y}|\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma})\, f(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma})\, d\boldsymbol{\beta}_\gamma \quad (1)$$

where $f(\boldsymbol{Y}|\boldsymbol{\gamma})$ is the marginal distribution of the data $\boldsymbol{Y}$ given the model $\boldsymbol{\gamma}$ after integrating out model-specific parameters $\boldsymbol{\beta}_\gamma$ with respect to the prior distribution, $f(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma})$. An equivalent representation of (1) based on Bayes factors (Kass and Raftery 1995) is

$$f(\boldsymbol{\gamma}|\boldsymbol{Y}) = \frac{B(\boldsymbol{\gamma}, \boldsymbol{\gamma}_F)\pi(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}' \in \Gamma} B(\boldsymbol{\gamma}', \boldsymbol{\gamma}_F)\pi(\boldsymbol{\gamma}')}, \quad (2)$$

where $B(\boldsymbol{\gamma}, \boldsymbol{\gamma}_F)$ is the Bayes factor for comparing model $\boldsymbol{\gamma}$ to model $\boldsymbol{\gamma}_F$, the full model. Under BMA, the distribution of quantities of interest Δ, such as relative risks or future observations, can be represented as a mixture distribution,

$$f(\Delta) = \sum_{\boldsymbol{\gamma} \in \Gamma} f(\Delta_\gamma|\boldsymbol{Y}, \boldsymbol{\gamma}) f(\boldsymbol{\gamma}|\boldsymbol{Y}) \quad (3)$$

where the model-specific distributions $f(\Delta_\gamma|\boldsymbol{Y}, \boldsymbol{\gamma})$ are weighted by the amount each model is supported by the data as measured by the posterior model probabilities, $f(\boldsymbol{\gamma}|\boldsymbol{Y})$.

In practice, there are two major problems with implementing BMA (in addition to the usual difficulty of specifying prior distributions). The first problem is that the integrals required to obtain the marginal distribution of the data in (2) may be analytically intractable, and approximate methods of integration such as the Laplace method or Monte Carlo methods are necessary (Kass and Raftery 1995). In nested models, such as in regression problems, a generalization of the Savage-Dickey density ratio (Dickey 1971, Verdinelli and Wasserman 1995) can be used to estimate Bayes factors using posterior simulations. The second major difficulty is that the number of models (2^p) is enormous, when there are many covariates. Even if Bayes factors or marginal distributions can be calculated analytically or accurately using simulation methods,

for p greater than 25-30 it is generally computationally infeasible to use all possible models in BMA, and the summation over Γ in (1) and (3) is replaced by a summation over some subset S. In high-dimensional problems where model averaging must be based on a subset of models, the challenging aspect is determining which subset of models S to use. Two approaches in the literature for determining a subset of models to approximate BMA estimates include Occam's Window or stochastic search via Monte Carlo methods.

2.1. Deterministic Search and Occam's Window

Madigan and Raftery (1994) proposed Occam's window as an approach for determining a subset of models to use in model averaging. To determine the models in the "window", any model that has a posterior probability far less than the best model is removed. Any models that have lower posterior probability than any of their simpler sub-models are also removed. For larger problems where enumeration of Γ is not possible, Volinsky *et al.* (1997) use the "leaps and bounds" algorithm to identify potential models in Occam's Window (for $p \leq 30$). In many cases, the number of models in S is reduced to fewer than 25. While Occam's window has provided better predictive performance than selecting the single "best" model (Madigan and Raftery 1994, Madigan *et al.* 1996, Raftery *et al.* 1997, Raftery *et al.* 1996, Volinsky *et al.* 1997), averaging over a larger set of models often leads to better predictive performance.

2.2. *Stochastic Search and Reversible Jump Markov Chain Monte Carlo*

Markov chain Monte Carlo (MCMC) methods are a popular approach for stochastically searching the model space for both variable selection and model averaging and providing samples from the posterior distributions $f(\boldsymbol{\gamma}|\boldsymbol{Y})$ and $f(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}, \boldsymbol{Y})$. Some examples include Carlin and Chib (1995), Dellaportas and Forster (1996), George and McCulloch (1993, 1997), George *et al.* (1995), Geweke (1996), Kuo and Mallick (1998), Madigan and York (1993), Phillips and Smith (1994), Raftery (1993), Raftery *et al.* (1993), Raftery *et al.* (1996), Volinsky *et al.* (1996). The reversible jump MCMC algorithm of Green (1995) generates samples from the joint posterior distribution $f(\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{Y})$ when the models have parameter spaces of different dimensions, and includes many of the above algorithms as special cases (see Dellaportas *et al.* 1997 or Godsill 1997 for discussion of the relationships among these sampling methods). A key component of these algorithms is the proposal distribution for "jumping" to new models.

In high-dimensional problems, convergence of MCMC algorithms is a critical issue, as the number of models in the model space often far exceeds the number of iterations of the MCMC sampler. The correlation structure of $\boldsymbol{X}$ can have a great impact on the convergence rate of MCMC methods (Geweke 1996). Orthogonalizing the explanatory variables can strongly improve convergence and mixing (Clyde *et al.* 1996, Gelfand *et al.* 1996, Gilks and Roberts 1996). In the next section, we show that, with orthogonality, normal errors and known error variance, we can sample directly from the posterior. This results in an independent proposal distribution for sampling from the model space, and provides the basis for approximations to posterior model probabilities for GLMs. Alternatives to MCMC for more efficient sampling are also discussed.

3. ORTHOGONAL REGRESSIONS WITH NORMAL ERRORS

Consider a normal linear model with known error variance σ^2 and orthogonal design matrix. Under independent prior distributions for $(\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma})$,

$$\beta_j|\gamma_j \sim N(b_j, c_j^2\gamma_j)$$
$$\gamma_j \sim \text{Be}(p_j)$$

the elements of $\boldsymbol{\gamma}$ are *a posteriori* independent Bernoulli random variables,

$$\pi_N(\boldsymbol{\gamma}|\boldsymbol{Y}) = \prod_{j=1}^{p} \rho_j^{\gamma_j}(1-\rho_j)^{1-\gamma_j} \qquad \rho_j = \rho(\boldsymbol{Y},\sigma)_j = \frac{O_j(\boldsymbol{Y},\sigma)}{1+O_j(\boldsymbol{Y},\sigma)}$$

$$O_j(\boldsymbol{Y},\sigma) = \left(\frac{p_j}{1-p_j}\right)\left(\frac{\boldsymbol{x}_j^T\boldsymbol{x}_j/\sigma^2 + 1/c_j^2}{1/c_j^2}\right)^{-1/2} \exp\left\{\frac{1}{2}\frac{(\hat{\beta}_j\boldsymbol{x}_j^T\boldsymbol{x}_j/\sigma^2 + b_j/c_j^2)^2}{\boldsymbol{x}_j^T\boldsymbol{x}_j/\sigma^2 + 1/c_j^2} - \frac{1}{2}\frac{b_j^2}{c_j^2}\right\} \tag{4}$$

where $\boldsymbol{x}_j$ is the j-th column of $\boldsymbol{X}$ and $\hat{\beta}_j = (\boldsymbol{x}_j'\boldsymbol{x}_j)^{-1}\boldsymbol{x}_j'\boldsymbol{Y}$ is the least-squares estimate of β_j. The term $O_j(\boldsymbol{Y},\sigma)$ represents the posterior odds of including variable j in the model and is the product of the prior odds times the Bayes factor (BF) for testing the hypothesis that β_j is not equal to zero against the hypothesis that β_j equals zero. For j greater than 1, typically the prior mean is zero. If we select c_j^2 to have the same scale as the likelihood so that $c_j^2 = c^2\sigma^2/(\boldsymbol{x}_j'\boldsymbol{x}_j)$, then the Bayes factor simplifies to

$$(1+c^2)^{-1/2}\exp\left(\frac{1}{2}t^2\frac{c^2}{1+c^2}\right) \tag{5}$$

where t is the t-statistic for testing the hypothesis that β_j equals 0. This formulation is useful for calibrating the prior hyperparameter c.

With an orthogonal design and known σ^2, one can sample directly from the posterior distribution of $\boldsymbol{\gamma}$ using the Gibbs sampler, which produces independent and identically distributed draws from the posterior. This is clearly more efficient than Metropolis algorithms, which result in some rejection of proposed models and correlation in the sample. Quantities such as posterior means and variances under BMA, however, can be calculated, for arbitrarily large p without Monte Carlo sampling,

$$E(\beta_j|Yv) = P(\gamma_j = 1|\boldsymbol{Y})\frac{c^2}{1+c^2}\hat{\beta}_j = \rho_j(\boldsymbol{Y},\sigma)\frac{c^2}{1+c^2}\hat{\beta}_j$$

$$\mathrm{var}(\beta_j|Yv,\sigma) = \rho_j(\boldsymbol{Y},\sigma)\frac{c^2}{1+c^2}\sigma^2(\boldsymbol{x}_j'\boldsymbol{x}_j)^{-1} + \rho_j(\boldsymbol{Y},\sigma)\left(1-\rho_j(\boldsymbol{Y},\sigma)\right)\left(\frac{c^2}{1+c^2}\hat{\beta}_j\right)^2.$$

Reasonable results can be obtained using a plug-in estimate for σ^2. When σ^2 is unknown and has an inverse gamma prior distribution, then it is straightforward to implement a blocked Gibbs sampler where $\boldsymbol{\gamma}|\boldsymbol{Y},\sigma^2$ is distributed as a product of independent Bernoulli distributions and $\sigma^2|\boldsymbol{\gamma},\boldsymbol{Y}$ has an inverse gamma distribution. A Rao-Blackwellized estimator of the marginal probability that γ_j equals one can obtained by averaging $\rho_j(\boldsymbol{Y},\sigma^2)$ over values of σ^2 from the Gibbs sampler. The Rao-Blackwellized estimator can be used in place of $\rho_j(\boldsymbol{Y},\sigma^2)$ to find the posterior mean under model averaging. For nonparametric curve estimation using wavelets (where $p = n$), Clyde *et al*. (1998) considered simulated data from several popular test functions, "Doppler", "Bumps", "Heavisine", and "Blocks". Using data-based estimates of σ^2 in the model probabilities was nearly as efficient (in terms of mean squared error) as the Rao-Blackwellized estimator.

For other quantities of interest, such as quantiles, one cannot average over the model space analytically and one must implement BMA by averaging over a subset of models. There are, however, more efficient sampling alternatives than MCMC. Models generated from the Gibbs sampler can be viewed as a sample drawn with replacement from a finite population. In conjugate models, however, there is no additional information provided by re-sampling models.

For the orthogonal design, Clyde and Littman (1998) develop an efficient sampling-without-replacement algorithm. This is a stochastic algorithm for enumerating all possible models, and is much more efficient than the Gibbs sampler for enumerating the model space. Clyde and Littman also develop an efficient deterministic algorithm for listing models in order of their posterior probabilities. For a sample of k models, both algorithms take $O(kp)$ operations, which is approximately equivalent to the time required to write out each model. These can both be used to identify a subset of models S for BMA, or can be used to identify potential models in Occam's Window. For variable selection problems, where the log of the expected utility can be approximated by a linear function in $\boldsymbol{\gamma}$, the deterministic sampling algorithm can also be used to identify the k models with highest expected utility. Likewise, this sampling-without-replacement algorithm can sample models proportional to their expected utility.

In generalized linear models, one typically cannot integrate out the model-specific parameters, and thus some type of stochastic or deterministic sampling scheme is required to identify the subset of models used in BMA in GLMs. In the next section, we develop several approximations to the posterior model probabilities for GLMs that have the same independent Bernoulli structure as in linear models under orthogonality.

