

Model Weights for Model Choice and Averaging

Peter Congdon, Department of Geography, Queen Mary,
University of London. p.congdon@qmul.ac.uk

Abstract A method is suggested to estimate posterior model probabilities and model averaged parameters via MCMC sampling under a Bayesian approach. The estimates use pooled output for J models ($J > 1$) whereby all models are updated at each iteration. Posterior probabilities are based on averages of continuous weights obtained for each model at each iteration, while samples of averaged parameters are obtained from iteration specific averages that are based on these weights. Parallel sampling of models assists in deriving posterior densities for parameter contrasts between models and in assessing hypotheses regarding model averaged parameters. Four worked examples illustrate application of the approach, two involving fixed effects regression, and two involving random effects.

Key Words: Bayesian. Model Weight. Model Averaging. Monte Carlo Estimation.

1 Introduction

Several procedures have been suggested for model choice and diagnosis based on Bayesian principles. For models $j=1, \dots, J$, let m be a multinomial model indica-

tor and θ_j be the parameters under each model. Then formal Bayesian model assessment is based on prior model probabilities $\pi_j = P(m=j)$ and posterior model probabilities $\rho_j = P(m=j|Y)$ after observing data Y , where

$$P(m = j|Y) = \frac{P(m=j)P(Y|m=j)}{\sum_{k=1}^J [P(m=k)P(Y|m=k)]}$$

and $P(Y|m=j)$ are marginal likelihoods.

Approximations to the marginal likelihood that are used in estimating posterior model probabilities include the importance sampling method of Gelfand and Dey (1994), and the method of Chib (1995) based on the identity $P(Y|m=j) = P(Y|\theta_j, m=j)P(\theta_j|m=j) / P(\theta_j|Y, m=j)$. With such methods, marginal likelihoods are estimated by fitting one model at a time and posterior model probability estimates $\hat{\rho}_j$ obtained by comparing marginal likelihood estimates. Furthermore model averaged estimates of quantities Δ are obtained as $\hat{\Delta} = \sum_{j=1}^J \hat{\rho}_j \hat{\Delta}_j$ where $\hat{\Delta}_j$ is the posterior mean of Δ under model j (Hoeting et al, 1999). Alternatively posterior model probabilities may be estimated using model search algorithms that consider more than one model at a time and allow switching between models. Examples are the Carlin & Chib (1995) algorithm, RJMCMC algorithms (Green, 2003), or the Metropolis-Hastings version of the Carlin & Chib (1995) algorithm mentioned by Dellaportas et al (2002). These approaches obtain model averaged quantities by a binary mechanism: if model j is selected at iteration t then $\Delta_j^{(t)}$ has a weight of 1 in the average at that iteration and parameters of other models have zero weight.

By contrast, the present paper proposes a model comparison approach in which several models are considered at a time, but there is no switching. Instead all J models are updated and continuous measures of their relative performance are the basis for model averaging at each iteration. This avoids the problems involved in tuning prior model probabilities or 'jump' proposal densities in product space algorithms to ensure that models are visited sufficiently often (Green & O'Hagan, 1998; Friel & Pettitt, 2006). Such tuning problems mean that model search algorithms may need a very large MCMC sample size in order to visit all models sufficiently often to provide efficient estimates of posterior model probabilities or parameter densities under different models (Chen, 2005). The second main difference follows from the first. Switching between models under product space search or a reversible jump mechanism implies a binary form of model averaging: if model 1 is selected at several successive iterations then the averaged parameters at these iterations are in fact the parameter values sampled from one model only. Under the approach here averaging can be undertaken at each iteration on the basis of continuous quantities, namely weights $w_k^{(t)}$ obtained for all models k , so some weight in the average at each iteration is given to inferior models. This may be important when models are closely competing.

A consequence is that profiles can be obtained of differences in parameters or model fit measures between models. For example, suppose parameters $\theta_1^{(t)}$ and $\theta_2^{(t)}$ are sampled from models 1 and 2 at iteration t of an MCMC chain. Then one may obtain the posterior density of $\Delta = \theta_1 - \theta_2$ by directly monitoring