4. APPROXIMATE POSTERIOR MODEL PROBABILITIES IN GLMS

In the generalized linear model, the Y_i's are taken as independent observations from an exponential family with canonical parameter θ_i:

$$f(y_i|\theta_i) = \exp\left\{\frac{w_i}{\phi}(y_i\theta_i - b(\theta_i)) + c(y, w, \phi)\right\}$$

for specific known functions $b(\cdot)$, and $c(\cdot)$ (see McCullagh and Nelder (1989)). The mean and variance of Y_i are given by

$$E(Y_i) = \mu_i = b'(\theta_i) \quad \text{var}(Y_i) = b''(b'^{-1}(\mu_i))\frac{\phi}{w_i},$$

where the linear predictor η_i is related to the mean μ_i via a link function g, $\eta_i = g(\mu_i)$. In the examples in Section 5, the Poisson distribution with a log and identity link, and a possible overdispersion parameter ϕ (Efron 1986, West 1985), is used.

Before routines for maximum likelihood estimation were routinely available, it was common to use transformations of the data and then ordinary least squares to analyse Poisson, binomial or gamma observations. Using the transformed data and normal theory results for model probabilities from Section 3 is a simple and effective way to approximate the posterior distribution of $\boldsymbol{\gamma}$ and leads to posterior independence for $\boldsymbol{\gamma}$. As an alternative approximation that does not assume approximate normality, we fit a "meta-model" for the model space which leads to posterior independence. Approximate posterior independence of $\boldsymbol{\gamma}$ under orthogonality provides computationally simple yet efficient approximations to model probabilities in generalized linear models. The approximations can be used as proposal distributions in reversible jump algorithms or used in the deterministic or stochastic sampling-without-replacement algorithms of Clyde and Littman (1998) to find a subset of models S for BMA.

4.1. Variance-Stabilizing Transformations

For many exponential families there exists a variance-stabilizing transformation h, so that the variance of the transformed response is approximately constant, say k. Additionally, if Y_i is approximately normally distributed $N(\mu_i, V(\mu_i))$, then $h(Y_i)$ is approximately normally

distributed $N(h(\mu_i), V(\mu_i)h'(\mu_i)^2)$, provided h is differentiable and $h'(\mu_i)$ is not zero. The variance-stabilizing transformation can be obtained as the indefinite integral,

$$h(\mu) = \int k^{1/2} V(\mu)^{-1/2} d\mu.$$

For Poisson data h is the square root transformation and $k = 1/4$. Under the variance-stabilizing transformation h, the mean is

$$E(h(\boldsymbol{Y})) \approx h(\boldsymbol{\mu}) = h(g^{-1}(\boldsymbol{\eta}_\gamma)),$$

which is a nonlinear model in terms of $\boldsymbol{\eta}_\gamma = \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma$, depending on the specific link function and variance-stabilizing function. To use normal linear model methods to approximate model probabilities, $h(g^{-1}(\boldsymbol{\eta}_\gamma))$ is replaced by the first two terms of the Taylor series expansion about η_{0i},

$$h(g^{-1}(\eta_i)) \approx h(g^{-1}(\eta_{0i})) + \frac{h'(g^{-1}(\eta_{0i}))}{g'(g^{-1}(\eta_{0i}))}(\eta_i - \eta_{0i}). \tag{6}$$

Substituting the approximate linear regression (6) for the mean results in an approximate normal regression model for the transformed variable $\boldsymbol{W} = (w_1, \ldots, w_n)^T$ with a mean $\boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma$ and known diagonal variance Σ, where

$$\begin{aligned} w_i &= \frac{g'(g^{-1}(\eta_{0i}))}{h'(g^{-1}(\eta_{0i}))} \left(h(Y_i) - h(g^{-1}(\eta_{0i}))\right) + \eta_{0i} \qquad (7) \\ \Sigma_{ii} &= k \left(\frac{g'(g^{-1}(\eta_{0i}))}{h'(g^{-1}(\eta_{0i}))}\right)^2. \end{aligned}$$

If η_{0i} does not depend on i, then Σ_{ii} is a known constant, σ^2. Using $\boldsymbol{W}$ in place of $\boldsymbol{Y}$, the posterior model probabilities in (4) can be used to approximate the posterior distribution of $\boldsymbol{\gamma}$ in generalized linear models.

Calculations and examples for Poisson regression with a log link, are considered by Clyde and DeSimone–Sasinowska (1997). In this case, the variance-stabilizing transformation is the square root transformation and the transformed variable $\boldsymbol{W}$ is

$$\boldsymbol{W} = \frac{2}{\sqrt{\bar{\boldsymbol{Y}}}} \left(\boldsymbol{Y}^{1/2} - \bar{\boldsymbol{Y}}^{1/2}\right) + \log(\bar{\boldsymbol{Y}})$$

where the Taylor series expansion is about $\eta_{0i} = \log(\bar{\boldsymbol{Y}})$.

4.2. *Other Transformations for Normality*

As the derivation of approximate model probabilities in Section 4.1 relies on approximate normality, it is natural to look for other transformations such that the distribution of $h(\boldsymbol{Y})$ is "close" to a normal distribution. Hougaard (1982) discussed various transformations in one-parameter exponential families given by

$$h(\theta) = \left(\frac{\phi}{w_i}\right)^\delta \int \left\{\frac{d^2}{d\theta^2} b(\theta)\right\}^\delta d\theta,$$

where δ is a constant that determines properties of the reparameterization. For example, $\delta = 0$ corresponds to the canonical parameterization, $\delta = 1/3$ corresponds to a quadratic loglikelihood

parameterization (the third derivative of the log likelihood vanishes), $\delta = 1/2$ is the variance-stabilizing transformation, $\delta = 2/3$ results in approximate zero skewness (symmetry), and $\delta = 1$ is the mean value parameterization. In terms of μ, the transformation is

$$h(\mu) = \frac{\phi}{w_i} \int V(\mu)^{\delta-1} d\mu \tag{8}$$

so that the approximate mean under transformation of Y_i is $h(\mu_i)$ and the approximate variance is $V(\mu_i)^{2\delta-1}$. For the Poisson model, the transformation is $h(Y_i) = Y_i^{\delta}/\delta$ with an approximate mean of μ_i^{δ}/δ and approximate variance of $\mu_i^{(2\delta-1)}$ for $\delta > 0$, and is the log transformation for $\delta = 0$.

If $h(Y_i)$ is approximately normal, then the above mean and variance determine the distribution. To achieve a mean that is linear in $\boldsymbol{\beta}_\gamma$, a Taylor series expansion of h about η_{0i} is carried out as in (6). The variable $\boldsymbol{W}$ is defined analogously to $\boldsymbol{W}$ in Section 4.1 and is approximately normal with mean $\boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma$, but Σ is diagonal with elements

$$\Sigma_{ii} = \left(\frac{g'(g^{-1}(\eta_{0i}))}{h'(g^{-1}(\eta_{0i}))} \right)^2 V(g^{-1}(\eta_i))^{2\delta-1}.$$

The approximate variance Σ varies with η_i, so unless a common value is substituted for all cases, the nonconstant variance destroys the orthogonality required for independence of the posterior model probabilities. One option involves replacing Σ_{ii} by a known constant σ^2, such as evaluating Σ_{ii} at η_0, the point where the Taylor series was evaluated.

4.3. *Transformation to Linearity*

While the above transformations may improve normal approximations, they often result in a mean that is nonlinear in $\boldsymbol{\beta}_\gamma$. Using the link function to define a transformation of $\boldsymbol{Y}$ results in approximate linearity, $E(g(\boldsymbol{Y}|\boldsymbol{\gamma})) \approx \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma$. When the link function is the same as the variance-stabilizing transformation, the variance of $g(\boldsymbol{Y})$ is approximately constant. Otherwise, Σ_{ii} may be replaced by a common value σ^2.

4.4. *Model Probabilities by Approximate Bayes Factors*

In generalized linear models, orthogonality of the covariates does not lead to posterior independence of elements of $\boldsymbol{\gamma}$. However, ignoring the dependence may lead to reasonable approximations of posterior model probabilities. Using the Savage-Dickey density ratio, the posterior odds that γ_j equals one can be approximated by

$$O_j = \frac{p_j}{1 - p_j} \frac{f_{\beta_j}(0)}{f_{\beta_j}(0|\boldsymbol{Y})}$$

where $f_{\beta_j}(\cdot|\boldsymbol{Y})$ is the marginal posterior distribution for β_j obtained from the full model and $f_{\beta_j}(\cdot)$ is the marginal prior distribution for β_j. One can approximate the marginal distribution using the asymptotic normal approximation for the maximum likelihood estimates (MLEs) or the Laplace method (Raftery 1996). The approximate posterior odds are then used in equation (4) to approximate the posterior distribution of $\boldsymbol{\gamma}$. In the examples that we have considered, when the correlations among the MLEs are less than 0.2 in absolute value, these approximations have worked reasonably well. With higher correlations, independent proposal distributions become less efficient.

4.5 *Model Probabilities via Loglinear Models on the Model Space*

One can view each model $\boldsymbol{\gamma}$ as a cell in a 2^p contingency table that represents the model space, where the probability of being in the cell $\boldsymbol{\gamma}$ is $\pi(\boldsymbol{\gamma}|Yv)$. In general, posterior model probabilities may be represented by a "meta-model" that is a saturated log-linear model for $\pi(\boldsymbol{\gamma}|Yv)$,

$$\begin{aligned}\log(\pi(\boldsymbol{\gamma}|\boldsymbol{Y})) &= \log(q_\gamma) - \log(\sum_{\boldsymbol{\gamma}'} q_{\boldsymbol{\gamma}'}) \\ &= \alpha_0 + \sum_j \alpha_j\gamma_j + \sum_{j,k} \alpha_{jk}\gamma_j\gamma_k + \sum_{j,k,l} \alpha_{jkl}\gamma_j\gamma_k\gamma_l + \ldots \alpha_{1\ldots p}\prod_{j=1}^{p}\gamma_j,\end{aligned} \tag{9}$$

where the vector $\boldsymbol{\alpha} = (\alpha_0, \ldots \alpha_{1\ldots p})^T$ is a function of the data $\boldsymbol{Y}$ and q_γ is the un-normalized posterior model probability.

In the linear regression model with known σ^2, the log model probability is

$$\log(\pi(\boldsymbol{\gamma}|\boldsymbol{Y})) = \sum_j (1 + O_j(\boldsymbol{Y}, \sigma))^{-1} + \sum_j O_j(\boldsymbol{Y}, \sigma)\gamma_j$$

which corresponds to a model of independence for the model space, where all the two-way (α_{jk}) and higher order interaction terms are zero. Likewise, in the approximations in 4.1-4.4, the elements of $\boldsymbol{\gamma}$ are all assumed to be independent and the "parameters" α_j have been estimated under different transformations.

Approximate posterior model probabilities in GLMs can be obtained by fitting a model of independence for the meta-model. Since $\log(\sum_{\boldsymbol{\gamma}'\in\Gamma} q_{\boldsymbol{\gamma}'})$ is a constant we can alternatively model $\log(q_\gamma)$ and then obtain the normalizing constant by summation afterwards. We estimate q_γ using the Laplace approximation for integrals (Raftery 1996),

$$\begin{aligned}q_\gamma &= \pi(\boldsymbol{\gamma}) \int p(\boldsymbol{Y}|\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}) p(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma})\, d\boldsymbol{\beta}_\gamma \\ &\approx \pi(\boldsymbol{\gamma})\ (2\pi)^{p/2} |\boldsymbol{\psi}_\gamma|^{-1/2} p(\boldsymbol{Y}|\tilde{\boldsymbol{\beta}}_\gamma, \boldsymbol{\gamma})\ p(\tilde{\boldsymbol{\beta}}_\gamma|\boldsymbol{\gamma}),\end{aligned} \tag{10}$$

where $\tilde{\beta}_\gamma$ is the posterior mode given model $\boldsymbol{\gamma}$ and $\boldsymbol{\psi}_\gamma$ is the negative Hessian of the log posterior with (i, j) element,

$$[\psi_\gamma]_{ij} = -\frac{\partial^2}{\partial\beta_i\, \partial\beta_j} \log(p(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\gamma}) p(\boldsymbol{\beta}|\boldsymbol{\gamma})).$$

To estimate $\boldsymbol{\alpha}$, let $\boldsymbol{Q}$ denote the vector of the Laplace approximations for a subset of l models, and let $\boldsymbol{U}$ denote the $l \times p$ "design" matrix based on the l models, where the rows of $\boldsymbol{U}$ are the corresponding vectors $\boldsymbol{\gamma}$. The log-linear model can be represented as

$$\log(\boldsymbol{Q}) = \boldsymbol{U}\boldsymbol{\alpha} + \boldsymbol{e} \tag{11}$$

where $\boldsymbol{e}$ represents an "error term" due to lack of independence. Estimates of $\boldsymbol{\alpha}$ using least squares are then transformed to obtain estimates of the marginal posterior probability that γ_j equals one:

$$p_j = \frac{\exp(\hat{\alpha}_j)}{1 + \exp(\hat{\alpha}_j)}.$$

To find the best subset of models $\boldsymbol{U}$ to estimate $\boldsymbol{\alpha}$ is a "meta-design" problem. One approach is to start with the best model γ^* based on one of the approximate model probabilities from

sections 4.1-4.4. The first row of $\boldsymbol{U}$ is γ^*. To obtain row j in $\boldsymbol{U}$, one starts with γ^* and switches γ_j^* to $1 - \gamma_j^*$. This ensures that all parameters are estimable, and has given better results than a random selection of models or using a "design" based on fractional factorials or orthogonal arrays, but is open to further research. One can include more models and extend this approach to incorporate higher order interactions to model dependence among the elements of $\boldsymbol{\gamma}$.

4.6. Posterior Inference

The approximations to posterior model probabilities in Sections 4.1-4.5 can be used directly for computing approximate posterior means or variances. They can also be used in the deterministic or sampling-without-replacement algorithms of Clyde and Littman (1998) to provide a subset of models, S. The approximations can be used as proposal distributions in reversible jump MCMC algorithms. Generation of a new model only involves sampling from Bernoulli distributions, and as long as the approximate model probabilities are strictly greater than 0 and less than 1, the chain is irreducible. One can potentially move to any model in the model space from the current state, unlike the default proposal in a reversible jump algorithm, which uses birth and death steps to move to new models and only allows moves to models that differ in one component of $\boldsymbol{\gamma}$. In the default birth step, each variable (not currently in the model) has the same probability of being added; using the approximate probabilities as a proposal distribution can be viewed as having different probabilities of birth/death for each variable. In large model spaces, this enables more rapid mixing, and since components of the new model are selected based on their approximate posterior distribution, this can be more efficient as the procedure focuses on important variables. The importance sampler in Clyde *et al.* (1996) also has this feature. Using mixtures of the proposal distributions in Sections 4.1-4.5 can be easily implemented, or one may even mix these with the usual birth/death proposal. This may be desirable in terms of "robustifying" the independent proposal sampler in problems with moderate correlation among the MLE's in the GLM. Standard methods can be used to generate $\boldsymbol{\beta}_\gamma$ given a proposed $\boldsymbol{\gamma}$ such as adaptive rejection sampling, independence proposals based on a normal approximation, or a random walk proposal (Clyde and DeSimone–Sasinowska 1997, Dellaportas and Forster 1996, George *et al.* 1994, Kuo and Mallick 1998).

5. EXAMPLES

5.1. *Tetanus*

Accuracy of the various approximations to the posterior model probabilities was examined using a log-linear model from Healy (1988, page 97) on deaths from tetanus, where we assume that the number of deaths has a Poisson distribution. The problem is small enough (a 2^3 contingency table with factors Mortality (M), Severity of Tetanus (S), Antitoxen Indicator (S)) that model probabilities for all log-linear models (2^{2^3}) can be calculated. While in practice one might only examine hierarchical models, for comparison purposes this restriction has not been imposed. The design matrix is constructed using plus and minus ones to code the effects (the ANOVA sum-to-zero constraint) and products of columns to create 2-way and 3-way interactions, so that $\boldsymbol{X}'\boldsymbol{X} = 8I_8$.

The prior mean for the regression coefficients was 0 and the covariance matrix was $16I_8$. The Laplace approximation was used as a "gold standard" for the comparison. Table 1 shows the Kullback-Leibler divergence between the approximate posterior model probabilities using the power transformations from Section 4.1-4.3 and the model probabilities under the Laplace approximation. For comparison, the approximation based on the Savage-Dickey ratio and a normal approximation based on the GLM estimates (Section 4.4) resulted in a divergence of 0.045. The model of independence (Section 4.5) was fitted to all models, giving a Kullback-Leibler

Table 1. *Kullback-Leibler divergence between approximate posterior model probabilities and posterior model probabilities based on the Laplace approximation for the tetanus data.*

Method:	Power Transformation δ					Savage-Dickey	Independence (all models)	Independence (8 models)
	0	1/3	1/2	2/3	1			
KL	0.045	0.030	0.025	0.018	0.011	0.045	0.011	0.015

divergence of 0.011. Estimating the model of independence using a minimum subset of 8 models results in a Kullback-Leibler divergence of 0.015. Overall, there is an excellent agreement between the approximations and the Laplace approximations; however the approximations can be calculated in a fraction of the time necessary to compute all Laplace approximations. Similar results were obtained with other prior covariance matrices and using other Monte Carlo methods to estimate the "exact" posterior model probabilities.

5.2. *Simulated Loglinear Model*

To compare model averaging to other model selection approaches, DeSimone–Sasinowska and Clyde (1998) consider a 2^4 contingency table, with the effects coded using the ANOVA sum-to-zero constraints as above, so that the design matrix for the saturated model has 16 columns and there are 2^{16} possible models. The response variable was generated according to the model

$$\eta = \log E(Y) = 2.7375X_1 + 0.15X_2 + 0.15X_4 + 0.15X_5 + 0.15X_9 + 0.15X_{10} + 0.15X_{16}$$

where $Y \sim \text{Poisson}\,(\exp(\eta))$. The prior mean for the intercept was $\log(\bar{Y})$, while the prior mean for the remaining coefficients was zero. The prior covariance was a diagonal matrix, $(c/\bar{Y})(X'X)^{-1}$ with $c = 249$ based on the George and Foster (1998) Risk Inflation Criterion. This provides one method of automatically calibrating the prior distribution: see George and Foster for other choices. A uniform prior distribution over the model space was also used.

Table 2. *Comparing Bayesian Model Averaging to maximum likelihood estimation*

Method	Bayesian Model Averaging			Maximum Likelihood		
	Independent Approx + RJ	Independent Approx	MCMC (KM)	BIC	STEP	FULL
avg MSE	0.258	0.319	0.254	1.102	0.805	1.188
min MSE	0.072	0.065	0.064	0.252	0.135	0.268
max MSE	1.461	3.044	1.497	2.735	2.523	4.450

Using 120 simulated data sets from the true model, we compared Bayesian Model Averaging to maximum likelihood estimation using mean squared error calculated for each simulation, MSE = $\sum_j (\beta_j - \hat{\beta}_j)^2/16$ (see Table 2). For the independent approximate posterior model probabilities, we used the variance-stabilizing transformation from Section 4.1. We used this independent approximation as a proposal distribution in a reversible jump MCMC and also directly to estimate the posterior mean of $\boldsymbol{\beta}$ without using MCMC. For comparison, we used a MCMC algorithm based on the method in Kuo and Mallick (KM) to estimate model probabilities and the BIC approximation from Volinsky *et al*. (1997). These four approaches to BMA were compared to the the maximum likelihood estimates from stepwise variable selection and the full model maximum likelihood estimates. The best results are obtained by the two MCMC approaches, with little difference in efficiency in terms of MSE. However, the computational time for the KM MCMC algorithm was 8.5 times longer than using the independent proposal

distribution in the reversible jump MCMC algorithm. Using the independent approximation directly for model averaging is roughly 2.5 to 4 times more efficient (in terms of MSE) than using BIC, the MLE under stepwise selection or the MLE from the full model, but with roughly the same computational burden as finding the MLE in the full model.

5.3. *Particulate matter and Mortality*

A number of scientific studies have found an association between mortality and PM_{10}, particulate matter less than 10μ in aerodynamic diameter, prompting the EPA to propose changes in the National Ambient Air Quality Standard (NAAQS) for PM_{10} and creating new standards for $PM_{2.5}$. Many analyses are based on Poisson or over-dispersed Poisson regression models for daily non-accidental mortality. As there is a large number of potential confounding variables, traditional model selection techniques can lead to the selection of very different models and ignore model uncertainty. One concern is that the positive association between PM_{10} and mortality found in many of these studies is a result of multiple testing and selection. We will use BMA to estimate the effect of PM_{10} on mortality taking into account uncertainty about which predictors should be included.

Data from Chicago and Cook County, Illinois, contain daily measurements of total non-accidental mortality for individuals age 65 and older, plus daily and lagged values of average temperature, maximum and minimum temperature, specific humidity, windchill, discomfort index, atmospheric pressure, and total solar radiation over the period 1985-1990. The PM_{10} data are based on daily values from a single daily monitoring station and the daily average of a subset of monitoring stations that report values every six days. Other measures are based on three-day means from the daily station, the subset of six stations, and all of the 20 available reporting stations. For additional information on the data see Styer *et al*. (1995).

Because of the correlation structure and its effect on sampling from the model space, principal component analysis was used to create a set of orthogonal meteorological and orthogonal particulate matter variables. This has the advantage of keeping all the information in the original variables, but eliminating problems due to multicollinearity within the two groups of variables. The first meteorological principal component (PC) combines temperature and humidity with windchill, where low values correspond to extreme windchill, low temperature and humidity combinations and high values correspond to high temperature and humidity days. The second PC measures average atmospheric pressure, while the third measures changes in pressure over 3-day intervals with high values of the component corresponding to extreme highs followed by a low, and low values related to low pressure followed by a high. The fourth principal component appears to be related predominately to solar radiation. The first PM_{10} principal component provides a weighted average of the daily and three-day means, resulting in more weight given to the daily stations. The second PM_{10} principal component captures days where the current day average differs from the three-day average. Additionally, fourth order orthogonal polynomials of the first two PM_{10} and meteorological principal components were included to allow for nonlinear functions. Interpretation of the remaining variables is not as clear but they capture departures from the main component directions, and are included in the design matrix so that there is a total of 35 variables. Other methods such as Gram-Schmidt or sliced inverse regression (Li 1992) could also be used, leading to a different set of orthogonal variables.

In general, the canonical log link has been used in such analyses. As data are often aggregated over time (daily mortality) and space (metropolitan area) this often leads to biased estimates, as the aggregate mean no longer has a log-linear form (Richardson 1992). As individual level covariates are unavailable, it is necessary to rely on aggregated data. Under an identity link with additive effects, the Poisson mean retains linearity under aggregation so

that results may be summarized at different levels of aggregation without re-fitting the model. While aggregation combined with measurement error may lead to bias with the identity link, simulations indicate the bias appears much worse with the log-linear formulation. To compare the assumption of an identity link to a log link, the log Bayes factor was approximated using the Schwarz criterion (Schwarz 1978, Kass and Raftery 1995) with deviances estimated under the full model allowing for overdispersion. Twice the log Bayes factor was 4.58 indicating positive support for the identity link. Using a reduced set of variables, twice the log Bayes factor was 8.04, indicating strong support for the identity link. In the following, we will explore using Poisson regression with the identity link, using BMA to potentially adjust for all of the meteorological variables. To allow for overdispersion, we incorporate an additional scale parameter ϕ (Clyde and DeSimone–Sasinowska 1997, Efron 1986, West 1985).