$\Delta^{(t)} = \theta_1^{(t)} - \theta_2^{(t)}$. This is not possible if averaging is based on a binary mechanism; under a binary mechanism one might form a density for Δ by constructing posterior kernels or some other density estimates (Chen, 2005) for $\theta_1|Y$ and $\theta_2|Y$ and sampling from them. This is more laborious (especially when there are many parameters for which density contrasts are required) than the direct approach possible under parallel sampling. Similarly in spatial health outcome models or survival for models, one could monitor the difference in the regression effect of a known risk factor between a model (model 1) with a random effect (or "frailty" in survival applications), and a model (model 2) without such an effect (Henderson & Oman, 1999). One might expect the impact β of a known risk factor to change (possibly increase) when random effects are added. With the samples $\beta_1^{(t)}$ and $\beta_2^{(t)}$ from the models one may obtain the posterior probability $\Pr(\beta_1 > \beta_2|Y)$ by counting the proportion of iterations where $\beta_1^{(t)} > \beta_2^{(t)}$. One may also obtain posterior probabilities on hypotheses relating to model averaged parameters. If $\theta_1^{(t)}$ and $\theta_2^{(t)}$ are sampled and their average at iteration t obtained as $\theta_a^{(t)} = w_1^{(t)} \theta_1^{(t)} + w_2^{(t)} \theta_2^{(t)}$ then one may obtain a posterior probability $\Pr(\theta_a > A|Y)$, where A is some threshold, by counting the proportion of iterations where $\theta_a^{(t)} > A$.

As pointed out by Wasserman (2001) and Kadane and Lazar (2004), an essential feature of Bayesian model choice and averaging is that it incorporates prior information on the parameters. This is true of the multiple model strategy proposed. The method has a similarity to the frequentist procedures of Burn-

ham and Anderson (2002) in estimating an average model weight, but differs in allowing for prior information on parameters to influence the relative weight attached to different models. Section 2 describes estimates for $P(m=j|Y)$ obtained by this approach, section 3 considers particular implementation issues, section 4 considers some alternative weighting procedures based on AIC and BIC criteria, section 5 considers how model averaging at each iteration is performed, Section 6 contains illustrative examples, while section 7 suggests possible avenues for further development of the paper's approach.

2 Posterior Model Probability Approximation from Multi-Model Sampling

In this section a simulation based method is developed for estimating posterior model probabilities based on independent MCMC sampling of two (or more) different models. The methodology builds on the paper of Congdon (2006) in adopting a more general cross model prior structure and in considering random effects models. Let $\theta = (\theta_1, \dots, \theta_J)$ denote the parameter set over all J models, with dimension (d_1, \dots, d_J) . Assume a model indicator $m \in (1, \dots, J)$, such that given $m=j$, θ_j defines the likelihood for $Y=(y_1, \dots, y_N)$ and Y is independent of parameters of other models $\theta_k, k \neq j$ (Godsill, 2001; Carlin & Chib, 1995). This means that the marginal likelihood given $m=j$ is

$$\begin{aligned} P(Y|m = j) &= \int P(Y|\theta, m = j)P(\theta|m = j)d\theta \\ &= \int P(Y|\theta_j, m = j)P(\theta|m = j)d\theta. \end{aligned}$$

This framework replicates that of section 2.1 in Carlin & Chib (1995). As these authors mention, the form of the cross-model priors $P(\theta_k|m = j, j \neq k)$ within the product $P(\theta|m = j) = \prod_k P(\theta_k|m = j)$ is arbitrary, though proper densities are required in order that $P(\theta|m = j)$ integrates to 1. Here the assumption is made that $P(\theta_k|m = j_1) = P(\theta_k|m = j_2)$ for all $k \neq j_1, k \neq j_2$, so that there will be J cross-model priors. So in a three model situation $P(\theta_3|m = 1) = P(\theta_3|m = 2) = g_3$, $P(\theta_2|m = 1) = P(\theta_2|m = 3) = g_2$ and $P(\theta_1|m = 2) = P(\theta_1|m = 3) = g_1$.

The form of the densities g_j is arbitrary and as long as they are proper they can be made arbitrarily diffuse; that is they may be taken as just proper but for practical purposes are close to being flat. Examples are $N(0, V)$ priors on regression parameters where V is taken very large, and gamma priors, namely $\text{Ga}(1, \varepsilon)$ or $\text{Ga}(\varepsilon, \varepsilon)$, on inverse variance parameters where ε is a very small positive number. To avoid arbitrary choices of just proper cross-model prior, Congdon (2006) suggests a flat option for $P(\theta_k|m = j, k \neq j)$. The flat prior is a simplification which no longer satisfies the requirement that $P(\theta|m = j)$ integrates to 1. However, it may be useful in exploratory data analysis and preliminary model sifting, and can be seen as approximating a limiting situation when just proper cross-model priors are made arbitrarily diffuse. Below we consider a different role for the cross model priors which is similar to that in importance sampling to obtain marginal likelihood estimates (Sinharay & Stern, 2005). Thus the g_j are not

diffuse but in fact taken as approximations to the posterior $P(\theta_j|m = j, Y)$. In fact the terms $H_j^{(t)} = \frac{P(Y|m=j, \theta_j^{(t)})P(\theta_j^{(t)}|m=j)}{g_j^{(t)}}$ used in the model weight derivation described below are related to the estimator used by Gelfand & Dey (1994).

Using weights based on parallel sampling, one may use the output $\{\theta_j^{(t)}, t = 1, \dots, T; j = 1, \dots, J\}$ from the J models to estimate $P(m=j|Y)$. Note that this estimate needs a sample of the same length (say T iterations) from the posteriors $P(\theta_j|Y, m = j)$ of all J models under consideration. Such samples might be obtained by running models in parallel or by running them separately and then pooling the output. This is conceptually distinct from product space search algorithms, such as that of Carlin and Chib (1995), when the model j parameters are only updated when model j is visited.

To see how a Monte Carlo estimate of the posterior model probability is obtained from such output, first write

$$P(m=j|Y) = \int P(m=j, \theta|Y) d\theta = \int P(m=j|Y, \theta) P(\theta|Y) d\theta.$$

Hence a Monte Carlo estimate of $P(m=j|Y)$ is obtainable as

$$\bar{w}_j = \sum_{t=1}^T P(m=j|Y, \theta^{(t)}) / T$$

where $\{\theta^{(t)} = (\theta_1^{(t)}, \dots, \theta_2^{(t)}, \dots, \theta_J^{(t)})$, $t=1, \dots, T\}$ are T samples of parameters in all models. For a particular iteration let

$$w_j^{(t)} = P(m = j|Y, \theta^{(t)}) = \frac{P(m=j, Y, \theta^{(t)})}{P(Y, \theta^{(t)})} = \frac{P(Y|m=j, \theta^{(t)})P(\theta^{(t)}|m=j)P(m=j)}{P(Y, \theta^{(t)})} \quad (1)$$

The numerator in (1) contains the term $P(\theta|m = j) = P(\theta_1, \theta_2, \dots, \theta_J|m = j)$.

From above the cross-model prior is arbitrary and the simplifying assumption

$$P(\theta_h|m = j) = g_h \quad (\text{all } j \neq h)$$

is made, where g_h is a proper density; so $P(\theta_h|m = 1) = P(\theta_h|m = 2) = \dots = P(\theta_h|m = h - 1) = P(\theta_h|m = h + 1) \dots P(\theta_h|m = J) = g_h$.

As in the model choice procedures of Gelfand and Dey (1994) and Carlin & Chib (1995), one might set g_h to be an approximation to $P(\theta_h|m = h, Y)$, namely the posterior density of θ_h given Y and $m=h$. It follows that

$$P(\theta|m = j) = P(\theta_j|m = j) \prod_{h \neq j} P(\theta_h|m = j) = P(\theta_j|m = j)[g_1 g_2 \dots g_{j-1} g_{j+1} \dots g_J].$$

Then

$$w_j^{(t)} = \frac{P(Y|m=j, \theta_j^{(t)})P(\theta_j^{(t)}|m=j) [\prod_{h \neq j} g_h^{(t)}] P(m=j)}{P(Y, \theta^{(t)})} \quad (2)$$

The denominator in (2) can be written

$$\begin{aligned} P(Y, \theta^{(t)}) &= \sum_{k=1}^J P(Y, \theta^{(t)}, m=k) \\ &= \sum_{k=1}^J \{P(Y|\theta^{(t)}, m=k) P(\theta_k^{(t)}|m=k) [\prod_{h \neq k} g_h^{(t)}] P(m=k)\} \\ &= \sum_{k=1}^J \{P(Y|\theta_k^{(t)}, m=k) P(\theta_k^{(t)}|m=k) [\prod_{h \neq k} g_h^{(t)}] P(m=k)\}. \end{aligned}$$

Then

$$w_j^{(t)} = \frac{P(Y|m=j, \theta_j^{(t)})P(\theta_j^{(t)}|m=j) [\prod_{h \neq j} g_h^{(t)}] P(m=j)}{\sum_{k=1}^J P(Y|m=k, \theta_k^{(t)})P(\theta_k^{(t)}|m=k) [\prod_{h \neq k} g_h^{(t)}] P(m=k)}$$

One may divide through by the product of the J cross-model priors $[g_1, g_2, \dots, g_J]$

giving

$$w_j^{(t)} = \frac{[\frac{P(Y|m=j, \theta_j^{(t)})P(\theta_j^{(t)}|m=j)P(m=j)}{g_j^{(t)}}]}{\sum_{k=1}^J [\frac{P(Y|m=k, \theta_k^{(t)})P(\theta_k^{(t)}|m=k)P(m=k)}{g_k^{(t)}}]} \quad (3)$$

For example when $J=2$, one obtains

$$w_1^{(t)} = \frac{[\frac{P(Y|m=1, \theta_1^{(t)})P(\theta_1^{(t)}|m=1)P(m=1)}{g_1^{(t)}}]}{[\frac{P(Y|m=1, \theta_1^{(t)})P(\theta_1^{(t)}|m=1)P(m=1)}{g_1^{(t)}}] + [\frac{P(Y|m=2, \theta_2^{(t)})P(\theta_2^{(t)}|m=2)P(m=2)}{g_2^{(t)}}]} \quad (4)$$