Independent normal prior distributions were used for each of the regression coefficients, with the prior mean for the intercept taken as the sample mean, 83 deaths/day; the other prior means were set to zero. The prior standard deviations were based on 10 times the estimated standard error from the GLM analysis of the full data. The prior probability for each component of the orthogonal polynomials of the first two meteorological and PM_{10} principal component variables was set to 0.75. For the remaining variables, the prior probability that the variable was included was set to 0.5. The prior distribution also incorporates a constraint that the Poisson mean must be positive. Alternative prior distributions, such as the data augmentation prior suggested by Bedrick *et al.* (1997) may be preferable when using the identity link. Development of alternative proposal distributions using this class of prior distributions is under way.

Approximate model probabilities were estimated using the Bayes factor method described in Section 4.4, using the MLEs from the Poisson regression model with an overdispersion parameter and an identity link. Because the standard errors of the MLEs were more variable than the OLS standard errors which ignore correlation, this approximation seemed preferable and had to a better acceptance rate in the MCMC sampler than the approximation based on the variance-stabilizing transformation. Using a burn-in of 2000 iterations, the Markov chain was run for 100,000 iterations with every 10th iteration saved to provide an approximately independent sample for making posterior inferences. Details of the MCMC algorithm and methods for checking convergence and fit are described in Clyde and DeSimone-Sasinowska (1997).

Because of the size of the model space, the Monte Carlo frequency of many models will be one or zero. For each model in the sample, the Laplace method was used to estimate the model probability and posterior mean of $\boldsymbol{\beta}_\gamma$. Figure 1 shows one representation of the space of sampled models, with models and variables arranged in a matrix, with variables represented by columns and models by rows. Variables that are excluded in a model are left blank. The models are sorted in order of their log posterior model probability, in this case, estimated using the Laplace approximation, with the best model in the sample at the bottom. The intensity of the color or grey-scale is proportional to log(posterior model probability) $-$ log(max(posterior model probability). We find such a plot helpful in visualizing important variables and uncertainty about which ones should be included. From this we see that PM_{10} variables have been included in almost all of the sampled models. The association does not appear to be the effect of model selection. While models in Figure 1 are ranked based on posterior model probabilities, such a plot can also be constructed using other utility functions.

Figure 2 shows the posterior distribution of the linear, quadratic and cubic coefficients for the orthogonal polynomial in the first PM_{10} principal component variable. The last plot in the figure shows the posterior mean for the linear predictor under model averaging plus 95% probability intervals. Because the variables are centered, this shows how the change in the

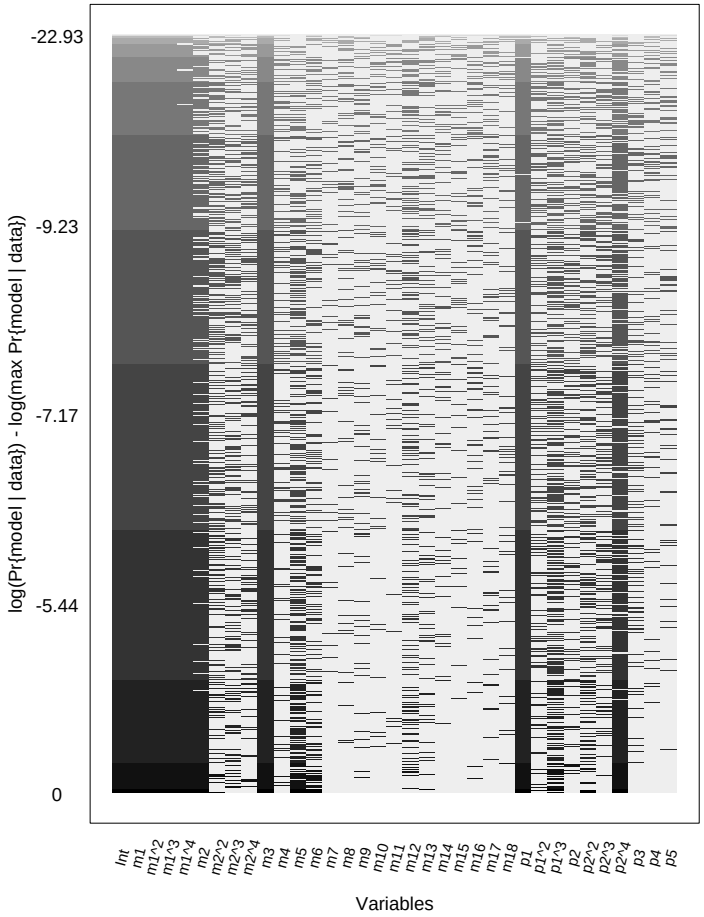

Figure 1. *Posterior distribution of γ. Each column corresponds to a column of the $\boldsymbol{X}$ and each row to a model obtained from the MCMC output. The intensities are proportional to the log of the approximate model probability estimated using the Laplace approximation, with the best model having a value of 0.*

number of deaths from the PM_{10} increases/decreases relative to the overall mean. These values can be converted to relative risks by dividing by the expected mortality under the average values of PM_{10} and other variables. A similar figure showing a nonlinear relative risk for PM_{10} is presented by Smith *et al.* (1997) using a generalized additive model with a log link. While there is some evidence of a nonlinear effect at higher PM_{10} levels (the posterior probability that the higher order terms are included is 0.68), this appears to be sensitive to the choice of the prior probabilities. The posterior probability that there is a particulate matter effect is close to 1. Without the higher order terms, the posterior mean number of deaths attributed to PM_{10} on a day with 150 $\mu g/m^3$ is 9.3, with a 95% probability interval of (4.2, 14.1). This corresponds to a 5 to 16 percent increase in mortality compared to the average level of PM_{10} and is at the highest permissible daily average for PM_{10}. Such conditions occurred, however, only on 2-3 days in the seven year period. For a day with average PM_{10} levels, 41 $\mu g/m^3$, the expected number of deaths attributed to PM_{10} ranges from 1.1 to 3.8. While it is not clear what effect the new standards will have on PM_{10} levels, Smith *et al.* (1997) have suggested that levels may decrease by 10 $\mu g/m^3$ on average. Under the model with the linear term for PM_{10}, 95% posterior intervals for the expected decrease in mortality due to such a change are 0.25 to 0.82 deaths/day, or roughly 91 to 300 deaths per year in the over 65 population in Cook County. One can use this to obtain the predictive distribution for the reduction in mortality. A 95% prediction interval suggests that the overall reduction is 70 to 340 deaths/year.

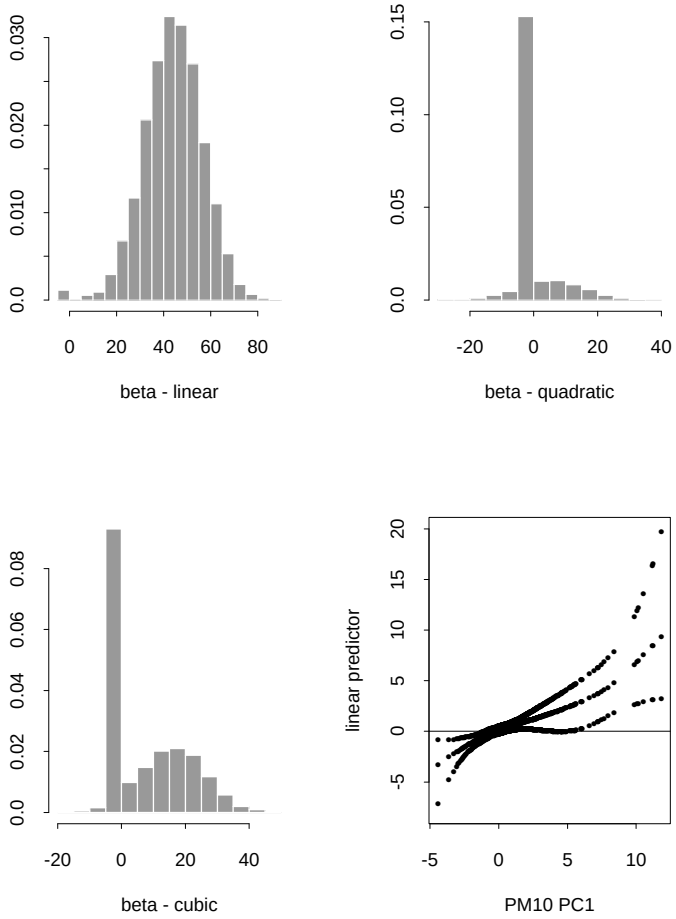

Figure 2. *Posterior distributions of the coefficients of of the first principal component for particulate matter variable using third order orthogonal polynomials. Estimated effect and* 95% *probability interval using BMA.*

These are still preliminary results and are subject to a number of caveats. If the PM_{10} measurements are not representative of the average ambient outdoor exposure in the population, then ecological bias is a concern in interpreting the results, and may lead to under- or over-estimation of the effect. With the non-zero probability that the PM_{10} effect is nonlinear, aggregation and measurement error become more serious issues.

6. DISCUSSION

Model uncertainty often dominates other forms of uncertainty, such as parameter uncertainty, and in almost every application of BMA model averaging has led to better predictive performance (Raftery, Madigan and Volinsky 1996). In the examples in Section 5, using orthogonal variables can lead to major improvements in computational efficiency for implementing BMA. In the PM_{10} example, our goal is to make predictions as well as to make inferences about the effect of PM_{10} on mortality. Because of the large number of models, orthogonalizing the meteorological variables leads to dramatic increases in computational efficiency, making it much easier to explore the model space. The use of principal components results in more efficient computations, but also still allows meaningful interpretation of several of the components.

There are still a number of open areas. The choice of orthogonal variables and what impact that has on the efficiency, and selection of prior distributions on the model space, are open

problems. The approximate posterior distributions are all based on a model of independence for the posterior distribution of γ. Simple diagnostics in addition to high correlations among the mle's that indicate when independence is not a reasonable approximation for a proposal distribution would be useful.

REFERENCES

Carlin, B.P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo. *J. Roy. Statist. Soc. B* **57**, 473–484.

Chipman, H., George, E. I. and McCulloch, R. (1998). Bayesian CART model search. *J. Amer. Statist. Assoc.* **93**, 935-960.

Clyde, M. and DeSimone–Sasinowska, H. (1997). Accounting for model uncertainty in Poisson regression models: Particulate matter and mortality in Birmingham, Alabama. *Tech. Rep.* **97–06**, Duke University.

Clyde, M., DeSimone, H. and Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *J. Amer. Statist. Assoc.* **91**, 1197–1208.

Clyde, M. and Littman, M. (1998). Computationally efficient sampling for Bayesian model averaging. *Tech. Rep.*, Duke University.

Clyde, M. and Parmigiani, G. (1996). Orthogonalizations and prior distributions for orthogonalized model mixing. *Modelling and Prediction: Essays in Honor of Seymour Geisser* (J. C. Lee *et al.* , eds.). New York: Springer, 206-227.

Clyde, M. and Parmigiani, G. (1998). Protein construct storage: Bayesian variable selection and prediction with mixtures. *J. Biopharmaceutical Statist.* **8**, 431–443.

Clyde, M., Parmigiani, G., Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika* **85**, 391–402.

Dellaportas, P. and Forster, J. J. (1996). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Tech. Rep.*, Southampton University.

Dellaportas, P. Forster, J. J, and Ntzoufras, I. (1997). On Bayesian model and variable selection using MCMC. *Tech. Rep.*, Southampton University.

Denison, DGT (1997). *Simulation Based Bayesian Nonparametric Regression Methods*. Ph.D. Thesis, Imperial College, London.

Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998). A Bayesian CART algorithm. *Biometrika* **85**, 363–377

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *Ann. Statist.* **42**, 204–223.

Draper, D. (1995). Assessment and propagation of model uncertainty. *J. Roy. Statist. Soc. B* **57**, 45–70 (with discussion).

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *J. Amer. Statist. Assoc.* **81**, 709–721.

Gelfand, A., Sahu, S. K. and Carlin, B. P. (1996). Efficient parameterizations for generalized linear mixed models. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 48–74 (with discussion).

George, E. I. and Foster, D. P. (1997). Calibration and empirical Bayes selection. *Tech. Rep.*, Univ. of Texas, Austin.

George, E.I. and McCulloch, R. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88**, 881–889.

George, E. I. and McCulloch, R. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–374.

George, E. I., McCulloch, R. and Tsay, R. (1995). Two approaches to Bayesian model selection with applications. *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner* (D. A.Berry, K. M. Chaloner and J. K. Geweke, eds.). New York: Wiley, 339–348.

Geweke, J. (1996). Variable selection and model comparison in regression. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 609–620.

Gilks, W. R. and Roberts, G. O. (1996). Strategies for improving MCMC. *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds.). London: Chapman and Hall, 89–114.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Hoeting, J., Raftery, A. E., and Madigan, D. M. (1996). A method for simultaneous variable selection and outlier identification in linear regression. *Comput. Statist. and Data Analysis* **22**, 251–270

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773–795.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā A* (to appear).

Leamer, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data* Chichester: Wiley

Madigan, D. M. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.* **89**, 1535–1546.

Madigan, D., Raftery, A. E., Volinsky, C., and Hoeting, J. (1996). Bayesian Model Averaging. *Integrating Multiple Learned Models (IMLM-96)*, (P. Chan, S. Stolofo, and D. Wolpert, eds.), 77–83.

Madigan, D. M. and York, J. (1995). Bayesian graphical models for discrete data. *Internat. Statist. Rev.* **63**, 215-232.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83**, 1023–1036.

Phillips, D. B. and Smith, A. F. M. (1994). Bayesian model comparison via jump diffusions. *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds.). London: Chapman and Hall, 215–238.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika* **83**, 251–266.

Raftery, A. E., Madigan, D. M., and Hoeting, J. (1997). Model selection and accounting for model uncertainty in linear regression models. *J. Amer. Statist. Assoc.* **92**, 179–191.

Raftery, A. E., Madigan, D. M. and Volinsky C. T. (1996). Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 323–350, (with discussion).

Richardson, S. (1992). Statistical methods for geographical correlation studies. In *Geographical and Environmental Epidemiology: Methods for Small-Area Studies* (P. Elliott, J. Cuzick, D. English and R. Stern, eds.). Oxford: Oxford University Press, 181–204.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75**, 317–344.

Smith, M. and Kohn, R. (1998). Nonparametric estimation of irregular functions with independent or autocorrelated errors. *Practical Nonparametric and Semiparametric Bayesian Statistics*, (D. Dey, P. Müller, and Sinha, D. eds.). New York: Springer-Verlag, 157–180.

Smith, R. L., Davis, J. M. and Speckman, P. (1997). Assessing the human health risk of atmospheric particles. *Tech. Rep.*, Univeristy of North Carolina, Chapel Hill.

Styer, P., McMillan, N., Gao, F., Davis, J. and Sacks,J. (1995). The effect of airborn particulate matter on daily death counts. *Environ. Health Persp.* **103**, 490–497.

Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *J. Amer. Statist. Assoc.* **90**, 614–618.

Volinsky, C., Madigan, D., Raftery, A. E. and Kronmal, R. (1997). Bayesian model averaging in proportional hazard models: Assessing stroke risk. *Appl. Statist.* **46**, 433–448.

West, M. (1985). Generalized linear models: Scale parameters, outlier accommodation and prior distributions. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 531–558.

DISCUSSION

PETROS DELLAPORTAS *(Athens University of Economics and Business, Greece)*

This is an interesting paper which collects interesting results and presents some new research avenues. It is based on a series of previous papers by the author and co-authors focused on Bayesian model averaging via orthogonalised model mixing.

There is plenty of empirical and some theoretical evidence that Bayesian model averaging yields better out-of-sample predictions than a single "best" model (Draper 1995, Raftery *et al.* 1995). The author focuses on prediction via orthogonalised model mixing (Clyde *et al.* 1996) which is *extremely* fast for linear models. The innovative idea of this paper is to exploit this algorithm for the more general family of generalised linear models (GLM) by transforming any

GLM to a normal regression model using asymptotic arguments and then applying orthogonalised mixing to achieve a fast model averaging. This can provide either estimates of posterior model probabilities or proposals for other MCMC algorithms.

There are two questions one needs to answer: when do the suggested transformations produce the desired result, and when should we prefer to obtain predictions with this method instead of adopting other existing model averaging algorithms? Clearly, if the method is robust or fail-safe, there is a big potential due to its amazing efficiency. We will try to motivate discussion by re-analysing some real data examples.

We would like to compare Clyde's algorithm with standard MCMC model choice algorithms such as reversible jump (Green, 1995) and variants, so an obvious choice is the log-linear model of Section 5.1. We used priors $\beta_j \sim N(0, 2)$ (Dellaportas and Forster, 1997) for all model parameters. Model averaging can be performed considering only hierarchical models or by considering all possible models. Clyde chooses the latter for illustration, but because the former might also be a feasible approach we chose to illustrate results for both cases. Tables 3 and 4 present the results, with A denoting severity, B denoting antitoxen medication, C denoting mortality and M denoting the overall mean. It seems that Clyde's algorithm performs well for a series of δ values.

Table 3. *Posterior Model Probabilities of Hierarhical Models in Healy Data.*

Model	MCMC Poisson	Clyde $\delta = 0$	$\delta = 1/3$	$\delta = 1/2$	$\delta = 2/3$	$\delta = 1$
AC+BC	0.59	0.55	0.49	0.47	0.46	0.45
AC+B	0.25	0.34	0.38	0.38	0.38	0.35
AB+AC+BC	0.07	0.06	0.07	0.07	0.08	0.01
AB+AC	0.07	0.04	0.05	0.06	0.07	0.08
A+BC	0.01	0.01	0.01	0.01	0.00	0.00
ABC	0.00	0.01	0.01	0.01	0.01	0.01
KL Distance	0.0290	0.0249	0.0371	0.0418	0.0443	0.0435

Table 4. *Posterior Model Probabilities of all Models in Healy Data*

Model	MCMC Poisson	Clyde $\delta = 0$	$\delta = 1/3$	$\delta = 1/2$	$\delta = 2/3$	$\delta = 1$
M+AC+BC	0.18	0.23	0.21	0.20	0.19	0.16
M+C+AC+BC	0.15	0.15	0.13	0.12	0.12	0.13
M+AC	0.13	0.14	0.16	0.16	0.16	0.13
M+C+AC	0.11	0.09	0.10	0.10	0.10	0.10
M+A+AC+BC	0.06	0.04	0.04	0.04	0.05	0.05
M+A+AC	0.05	0.02	0.03	0.04	0.04	0.04
M+AB+AC	0.03	0.02	0.02	0.02	0.03	0.03
M+AB+C+AC	0.03	0.01	0.01	0.01	0.02	0.02
M+B+C+AC+BC	0.02	0.02	0.01	0.01	0.01	0.01
KL Distance	0.0638	0.0626	0.0425	0.0374	0.0354	0.0486

A question that someone might pose looking at Tables 3 and 4, and in particular how the posterior probabilities vary with δ, is whether the prior $\pi(\beta, \gamma)$ should depend on δ. Smith and Kohn (1996) suggest the use of a transformation $h(Y_i)$ that has, approximately, the same

distribution for every δ. Their approach is not readily applicable in Clyde's case, but some thoughts on this direction might provide ideas to robustify the methodology put forward.

A more challenging problem is to compare the two methodologies above in a larger log-linear model. We use the example in Dellaportas and Forster (1996) which is a 2^6 contingency table of risk factors for coronary heart disease presented by Edwards and Havránek (1985). The labels represent: A, smoking; B, strenuous mental work; C, strenuous physical work; D, systolic blood pressure; E, ratio of α and β lipoproteins; F, family history of coronary heart disease. The results in Table 5 provide evidence that Clyde's method fails to detect the most probable models. In fact it seems that there is tendency to support more complicated models. The reason might be that more complicated models are closer to the underlying assumption of constant variance.

Table 5. *Posterior Model Probabilities for Edwards and Havránek data.*

Model	MCMC
AC+BC+AD+AE+CE+DE+F	0.28
AC+BC+AD+AE+BE+DE+F	0.16
AC+BC+AD+AE+BE+CE+DE+F	0.07
AC+BC+AD+AE+CE+DE+BF	0.07
Model	Clyde ($\delta = 1/3$)
AC+AD+BE+BCF	0.16
AC+AD+CD+BE+BCF	0.13
AC+AD+AE+BE+BCF	0.06
AC+AD+CD+AE+BE+BCF	0.05
Model	Clyde ($\delta = 1/2$)
AC+AD+AE+BE+BCF+DEF	0.23
AC+AE+BE+BCF+DEF	0.12
AC+AD+BE+BCF+DEF	0.08
AC+AD+AE+BE+CE+BCF+DEF	0.06

Further experimentation consisted of repeated construction of various marginal tables derived from the Edwards and Havránek data, and visual comparison of the results of both Clyde's and MCMC methodologies. It was revealed that in many cases Clyde's method provides adequate results but sometimes the results are very poor; see for example the 5-dimensional marginal table ABCDE results in Table 6 and the 3-dimensional BCF results in Table 8. In this latter case, Clyde's method visits only the BCF model.

Table 6. *Posterior Model Probabilities for ABCDE Marginal Table*

Model	MCMC	Clyde ($\delta = 1/2$)
AC+BC+AD+AE+CE+DE	0.41	0.06
AC+BC+AD+AE+BE+DE	0.24	0.08
AC+BC+AD+AE+BE+CE+DE	0.13	0.25
AC+AD+AE+BCE+DE	*	0.32
AC+AE+BCE+DE	*	0.08

What is now interesting is to try to detect when and why the method is not working. Table 7 suggests that the assumption of normality together with retaining Σ_{ii} constant is a difficult task

Table 7. *BCF Marginal Table*

	C	No		Yes	
F	B	No	Yes	No	Yes
Negative		235	558	694	94
Positive		33	101	101	25

Table 8. *Posterior Model Probabilities for BCF Marginal Table*

Model	MCMC	Clyde ($\delta = 1/2$)
BC+F	0.77	*
BC+FB	0.21	*
BC+FC	0.02	*
BCF	*	1.00

and in some rather ill-conditioned problems nearly impossible to achieve. Some ideas to deal with it are as follows. If we use a hierarchical stage $\sigma^2 \sim IG(a, b)$ then the algorithm requires sampling from the full conditional density of σ^2 which involves quite a lot of computing effort. If we use, as Clyde suggests, the prior

$$\beta_j \sim N(0, c^2\sigma^2(X'X)^{-1})$$

then we can integrate out σ^2 and follow the steps of Smith and Kohn (1996). Another idea is to model outliers with heavy-tailed error distributions. This would lead to more robust methods but again, the price will be that the algorithm will be less eficient. Finally, a definite way to proceed is with the usual suspects: diagnostic checking, residual analysis, graphical displays.

It gives me great pleasure to congratulate the author for a very nice paper.

E. I. GEORGE *(University of Texas at Austin, USA)*

I would like to begin by congratulating the author on a very stimulating synthesis of recent work on model averaging and model search. In my opinion, an essential message of the paper is that, in large problems, the desired Bayesian model average or posterior mean cannot be computed exactly and so must be approximated. For example, in variable selection with Y and $X_1, \ldots, X_p$, when p is large (e.g. larger than 40), one cannot compute

$$f(\Delta) = \sum_{\gamma \in \Gamma} f(\gamma|Y) f(\Delta_\gamma|Y, \gamma)$$

in (3) because the 2^p models in Γ are too many for exhaustive enumeration. Instead, the author and others approximate $f(\Delta)$ by something of the form

$$\hat{f}(\Delta) = \sum_{\gamma \in S} \hat{f}(\gamma|Y, S) f(\Delta_\gamma|Y, \gamma)$$

where S is a manageable subset of models. (In general, $f(\Delta_\gamma|Y, \gamma)$ is also approximated by $\hat{f}(\Delta_\gamma|Y, \gamma)$ but I will ignore that aspect here). Model search enters the picture as a strategy for selecting S. I would like to focus my discussion on issues concerning the use of model search for this purpose, and how this can affect the approximation.

I would like to begin by discussing a posterior phenomenon which I call *dilution*. Basically, dilution occurs when posterior probability is spread out over many similar models. For example, suppose only X_1 and X_2 are considered as possible linear predictors of Y, yielding posterior probabilities:

Variables in γ	X_1	X_2	X_1, X_2
$\pi(\gamma\|Y)$	0.3	0.4	0.15

Suppose now a new potential variable X_3, very highly correlated with X_2 but not with X_1, is introduced. If the model prior is elaborated in a sensible way, as is discussed below, the posterior may well look something like

Variables in γ	X_1	X_2	X_3	X_1, X_2	X_1, X_3	X_1, X_2, X_3
$\pi(\gamma\|Y)$	0.3	0.2	0.2	0.05	0.05	0.05

The probability allocated to X_2 and (X_1, X_2) has been "diluted" across the new models containing X_3. Is dilution a reasonable phenomenon? I would argue yes, because dilution does not change the allocation of posterior probability across neighborhoods of similar models. In the example above, the introduction of X_3 has not really added any new models to the mix. Instead, models containing X_3 are merely equivalent substitutes for the corresponding models containing X_2. Introducing X_3 has essentially resulted in relabeling a set of equivalent models. The probability of such a set should not increase as a result of this relabeling, and it is dilution that prevents this from happening.

Can dilution be controlled? Again the answer is yes, because dilution results from prior probability allocation across models. This follows by noting that $f(\gamma|Y) \propto f(Y|\gamma)\pi(\gamma)$ and the likelihood $f(Y|\gamma)$ does not depend on the model space under consideration. Thus the prior needs to be adjusted to dilute properly to compensate for the presence of similar models. Note that the commonly used uniform prior on all models does not yield satisfactory dilution. Indeed, under uniform priors, the probability of a set of similar models can be increased merely by adding more of the same models. An interesting example of a prior construction which dilutes naturally is the tree generating process priors proposed for Bayesian CART by Chipman *et al*. (1998).

Although dilution seems to be a reasonable phenomenon, it can lead to paradoxical conclusions when trying to interpret individual model probabilities. This happens because dilution changes the relative posterior probability allocation to individual models. In the example above, for instance, the relative strengths of X_1 and X_2 should not depend on whether X_3 is considered. When dilution is present, model selection based on the largest posterior probabilities can be unreliable when there is dilution. If selection is desired in such a case, it may be best to use Bayes factors, which corresponds to Bayesian selection under a uniform prior over the model space.

For the purpose of obtaining a good model averaging approximation $\hat{f}(\Delta)$ of $f(\Delta)$, dilution is not a problem if the model subset S is randomly sampled from $f(\gamma|Y)$. Although i.i.d. sampling can more easily be used to obtain unbiased estimates of $f(\Delta)$, stochastic search strategies using MCMC methods can also approximately serve this purpose. Random sampling from $f(\gamma|Y)$ will avoid the dilution pitfalls because neighborhoods of models will tend to be represented in S according to their posterior probability. The substitution of one similar model

for another will not have a substantial affect on the model average. In sharp contrast, dilution can have a very detrimental affect on deterministic strategies for choosing S. For example, choosing S to contain models with largest posterior probability $f(\gamma|Y)$, such as happens with Occam's Window, can neglect diluted neighborhoods with many similar models. In a sense, dilution causes skewness of the model posterior distribution. From this perspective, such approximation failures result from a discrepancy between the mean and the mode.

Closely tied to the choice of S is the choice of $\hat{f}(\gamma|Y,S)$. When no special structure is available, and S has been randomly sampled, a natural choice is the relative frequency of γ in the sample. However, in many problems, the choice of conjugate forms for the parameter priors allows for analytical simplification to obtain $f(\gamma|Y) = Cg(\gamma)$ with some easily computable form g. In this case, each sampled γ includes the information $g(\gamma)$. When S is randomly sampled, a natural estimate of $f(\Delta)$ is then

$$\hat{f}_g(\Delta) = \sum_{\gamma \in S} \hat{f}_g(\gamma|Y,S) f(\Delta_\gamma|Y,\gamma)$$

where

$$\hat{f}_g(\gamma|Y,S) = g(\gamma)/g(S)$$

is just the renormalized value of $g(\gamma)$.

When available, the estimator $\hat{f}_g(\Delta)$ appears to be superior to using relative frequencies. For example, under i.i.d. sampling, it approximates the best unbiased estimator of $f(\Delta)$. To see this, consider $\bar{f}(\Delta) = \sum_{\gamma \in S} \hat{f}_{freq}(\gamma \,|\, Y,S) f(\Delta_\gamma \,|\, Y,\gamma)$ where $\hat{f}_{freq}(\gamma \,|\, Y,S)$ is the relative frequency of γ. Note that $\bar{f}(\Delta)$ is unbiased for $f(\Delta)$. Since S (together with g) is sufficient, the Rao-Blackwellized estimate $E(\bar{f}(\Delta) \,|\, S)$ is best unbiased. But when n is large, $E(\bar{f}(\Delta) \,|\, S) \approx \hat{f}_g(\Delta)$.

Although sampling-with-replacement is not needed for the calculation of $\hat{f}_g(\Delta)$ when $f(\gamma|Y) = Cg(\gamma)$ with g computable, the relative model frequencies can still be used to estimate C. This can be useful because (i) it yields improved estimates of the probability of individual γ values, $\hat{f}(\gamma|Y) = \hat{C}g(\gamma)$, and (ii) it allows for an estimate of the total visited probability,

$$\hat{f}(S|Y) = \hat{C}\, g(S).$$

Note that $\hat{f}(S|Y)$ can provide valuable information about when to stop a MCMC simulation.

Interestingly, Bayesian methods appear to be unavailable for estimating C from the sampled information in S because there is no likelihood for C. This is a consequence of the fact that the value of C, although unknown, does not affect the probability distribution of S. Fortunately, frequentist methods can still be used to estimate C. In particular, George & McCulloch (1997) propose the following. Let A be a *preselected* subset of γ values. If $\gamma_1, \ldots, \gamma_K$ is obtained by MCMC sampling from $f(\gamma|Y)$, a consistent estimate of C is obtained as

$$\hat{C} = \frac{1}{g(A)K} \sum_{k=1}^{K} I_A(\gamma^{(k)})$$

where $I_A(\,)$ is the indicator of the set A. If $\gamma_1, \ldots, \gamma_K$ were i.i.d. , then $\mathrm{var}(\hat{C}) = (C^2/K)(1 - f(A|Y))/f(A|Y)$, suggesting that A should be chosen so that $f(A|Y)$ is large. It is also desirable to choose A such that $I_A(\gamma)$ will be inexpensive to evaluate. Current joint work with my student Linghua Peng extends and generalizes these ideas to obtain improved estimators of C, and will be reported on elsewhere.

STEPHEN FIENBERG *(Carnegie Mellon University, USA)*

I found the author's approach to both the model average and model selection problems quite fascinating, and was pleased to see the link to other statistical literature such as that arising from sampling from finite populations. Because of the formal relationship between sampling theory and experimental design, one might think that there would be some scope for use of ideas from the classical design of experiments, e.g., fractional factorials, in this particular Bayesian context. Has the author considered such possibilities?

Doing model search for a generalized linear model with $p = 35$ possible predictors is an impressive feat, but I have colleagues whose problems are much more complex and who justifiably begin with $p = 2,000$ or even more predictors, while they are hoping to work ultimately with a value of p one or two orders of magnitude smaller. They would like to achieve the reduction in dimensionality in an automated or at least semi-automated way. To what extent could we expect the methods described in the paper to scale up for such problems?

Finally, the author describes several interesting applications involving loglinear and logistic models for count data. In my experience, the most difficult contingency table problems are those with large values of p, where virtually all of entries of the resulting cross-classification tend to be small, with many cells containing counts of 0 or 1. The distributional approach and approximations used by the author in her examples do not seem especially appropriate for such circumstances. Could she comment on this class of statistical problems and how it relates to the methods in the paper?

PAOLO GIUDICI *(University of Pavia, Italy)*

The paper is an important contribution in the rather challenging area of model determination. I would like to add one comment and one suggestion.

The comment concerns the proposed orthogonalisation procedure. The author shows that orthogonalising the explanatory variables may lead to a considerable increase in computational efficiency. I agree with this; however, as is well-known in multivariate analysis, doing so one typically loses in interpretation: how can we elicit a prior distribution on the regression coefficient of a principal component? Furthermore, in situations where one is interested in *interactions* among the explanatory variables (for instance in log-linear models), interaction effects are typically lost. Finally, it is known that conditional independences among the explanatory variables may considerably improve computational efficiency, via *local computations* (for instance in graphical models, see e.g. Lauritzen, 1996). I wonder whether orthogonalisation is a real computational advantage in this case.

The suggestion concerns the practical implementation of the proposed methodology. It is clearly very important to measure what we can gain in computational efficiency, for instance with respect to a "non orthogonal" reversible jump MCMC approach. It would be interesting to see a comparison in terms of convergence performance, both for model parameters and for model averaged quantities. A recent paper (Brooks and Giudici, 1998) has proposed a convergence diagnostic which can be used for both purposes.

I. NTZOUFRAS *(Athens University of Economics and Business, Greece)*

This nice work provides us with a valuable and fast tool for the calculation of posterior weights used in Bayesian model selection and model averaging. It uses clever ideas similar to Foster and George (1994) where they use information criteria in orthogonal data to select variables rather than models.

In model selection, our aim is either the interpretation of casual relationships or prediction of future outcomes. Transformation to orthogonality changes the model space and therefore the simple and natural model interpretation. This makes the method appropriate only for model

averaging where prediction is of main interest. Since, in real life practice, our interest may lie in examining casual relationships, I wonder in which cases we can avoid orthogonalization and how robust is the method to deviations from orthogonality.