Let $H_j^{(t)} = \frac{P(Y|m=j, \theta_j^{(t)})P(\theta_j^{(t)}|m=j)}{g_j^{(t)}}$ so that (3) becomes

$$w_j^{(t)} = \frac{H_j^{(t)} P(m=j)}{\sum_{k=1}^J [H_k^{(t)} P(m=k)]}. \quad (5)$$

The marginal likelihood estimator of Gelfand & Dey (1994) is obtained using the harmonic average of the $H_j^{(t)}$. In the Gelfand-Dey estimator the g_j plays a role analogous to the reciprocal of an importance function (Rossi et al, 2005, Chapter 6). Specifically to minimise the variance of the inverse $H_j^{(t)}$, namely

$$K_j^{(t)} = 1/H_j^{(t)} = \frac{g_j^{(t)}}{P(Y|m=j, \theta_j^{(t)})P(\theta_j^{(t)}|m=j)}$$

it is desirable that g_j have thin tails relative to the posterior $P(\theta_j|m = j, Y)$. In the $H_j^{(t)}$, by contrast, the role of the g_j is similar to that of an importance function, and so densities g_j which are heavy tailed relative to the posterior may be preferred in order to minimise the variance of the components $H_j^{(t)}$ of $w_j^{(t)}$ (Yuan & Dresdel, 2005). This might mean, for example, using a Student t with low degrees of freedom rather than a normal density for g_j ; or if the parameter θ_j is positive, it could mean using a log Student t rather than lognormal density for g_j . So if a preliminary run provided posterior means and variances (M_j, V_j) for a regression parameter on the real line, one might compare $N(M_j, V_j)$ to $T(M_j, V_j, \nu)$ densities for g_j (where ν is a degrees of freedom parameter). An assessment of stability of inferences according to diffuseness of the g_j may be gained by downweighting the precision $F_j=1/V_j$ so that $N(M_j, 1/F_j)$ densities for g_j are compared to $N(M_j, k/F_j)$ densities, where (for example) $k=10$ or $k=100$.

Let $Q_j^{(t)} = P(Y|m = j, \theta_j^{(t)})P(\theta_j^{(t)}|m = j)$ denote the unnormalized posterior of $\theta_j^{(t)}$ at iteration t . Then in the case of $J=2$ models with equal prior

model probabilities, one may multiply the numerator and denominator in (4)

through by $g_1^{(t)} g_2^{(t)} / (Q_1^{(t)} Q_2^{(t)})$. Then (4) can be expressed as

$$w_1^{(t)} = \frac{K_2^{(t)}}{K_1^{(t)} + K_2^{(t)}}$$

making clear a further link to importance sampling estimators of the marginal likelihood.

3. Implementation Issues

In practice the weights are obtained by considering the relativities in $\log[P(Y, \theta, m = k)]$, involving the logs of the product of the likelihood, of the own model prior $P(\theta_k | m = k)$ and of the product of cross model priors for $\theta_h, h \neq k$. Thus at iteration t , one obtains $q_k^{(t)} = \log\{P(Y | \theta_k^{(t)}, m = k) P(\theta_k^{(t)} | m = k) [\prod_{h \neq k} g_h] P(m = k)\}$, and deviations $\Delta q_k^{(t)} = q_k^{(t)} - \max_k(q_k^{(t)})$, with $w_j^{(t)}$ obtained by exponentiating:

$$w_j^{(t)} = \frac{\exp(\Delta q_j^{(t)})}{\sum_k \exp(\Delta q_k^{(t)})}$$

An indication of the uncertainty attaching to the estimated model weights can be obtained by batch sampling (Lewis and Raftery, 1997). This involves dividing the pooled output over T iterations from the J models into B batches of equal size. A 95% interval for the posterior model probability approximations \bar{w}_j is based on an approximate t density with $B-1$ degrees of freedom for the batch means $\bar{w}_{jb}, b = 1, \dots, B$.

We also briefly mention how the method extends to random effects models with means μ_i of y_i linked to predictors (X_i, Z_i) via $g(\mu_i) = X_i\beta + Z_i b_i$ where g is a link function, and b_i is a vector of random effects with prior $p(b_i|\Phi)$. Bayes factors and/or posterior model probabilities for such models are based on marginal densities with the random effects b integrated out (Fruhwirth-Schnatter, 2004; Sinharay and Stern, 2004; Albert, 1999). Then with $\theta_j = (\beta_j, \Phi_j)$, the posterior model probability is estimated as $\hat{P}(m=j|Y) = \bar{w}_j = \Sigma w_j^{(t)} / T$, where

$$w_j^{(t)} = \frac{P(Y|m=j, \theta_j^{(t)}) P(\theta_j^{(t)}|m=j) [\prod_{h \neq j} g_h^{(t)}] P(m=j)}{P(Y, \theta^{(t)})}$$