The author's proposed method is ideal for analysis of variance model selection using sum-to-zero constraints. An ANOVA model can be written as $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$, where $\boldsymbol{y}$ is the data vector, $\boldsymbol{X}$ is the orthogonal design matrix and $\boldsymbol{\beta}$ is the parameter vector. The design matrix and the parameter vector can be divided in sub-matrices $\boldsymbol{x}_i$ and sub-vectors $\boldsymbol{\beta}_i$ of dimension $n \times d_i$ and $d_i \times 1$, respectively, which correspond to a i factor or interaction term. Assuming known σ^2 and prior $\boldsymbol{\beta}_i \sim N_{d_i}(\boldsymbol{\theta}_i, \boldsymbol{V}_i)$ the sampler steps are

$$\gamma_i \,|\, \boldsymbol{y}, \sigma^2, \boldsymbol{\gamma}_{\setminus i} \sim \text{Be}(\pi_i), \; \frac{\pi_i}{1-\pi_i} = \frac{p(\gamma_i = 1, \gamma_{\setminus i})}{p(\gamma_i = 0, \gamma_{\setminus i})} O_i$$

with

$$O_i = \left(\frac{|\, \boldsymbol{x}_i^T \boldsymbol{x}_i/\sigma^2 + \boldsymbol{V}_i^{-1} \,|}{|\, \boldsymbol{V}_i^{-1} \,|} \right)^{-1/2} \exp\left(\frac{1}{2} \boldsymbol{A}_i^T (\boldsymbol{x}_i^T \boldsymbol{x}_i/\sigma^2 + \boldsymbol{V}_i^{-1})^{-1} \boldsymbol{A}_i - \frac{1}{2} \boldsymbol{\theta}_i \boldsymbol{V}_i^{-1} \boldsymbol{\theta}_i \right),$$

$$\boldsymbol{A}_i = (\boldsymbol{x}_i^T \boldsymbol{x}_i \hat{\boldsymbol{\beta}}_i/\sigma^2 + \boldsymbol{V}_i^{-1} \boldsymbol{\theta}_i),$$

where O_i is the Bayes factor to include the i term, and $\hat{\boldsymbol{\beta}}_i$ are the maximum likelihood estimates of the parameters of the i term, $\boldsymbol{\gamma}_{\setminus i}$ is the vector of γ excluding γ_i.

An alternative prior specification is $\boldsymbol{\beta}_i \sim N_{d_i}(0, c^2(\boldsymbol{x}_i^T \boldsymbol{x}_i)^{-1}\sigma^2)$ and $\sigma^{-2} \sim \Gamma(a_0, b_0)$, resulting to the Gibbs sampler steps

$$\gamma_i \,|\, \boldsymbol{y}, \sigma^2, \boldsymbol{\gamma}_{\setminus i} \sim \text{Be}(\pi_i),$$

$$\frac{\pi_i}{1-\pi_i} = \frac{p(\gamma_i = 1, \gamma_{\setminus i})}{p(\gamma_i = 0, \gamma_{\setminus i})} (c^2+1)^{-d_i/2} \exp\left(\frac{1}{2\sigma^2} \frac{c^2}{c^2+1} F_i \right), \; F_i = \hat{\boldsymbol{\beta}}_i^T \boldsymbol{x}_i^T \boldsymbol{x}_i \hat{\boldsymbol{\beta}}_i,$$

and

$$\sigma^{-2} \,|\, \boldsymbol{y}, \boldsymbol{\gamma} \sim \Gamma\left(a_0 + n/2, b_0 + (\boldsymbol{y}^T \boldsymbol{y} - \frac{c^2}{c^2+1} \sum_{i=1}^{p} \gamma_i F_i)/2 \right).$$

The author has drawn attention only to the former case. I wonder whether the latter is also fast and flexible.

As a final comment, I would like to point out that Bayesian model selection in GLMs can be routinely applied using either 'sophisticated' techniques such as reversible jump (Green, 1995), or easy-to-use Gibbs samplers which can be implemented in any standard MCMC software such as BUGS; for details see Dellaportas *et al.* (1998).

REPLY TO THE DISCUSSION

I would like to thank the discussants for their interesting comments and questions. Several issues were raised relating to orthogonalization, choice of prior distributions, and interpretations with orthogonal variables, as well as estimation of posterior model probabilities based on a sample of models. Rather than addressing each set of comments individually, I will try to address them by discussing the issues of prior choice, robustness to non-orthogonality, and sampling issues.

Prof. Giudici asked how to specify prior distributions on regression coefficients after orthogonalizing the explanatory variables and whether interaction effects are lost in such a transformation. Orthogonal designs naturally arise in applications such as balanced factorial experiments

and loglinear models with the parameterization based on the sum-to-zero constraints. One does not lose the ability to test for interactions by using these orthogonal designs/parameterizations. In other situations, one may restrict the model to allow interactions only if the "main" effects are included. In such a case, orthogonalizing the variables in the order of main effects then interactions via Gram-Schmidt orthogonalization leads to the usual interpretation of the coefficient for the interaction. Thus, one can still learn about interactions with an orthogonal design.

Specifying a prior distribution is often difficult enough when dealing with the original variables, and even more difficult if the variables do not have an intuitive meaning. All of the methods for specifying prior distributions on the regression coefficients in the original variables, such as default priors, priors based on historical data, or a subjective prior specification, still apply with orthogonal variables. If one can specify a prior distribution on the regression coefficients with the original variables (or on the mean using a coordinate-free approach), then one can easily obtain the prior distribution for the coefficients for the orthogonalized variables as the transformation to orthogonal explanatory variables is a linear transformation (Clyde and Parmigiani 1996). For example, if we start with a mean $\boldsymbol{X\beta}$, and the orthogonal variables are $\boldsymbol{W} = \boldsymbol{XU}'$ where $\boldsymbol{U}'\boldsymbol{U}$ is the identity matrix, then the resulting mean is $\boldsymbol{XU}'\boldsymbol{U\beta} = \boldsymbol{W\alpha}$, where $\boldsymbol{\alpha} = \boldsymbol{U\beta}$. When the orthogonalization is based on using principal components and the prior distribution for $\boldsymbol{\beta}$ is normal with covariance matrix $c(\boldsymbol{X}'\boldsymbol{X})^{-1}$ (a commonly used default choice), then the resulting covariance for $\boldsymbol{\alpha}$ is diagonal. If, in addition, one standardizes $\boldsymbol{W}$ so that $\boldsymbol{W}'\boldsymbol{W} = I$ by dividing each column by the square root of the corresponding eigenvalue, then the resulting prior covariance matrix is proportional to the identity matrix. It does not matter which coordinate system is used initially. Related to the choice of prior distributions on $\boldsymbol{\beta}$, Prof. Dellaportas asks whether the prior distribution for $\boldsymbol{\beta}$ should depend on the transformation determined by δ. The answer is no, as the prior distribution for $\boldsymbol{\beta}$ should reflect the GLM parameterization; δ determines an approximate model used to derive the proposal distribution. One can take into account the effect of the transformation which leads to different normal error variances σ by changing the value of c used to derive the proposal distribution.

A more difficult issue concerns how to construct the prior distribution on the model space. While one can obtain the prior distribution on the regression coefficients by a simple change of variables, the prior and posterior distributions on the model space depend on the model parameterization or coordinate system. The uniform prior distribution is often the default choice in most BMA applications, as the number of variables and correlation structure often makes subjective prior elicitation for 2^p models intractable. When variables are correlated, however, it is not clear that the independent uniform prior on the model space is sensible. This is directly related to the dilution issue that Prof. George discussed. For example, suppose we start off with one variable, X_1, and consider two models $(\{1\}, \{1, X_1\})$ where $\{1\}$ represents the model with just an intercept. Assign both models equal prior probabilities 0.5. Now consider adding a second variable X_2 that is highly correlated (or even perfectly correlated) with X_1, with possible models $(\{1\}, \{1, X_1\}, \{1, X_2\}, \{1, X_1, X_2\})$ and uniform prior probabilities $(0.25, 0.25, 0.25, 0.25)$. The total prior probability mass of the last three models is 0.75, while, if X_2 is really a proxy for X_1, the mass should be closer to 0.5, as these three models are approximately equivalent (or exactly with perfect collinearity), and should have the same weight as in the original model space with just X_1. The uniform prior has inflated the importance of X_1 while a more sensible prior should dilute the 0.5 mass over the three equivalent models, in order to be consistent with the first prior distribution. This inflation in the prior distribution carries over to the posterior distribution and hence has an effect on both model selection and model averaging.

As an alternative, consider what happens when we change to orthogonal variables $Z_1 = 0.5(X_1 + X_2)$ and $Z_2 = 0.5(X_1 - X_2)$. The resulting model space is $(\{1\}, \{1, Z_1\}, \{1, Z_2\}, \{1, Z_1, Z_2\})$ with uniform prior probabilities (0.25, 0.25, 0.25, 0.25). With highly positively correlated variables, Z_1 and X_1 are effectively measuring the same quantity, but Z_2 should be close to a constant. If we look at the mass on sets of equivalent models, we find that there is equal mass on $(\{1\}, \{1, Z_2\})$ and $(\{1, Z_1\}, \{1, Z_1, Z_2\})$, corresponding to the probabilities that we assigned originally to the two models $(\{1\}, \{1, X_1\})$. In this case, the independent uniform prior probabilities have diluted in an appropriate manner, and result in the appropriate dilution of the posterior model probabilities. While model selection can be affected by the dilution caused by adding a proxy variable, the posterior means under model averaging with $(\{1\}, \{1, X_1\})$ will be approximately the same as with $(\{1\}, \{1, Z_2\}, \{1, Z_1\}, \{1, Z_1, Z_2\}$.

In model selection where the aim is interpretation rather than prediction, using arbitrary principal components or other orthogonal variables does not make sense for the goals of the analysis. It is not always clear, however, that using the original variables leads to a simpler and more natural interpretation of the results, as suggested by Dr. Ntzoufras, because of the difficulties that can occur with multicollinearity. For example, Weisberg (1985) discusses an analysis from the Berkeley Guidance study for predicting somatypes where three of the explanatory variables are positively correlated (weights at age 2, 9 and 18). Using the original variables in the regression results in an unexpected sign on the regression coefficient for weight at age 2, and the conclusion that heavier girls at age 2 have thinner somatypes at age 18. He describes alternative reparameterizations using the weight at age 2, and then the weight gains between age 9 and 2, and between 9 and 18, and finds that while the coefficient for weight at age 2 still has the wrong sign, it is no longer statistically significant, leading to a simpler interpretation and a more natural conclusion. Weisberg concludes with "The interpretation of the effect of a variable depends not only on the other variables in the model, but also upon which linear transformation of those variables is used." One advantage of orthogonal variables is that the values of the coefficients do not depend on which other variables are included in the model. Principal components analysis in the example in Weisberg results in new variables with simple interpretations, with three components that approximately measure the average, linear, and quadratic time trends in weight. Many of the PCs in the PM_{10} example also have simple interpretations as average temperature, average pressure, the change in temperature and change in pressure, thus we can obtain meaningful interpretations of effects with the computational advantage of orthogonality. After either model selection or model averaging with orthogonal variables, one can still compute the coefficients for each of the original variables.

Both Prof. Dellaportas and Dr. Ntzoufras raise the issue of how robust the method is to deviations from orthogonality, and Dellaportas shows examples for the Edwards and Havránek data where some of the approximations fail. He suggests that the assumption of normality along with the constant variances Σ_{ii} may explain why the method is not working. However, the cell counts in Table 5 of Dellaportas are large enough for approximate normality and the variance-stabilizing transformation to result in constant variance, yet the approximation gives very different answers than those obtained by MCMC. The problems, in my view, arise because of dependence in the posterior distribution of $\boldsymbol{\beta}$, even when posterior normality is satisfied. The methods in the paper assume that the posterior distribution for $\boldsymbol{\gamma}$ can be adequately approximated by a model of independence, using a product of Bernoulli distributions. As pointed out in the paper, high correlations among the MLEs from the GLM analysis are a simple indication of posterior dependence. As the correlations increase, the accuracy of the approximation decreases, which also decreases the efficiency of MCMC algorithms using the approximation as a proposal distribution.