with

$$P(Y|m=j, \theta_j) = \prod_{i=1}^N \int P(y_i|\beta_j, b_{ij}) P(b_{ij}|\Phi_j) db_{ij}$$

where b_{ij} are random effects for case i under model j . The cross-model priors g_h involve (proper density) approximations for $P(\theta_h|Y, m=h) = P(\beta_h, \Phi_h|Y, m=h)$. In conjugate models the integrated likelihoods $P(Y|m=j, \theta_j)$ are available analytically, whereas in general linear mixed models devices such as numerical integration, Gaussian quadrature, or importance sampling (Geweke, 1989) are required, and are applied at each iteration to obtain $P(Y|m=j, \theta_j^{(t)})$.

4. AIC and BIC Model Weights

Results from the approximation to posterior model probabilities outlined above may be compared to choices based on the AIC and BIC. For large sample sizes the BIC approximates the Bayes factor (Raftery, 1996) while the AIC has been proposed as an alternative to the BIC that is less likely to choose models that

are too parsimonious (Burnham and Anderson, 2002, p 288; Fitzmaurice et al, 2004, p 177). Let $L_k^{(t)} = P(Y|\theta_k^{(t)}, m = k)$ be the likelihood for model k at iteration t. Similar to the comparisons in section 2 involving the total model likelihoods $P(m = j, Y, \theta^{(t)})$, one may compare $AIC_k^{(t)} = L_k^{(t)} - d_k$ and $BIC_k^{(t)} = L_k^{(t)} - 0.5d_k \log(N)$ between models. One possibility involving the AIC comparisons is the evidence ratio $E_{jk}^{(t)} = (L_j^{(t)} / L_k^{(t)}) \exp(d_k - d_j)$ (Burnham and Anderson, 2002). AIC and BIC based model probabilities are obtained by taking differences against the maximum AIC and BIC models (m^* and m^{**}) at a given iteration

$$\Delta AIC_j^{(t)} = AIC_j^{(t)} - AIC_{m^*}^{(t)}$$

$$\Delta BIC_j^{(t)} = BIC_j^{(t)} - BIC_{m^{**}}^{(t)}$$

and calculating weights

$$\omega_{j,AIC}^{(t)} = \exp(\Delta AIC_j^{(t)}) / \sum_{k=1}^J \exp(\Delta AIC_k^{(t)})$$

$$\omega_{j,BIC}^{(t)} = \exp(\Delta BIC_j^{(t)}) / \sum_{k=1}^J \exp(\Delta BIC_k^{(t)})$$

The average over all iterations of these weights provides estimated AIC and BIC model weights (cf. Brooks, 2002).

5. Iteration Specific Model Averaging

The iteration specific model weights $w_j^{(t)}$ (or those of section 4) may be used in model averaging. For example, consider a quantity $\Delta(\theta_j)$ depending on the parameters. Assuming equal prior model probabilities, namely $\pi_j = 1/J$, the average of $\Delta(\theta_j)$ over models $j=1, \dots, J$ at iteration t is

$$\Delta^{(t)} = \sum_{j=1}^J w_j^{(t)} \Delta(\theta_j^{(t)}) \quad (6)$$

with an average over all iterations

$$\bar{\Delta} = \sum_{t=1}^T \Delta^{(t)} / T.$$

In the case of unequal $\pi_j = \Pr(m=j)$ the appropriate weights in (6) would be $w_j^* = w_j / [J\pi_j]$.

This form of model averaging is carried out at each iteration and so a full posterior summary is obtained for Δ . Because all models are monitored at each iteration one may also monitor differences in particular parameters or other model outputs between models (e.g. differences in predicted means μ_{ij} for cases $i=1, \dots, n$ under models j). One might monitor the differences $\mu_{ik}^{(t)} - \mu_{ij}^{(t)}$ between model means for case i under models j and k , and so obtain posterior probabilities such as $P(\mu_{ik} > \mu_{ij})$. Letting $\mu_{ia}^{(t)}$ denote the model averaged regression mean for case i at iteration t one may also obtain inferences on the model averaged mean vs. a model specific mean, for example, $P(\mu_{ia} > \mu_{ij} | Y)$. This may be important when particular inferences are sensitive to the model used (O'Hagan, 2003).

6. Illustrative Examples

Four worked examples are considered, with implementation via the WINBUGS

package (Spiegelhalter et al, 2003), and with convergence of multiple chains assessed using the scale reduction factors of Gelman et al (1995). The first two examples involve fixed effects regressions, while the third and fourth are random effects models, the first count data relating to transplant surgery (Albert, 1999), and the second relating to suicide deaths in geographic subdivisions of London.