With the Savage-Dickey ratio discussed in section 4.4, the posterior odds that γ_j equals one were approximated by

$$O_j = \frac{p_j}{1-p_j} \frac{f_{\beta_j}(0)}{f_{\beta_j}(0|\boldsymbol{Y})},$$

where $f_{\beta_j}(0|\boldsymbol{Y})$ is the marginal posterior density for β_j obtained from the full model and $f_{\beta_j}(0)$ is the marginal prior density for β_j where both densities are evaluated at 0. As these marginals are obtained from the full model, this should really be interpreted as the posterior odds that γ_j equals one conditional on $\gamma_k = 1$, for all k not equal to j. In the normal linear model with known σ and orthogonality, this does not depend on the other values of γ_k, as the joint posterior distribution of $\boldsymbol{\beta}$ factors into the product of the marginal distributions. In GLMs this is not the case, and even with orthogonality, the distribution of β_j depends on the values of the other $\boldsymbol{\beta}$s. Assume for now that the joint posterior distribution of $\boldsymbol{\beta}$ can be well-approximated by a multivariate normal distribution centered at the MLE and with a covariance matrix based on the inverse Fisher information. In the saturated loglinear model, the covariance is $(\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{X})^{-1}$. In the Healy data, the range of $\boldsymbol{Y}$ is 4 to 22 so the covariance matrix is near diagonal, consistent with posterior independence of $\boldsymbol{\gamma}$. In fact, we found that the model of independence for the model space was a very good approximation to the actual posterior distribution of $\boldsymbol{\gamma}$ in the Healy data.

In the Edwards and Havránek example, the range of $\boldsymbol{Y}$ is much greater, 0 to 145, and the correlation matrix from the GLM analysis is not close to diagonal. If we compare the t-statistics for models with up to three-way interactions, we find that, under the variance-stabilizing transformation, the interaction BCF has a t-statistic over 8, which results in its approximate posterior probability of inclusion being near one. In the GLM analysis, the t-statistic is 0.78, so that the approximate posterior inclusion probability is much smaller. The difference is not as extreme in the case of the DEF interactions, with t-statistics of -2.99 and -2.05 for the variance-stabilizing transformation and the GLM analysis respectively. This explains some of the differences in Tables 3, 4, and 6 in Dellaportas's discussion.

Differences in the results do not disappear when the table is collapsed, as in the BCF table summarized by Dellaportas. Here, cell counts are large enough that we expect posterior normality to be a reasonable assumption; however posterior dependence between components of $\boldsymbol{\beta}$ is problematic. In this table, the parameter estimates with the variance-stabilizing transformation and the GLM parameter estimates are very similar, except for the BCF interaction, which is highly significant under the variance-stabilizing transformation but not in the GLM. There are large correlations between the MLEs (e.g. 0.7 which is much higher than what the paper recommends for implementing the method). Table 9 shows posterior model probabilities estimated using the variance-stabilizing and the Savage-Dickey ratio approximations with the independent Bernoulli model, and estimates using the Savage-Dickey ratio as a proposal distribution in a RJMCMC algorithm.

The approximate model probabilities computed using the independent Bernoulli approximation with the variance-stabilizing transformation put almost of the mass on the model with the three factor interaction. When this is used as a proposal distribution, the Markov chain will be very slow to converge, because of the dominance of this one model. The probability of the most probable model under the approximation can be easily calculated before running the MCMC chain and could identify this problem. The approximate model probabilities computed using the independent Bernoulli approximation with the Savage-Dickey estimates are better, and rank the models identically to the MCMC methods. But it is clear that, because of the correlation among the MLEs, the elements of $\boldsymbol{\gamma}$ will not be independent *a posteriori*, and the discrepancy between this approximation and the MCMC methods is not surprising. Using the

Table 9 *Estimated posterior model probabilities for the BCF marginal table. An asterisk denotes a probability less than 0.01.*

Model	Dellaportas MCMC	Savage-Dickey MCMC	Savage-Dickey without MCMC	variance-stabilizing without MCMC
$[BC][F]$	0.77	0.77	0.50	*
$[BC][BF]$	0.21	0.20	0.44	*
$[BC][CF]$	0.02	0.02	0.03	*
$[BC][BF][CF]$	*	0.01	0.03	*
$[BCF]$	*	*	*	1.00

Savage-Dickey ratio approximation as a proposal distribution does give results that are virtually identical to Dellaportas, but can be computed in much less time.

Going back to the more challenging problem of the larger ABCDEF table, we used the Savage-Dickey approximation as the proposal distribution for the RJMCMC algorithm. The approximate probabilities of inclusion for the proposal distribution for the two-way interactions AC, AD, AE, BC, DE are all over 0.96. The other terms identified by Dellaportas in the most probable models are CE, BE, and BF, which have approximate inclusion probabilities of 0.39, 0.17, and 0.29, respectively. The relative importance of the two groups of terms does agree with Dellaportas and Forster (1996). While the proposal probability for including the DEF interaction is 0.167, we find that, based on the MCMC output, it is more likely to be included than DE, and that there is high correlation between the two terms. Substituting DEF for DE, our results agree with the most probable models identified by Dellaportas. One might be concerned that the MCMC algorithm is getting stuck in models with DEF, however, the model $[AC][BC][AD][AE][CE][F]$ with $[DE]$ has a residual deviance of 64.9 while the non-hierarchical model $[AC][BC][AD][AE][CE][F]$ with the interaction DEF has a residual deviance of 62.3. As both models have the same number of parameters, there is slight evidence (using the Schwartz criterion) in favour of the model with DEF. If one restricts the class of models to hierarchical models rather than all log-linear models, then the two methods agree. If there is evidence of high correlation (which is easy to check from the GLM analysis), then, as is well known, an independent proposal distribution can be inefficient. In the loglinear models, our next step is to develop a more efficient approximation that takes into account the hierarchical structure and accounts for the dependence.

As seen above, different MCMC implementations can lead to different answers, although, theoretically, they should reach the same stationary distribution if run long enough. Dr. Ntzoufras contrasts two Gibbs sampling schemes in the normal linear model. The first Gibbs sampler treats σ^2 as known and only generates $\boldsymbol{\gamma}$, as in Section 3 of the paper. In the second case, there is an inverse gamma prior distribution on σ^2 and it is assumed that $V_i = c^2(\boldsymbol{x}_i^T\boldsymbol{x}_i)^{-1}\sigma^2$ (Clyde *et al.* 1998). Using the same choice of V_i in both, the two Gibbs samplers result in equivalent full conditional distributions for γ_i. Thus the only real difference between the two samplers is whether one needs to sample from the full conditional distribution for σ^2. As the additional time to generate σ^2 is only slightly more than the time to generate a single γ_i, in large problems both Gibbs samplers take the same order of time to run. In the case of unknown σ^2, one can run an alternative Gibbs sampler that generates each γ_i from the posterior conditional distribution of $\gamma_i|\gamma_{(i)}, \boldsymbol{Y}$ after integrating out σ^2 (George and McCulloch 1997). In this case the (γ_i) are no longer independent even with an orthogonal design; however, it is unclear whether this sampler is more or less efficient than the second Gibbs sampler that generates $\boldsymbol{\gamma}|\sigma^2, \boldsymbol{Y}$ and $\sigma^2|\boldsymbol{\gamma}, \boldsymbol{Y}$.

As σ^2 is typically unknown, in most regression problems one would use either the second or third Gibbs samplers. However, in applications with wavelets (where there is an orthonormal

basis) speed of algorithms is a serious concern and Bayesian methods must compete with algorithms that run in $O(n)$ time (in wavelets the number of parameters is n). Running any of the three Gibbs samplers described above for M iterations to estimate the posterior mean requires at least $O(nM)$ time ($M > n$). By using an estimate of σ^2, one can bypass using MCMC methods altogether for computing posterior means and variances under model averaging or model selection, so that Bayesian methods are computationally competitive with classical estimators. As n is usually large, the posterior distribution for σ^2 is often highly concentrated around the mode and there is very little gain in efficiency (in terms of mean squared error) using the Gibbs sampler compared to treating σ^2 as known when computing the posterior mean (see Clyde *et al*. 1998). In situations with high levels of noise, small sample sizes or when there is additional prior information about σ^2 that is not overwhelmed by the data, the additional gains in efficiency using the Gibbs sampler can be impressive enough to sway all concerns about the run time of the Gibbs sampler.

Estimated model probabilities using Monte Carlo frequencies can converge very slowly in large problems and posterior dependence can create problems with multi-modalities. Prof. Fienberg asks whether the methods here scale up as problems increase in size. With orthogonality in linear models (as in the wavelet example), instead of an optimization problem with dimension 2^p, the problem is replaced by p one-dimensional problems, and thus scales up linearly with the number of variables. In the case of the loglinear model with 2000 variables that Fienberg describes, orthogonality of the design is not likely to lead to independence of the posterior distribution of $\boldsymbol{\beta}$, and the current approximations will not be accurate enough for use in model averaging on their own. In order to avoid problems of convergence of MCMC algorithms in the GLM setting, we can use other data-dependent reparameterizations that lead to parameter orthogonality. This should lead to better approximations in larger problems; however this does lead to data-dependent prior distributions on the model and parameter spaces. While it is not clear that this is directly useful for the problem of variable selection, it may lead to efficient semi-automatic approaches for dimension reduction and model averaging by converting the problem to p independent ones.

Professor Fienberg also brings up the connection between sampling theory and experimental design, and questions whether ideas from experimental design can be used here. As an alternative to stochastic search of model spaces via MCMC algorithms, ideas from finite population sampling and experimental design (*e.g.*, fractional factorial designs, response surface estimation, sequential designs, and design for computer experiments) may be adapted to BMA. In the methodology outlined in Section 4.5, the posterior model probabilities are re-written in terms of a "meta-model" that is a saturated loglinear model for the 2^p contingency table that represents the model space. This representation is applicable to any variable selection problem, not just normal regression or generalized linear models. Estimating all of the parameters in the "meta-model" is equivalent to enumerating all models and is generally intractable. We can approximate the posterior model probabilities by setting some higher order terms in the meta-model to zero. The resulting problem of how to choose the models to best estimate the parameters of the meta-model falls within the usual domain of optimal experimental design. I have explored using 2^k fractional factorial designs for estimating "main effects" (the model of independence) as well as some orthogonal array designs, but have found that the designs obtained by the approach outlined in Section 4.5 lead to smaller Kullback-Leibler divergences. The approach in Section 4.5 informally uses the normal approximations or Savage-Dickey density ratios to choose a set of models for the design. More formal approaches may be brought to bear on this design/estimation problem. For example, sequential designs and sequential updat-

ing of the "meta-model" may lead to interesting new approximations for estimating posterior model probabilities.

Given that we are in the same position as the U.S. Census Bureau and cannot enumerate the population (but the Republicans cannot stop us from sampling), what type of sampling plan should we use? MCMC is one approach for random sampling. Even in moderate problems ($p = 15$), it is not clear that MCMC methods actually perform significantly better than simple random sampling (Clyde *et al*. 1996). Given the set of sampled models, Prof. George raised the question of how we should estimate posterior model probabilities, posterior means, and normalizing constants in this situation. While the relative frequencies are unbiased in sampling with replacement, other approaches may be more efficient and take advantage of the information available in the un-normalized posterior model probabilities. The Rao-Blackwellized estimator of Prof. George uses this information and is more efficient than the relative frequencies; however, computing it in in large problems may turn out to be as computationally difficult as enumerating all models. Sampling without replacement appears to be more efficient from a computational perspective, but more work is needed in finding the best estimators or properties of estimators. Normalizing the un-normalized posterior model probabilities over the set of unique sampled models leads to the correct answer when all models are sampled, but is biased otherwise. While there are some twists that make the BMA estimation problem slightly different from the usual finite population sampling problem, finite population sampling methods may yield new insights for improving Bayesian model averaging as we deal with larger scale problems.

ADDITIONAL REFERENCES IN THE DISCUSSION

Brooks, S .P. and Giudici, P. (1998). Convergence assessment for reversible jump MCMC simulations. *Tech. Rep*., University of Pavia.

Dellaportas, P., Forster, J. J. and Ntzoufras, I.(1998). Bayesian Variable Selection Using the Gibbs Sampler. *Tech. Rep*. **39**, Athens University of Economics and Business.

Lauritzen, S .L. (1996). *Graphical models*. Oxford: Oxford University Press