6.1 Binary CHD Symptoms

Selvin (1998, page 243-248) discusses an analysis of binary coronary event data for 100 men aged 58-60 in terms of seven risk factors (X1=height in inches, X2=weight in pounds, X3=systolic blood pressure in mm of HG, X4=diastolic blood pressure, X5=cholesterol in %/100ml, X6=1 for smoker (0 for non-smoker), and X7=1 for type A behaviour (0 otherwise). Selvin concludes that none of the risk factors show a substantial impact on the probability of a coronary event within an 8 year period, namely that a constant only model provides a comparable fit to any models including predictors. The data are sparse in that there are only 16 subjects with $y_i=1$.

Here a logit link is assumed, and priors are based on the maximum likelihood analysis of Selvin, with a $N(-2,100)$ prior on the intercept and $N(0,1)$ priors on the predictor effects. Application of predictive Bayesian variable selection (Meyer and Laud, 2002) suggested a gain in fit over the intercept only model can be gained by adding X1, or X4, or both. For example the model $1+X1$ gives a co-

efficient on X1 with posterior mean 0.29 and 95% interval (0.03,0.57), suggesting an increase in risk with height. So J=4 possible models are considered: 1, 1+X1, 1+X4, and 1+X1+X4, with equal prior model weights $P(m=j)=0.25$. Posterior means and variances of the parameters in the four models, namely $\{(M_{jk}, V_{jk}), j=1, \dots, 4; k=1, \dots, d_j\}$ where $d_1=1, d_2=2$, etc are used in forming cross-model prior densities g_{jk} and are obtained from preliminary single model runs of 10,000 iterations.

A two chain run of 10000 iterations is then made with comparisons between four models to provide $w_j^{(t)}$ at the t^{th} iteration. This run shows convergence at under 500 iterations. The averages \bar{w}_j from iterations 501-10000, namely $\{0.32, 0.53, 0.06, 0.09\}$, slightly favour the model 1+X1. By contrast, the BIC weights $\bar{\omega}_{BIC,j}$, namely $\{0.41, 0.31, 0.17, 0.11\}$ favour the constant only model, while AIC weights $\{0.13, 0.32, 0.18, 0.37\}$ slightly favour the most complex model. The DIC of Spiegelhalter et al (2002) also favours model 4, with the respective values being $\{89.9, 87.2, 88.4, 85.9\}$. The variability in the estimates $\bar{w}_k (j = 1, 2)$ was assessed using batches of 500 iterations (38 batches pooling over two chains). Thus with an inverse $t_{0.05}$ point of 2.03 for 37 d.f., the 95% intervals on \bar{w}_1 and \bar{w}_2 are (0.3235,0.3250) and (0.5252,0.5276).

To assess the role of the specification of the cross-model priors, the posterior variances ξ_j , $j=1, \dots, J$, of the $H_j^{(t)} = \frac{P(Y|m=j, \theta_j^{(t)})P(\theta_j^{(t)}|m=j)}{g_j^{(t)}}$ are compared when univariate normal densities $N(M_{jk}, V_{jk})$ are assumed for g_{jk} , as against Student

t densities $T(M_{jk}, V_{jk}, \nu)$ with $\nu = 2.5, 5, 15, 30$ and 50 degrees of freedom. A summary performance measure is the average percent reduction in the variances of $H_j^{(t)}$ over the four models (for each preset ν) through using Student t rather than normal cross-model priors. The greatest average reduction in variance is for $\nu=5$ though not all the ξ_j are lower than those obtained with normal g_{jk} (the variances are approximately halved for the less favoured models 3 and 4 but very slightly higher for models 1 and 2). To assess stability of inferences to downweighting the precisions $F_{jk}=1/V_{jk}$, $N(M_{jk}, V_{jk})$ densities for g_{jk} are compared to $N(M_{jk}, 10V_{jk})$ densities. This increases the posterior weight on the most complex model 4, and reduces that on the simplest model, though model 2 still has the highest weight: so $\bar{w} = (0.17, 0.56, 0.07, 0.20)$. The variances ξ_j are considerably inflated under the reduced precision option as compared to the $N(M_{jk}, V_{jk})$ option for g_{jk} . So while a heavy tailed density using posterior parameter variances V_{jk} obtained through earlier runs improves the sampling performance of the importance densities g_{jk} , inflating the V_{jk} leads to a deterioration in performance.

Model averaging is illustrated by density plots based on monitoring the iteration specific averages of the subject level CHD probabilities π_i under different forms of weight

$$\pi_{ia}^{(t)} = w_1^{(t)} \pi_{i1}^{(t)} + w_2^{(t)} \pi_{i2}^{(t)} + w_3^{(t)} \pi_{i3}^{(t)} + w_4^{(t)} \pi_{i4}^{(t)}$$

with $\pi_{ij}^{(t)}$ the probability under model j (with normal g_{jk} adopted) and π_{ia} denoting the averaged probability. Inferences on cardiac risk vary between the

models, and O’Hagan (2003, p. 425) mentions the particular problem with model averaging when two or more models are competitive but inferences on particular quantities are model sensitive. A measure of discrepancy between the parameters sampled under different models is provided by the coefficient of variation in the $\pi_{ik}^{(t)}$ and this shows the highest coefficient for subject 28. This subject has a relatively high value on X_1 and the highest value on X_4 of all subjects, and in consequence has widely contrasting mean probabilities of CHD under the four models, namely 0.16, 0.31, 0.46 and 0.64, while the model averaged mean CHD risk is 0.30. Figure 1 shows the Epanechnikov density of the model averaged probability $\pi_{ia}|Y$ for subject 28 obtained from 19000 iteration specific averages. The elongated right tail reflects the high CHD probabilities under models 3 and 4 which have a total posterior probability of 0.15.

6.2 Radiata Pine Data

The second illustration of the methodology of the paper is to the radiata pine data analysed by Song & Lee (2004), Chib & Carlin (1995), and Green & O’Hagan (1998). The observations are y (maximum compression strength parallel to the grain), x (density), and z (resin-adjusted density) for 42 specimens of radiata pine. The alternative models are

$$M1 : y_i = \alpha_1 + \beta_1(x_i - \bar{x}) + u_{1i}, u_{1i} \sim N(0, 1/\tau_1)$$

$$M2 : y_i = \alpha_2 + \beta_2(z_i - \bar{z}) + u_{2i}, u_{2i} \sim N(0, 1/\tau_2).$$

Estimates of the cross-model priors g_{jk} for parameters k and models j are based on a preliminary run. Initially a lognormal approximation is used for the densi-

ties g_{j3} of the precisions τ_j , and univariate normal approximations for the densities of α_j and β_j . The comparison between models is coded as in (4) above. With prior model probabilities $P(m=1)=0.9995$ and $P(m=2)=0.0005$, two chains of 10000 iterations were run with the last 9000 for inference. We find the posterior probability of model 1 is estimated as $\bar{w}_1 = 0.2953$, similar to the value of 0.2914 reported by Green and O'Hagan (1998, p 8).

To assess the role of the specification of the cross-model priors, the posterior variances ξ_j of the $H_j^{(t)} = \frac{P(Y|m=j, \theta_j^{(t)})P(\theta_j^{(t)}|m=j)}{g_j^{(t)}}$ are compared as in the first case study. As heavy tailed options for g_{jk} , a log Student t approximation is used for the density of the precisions τ_j , and univariate Student t approximations for the densities of α_j and β_j . As in section 6.1, alternative degrees of freedom considered are $\nu = 2.5, 5, 15, 30$ and 50. There is less advantage in using markedly heavy tailed Student t cross prior densities here, with the greatest average reduction in variance being for $\nu=30$. To assess stability of inferences to cross-model prior diffuseness, $N(M_{jk}, V_{jk})$ densities for g_{jk} are compared to $N(M_{jk}, 10V_{jk})$ densities (and similarly for the lognormal g densities for τ_j). This slightly increases the posterior weight on model 1 to $\bar{w}_1 = 0.34$ while the variances ξ_1 and ξ_2 are inflated when less precise g_{jk} are used.

6.3 Poisson-Gamma Mixture Models for Hospital Mortality

Poisson-gamma and binomial-beta mixtures are often used in hierarchical analy-

sis of institutional variation, e.g. hospital death rate differences (Kahn and Raftery, 1996; Albert, 1999). Let Y_i denote deaths in hospital i and E_i the deaths anticipated or expected in that hospital on the basis of the patient case-mix. Then for Y taken as Poisson one may model extra-variation using conjugate gamma mixing

$$Y_i \sim \text{Poi}(\lambda_i E_i)$$

$$\lambda_i \sim \text{Ga}(\alpha, \alpha/\mu_i)$$

$$\log(\mu_i) = X_i \beta$$

As α tends to infinity, a Poisson regression $Y_i \sim \text{Poi}(\mu_i E_i)$ is obtained, without an extravariation model being needed. Albert (1999) considers deaths in 94 US hospitals following heart transplant surgery between October 1987 and December 1989, using Bayes factors to criticise alternative model structures obtained with a negative binomial likelihood with gamma random effects integrated out.

In generic form with $\lambda_i \sim \text{Ga}(\alpha, \beta)$, the integrated likelihood for hospital i is

$$P(Y_i|\alpha, \beta) = \frac{\beta^\alpha \Gamma(\alpha + Y_i) E_i^{Y_i}}{\Gamma(\alpha) \Gamma(1 + Y_i) (\beta + E_i)^{\alpha + Y_i}}$$

If $\beta = \alpha/\mu$ (i.e. with no predictors X_i used) then hierarchical models involve taking α an extra parameter, or possibly comparing over a profile of fixed α values, as in Albert (1999, pp 294-295). Here (in model 1) α is first allowed to be a free parameter with proper priors on both α and μ assumed, specifically $\alpha \sim \text{Ga}(1, 0.1)$, $\mu \sim \text{Ga}(1, 0.1)$. A hierarchical model may be compared to a Poisson model (model 2) with mean μ and likelihood

$$P(Y_i|\mu) = \exp(-\mu E_i)(\mu E_i)^{Y_i}/Y_i!$$

Equal prior model weights $P(m=j)=0.5$ are assumed, with cross-model priors for g_1 and g_2 based on log-normal densities with the parameters of those LN densities obtained from a preliminary run.

Albert considers Bayes factors over a profile of fixed values of α and finds the maximum Bayes factor when $\log\alpha$ is approximately 2 (Albert, 1999, Figure 3). He finds a Bayes factor B_{12} of around 10 in favour of the gamma-Poisson mixture as against the fixed mean Poisson. By contrast, when uncertainty in α is allowed for as here, the posterior weights for the mixture model (model 1) against the Poisson model are less decisive. A 10,000 iteration run of two chains converges at under 1000 iterations and gives $(\bar{w}_1, \bar{w}_2) = (0.783, 0.217)$ from the last 9000 iterations. The posterior mean for α from this analysis is 11.03, but with a relatively wide 95% interval from 3.75 to 28.9. As well as the negative binomial likelihood with λ_i integrated out, Simpsons rule (Aitkin and Alfö, 2003) was applied to the poisson-gamma density product with 50 intervals for λ between 0 and 5. This gives the same posterior mean of -181.4 for the integrated likelihood as for the analytic negative binomial likelihood.

To assess cross-model prior specification, the posterior variances ξ_j of the $H_j^{(t)} = \frac{P(Y|m=j, \theta_j^{(t)})P(\theta_j^{(t)}|m=j)}{g_j^{(t)}}$ are compared using lognormal and log Student t approximations for the unknowns (two in model 1, one in model 2). The greatest average reduction in the ξ_j is for a heavy tailed density as compared to the normal,

namely one with $\nu=2.5$.

In a second comparison the approach of Albert is replicated by taking α fixed at 11 (close to its posterior mean in the above analysis). Therefore only one parameter is now unknown in both models 1 and 2. In this situation the mean weights clearly favour the Poisson-gamma mixture model (model 1) with $\bar{w}_1=0.904$. This illustrates the sensitivity of posterior model probabilities to alternative priors, specifically that the posterior probability on a model becomes higher as the informativeness on the priors in its parameters is increased. Setting α to a fixed value amounts to assuming a highly informative prior on α .

6.4 Poisson-LogNormal Model for Suicide Deaths

Here Y_i denotes suicide deaths in 32 London boroughs over 1989-93 with E_i denoting expected deaths using England & Wales age and sex specific suicide rates for 1991 (see Congdon, 2004 for data). With $Y_i \sim \text{Poi}(\lambda_i E_i)$, possible extra-variation is modelled via an unstructured error term in a log link regression for the relative risk λ_i . There are two predictors, X1 an index of socio-economic deprivation, and X2 an index of social fragmentation, both expected to be positive risk factors for suicide (Whitley et al, 1999). So model 1 is

$$\log(\lambda_i) = \alpha_1 + \beta_{11}X_{1i} + \beta_{21}X_{2i} + u_i$$

where $u_i \sim N(0,1/\tau_u)$, a LN(0,2) prior is assumed for τ_u , and N(0,1) priors for the fixed effects. The prior for τ_u implies approximate 90% intervals for u from

-2.3 to 2.3; these are obtained by sampling from $\tau_u \sim \text{LN}(0,2)$ and then drawing 32 unstructured effects (without reference to the observed suicide counts) at the sampled value of τ_u , with the interval approximated by monitoring the 3rd and 30th ranked effects. This interval comfortably includes historical differences in suicide relative risk in these areas, which typically range from 0.3 to 3 (or from -1.2 to 1.1 in terms of log relative risk). The X_j are standardised and $N(0,1)$ priors are also consistent with historic relative risks. For the actual 1989-93 data, the range in relative risk (crude standard mortality ratios) is from 0.65 to 1.7 (or from -0.42 to 0.53 in terms of log relative risk).

A reduced model (model 2) is considered excluding the unstructured random effect, namely

$$\log(\lambda_i) = \alpha_2 + \beta_{12}X_{1i} + \beta_{22}X_{2i}$$

since there seems to be some doubt about the need for such effects. Preliminary analysis of the random effects model shows that for all but one area the posterior 95% credible intervals of u_i straddle zero (Knorr-Held & Rainer, 2001, p 116). Standard penalized fit criteria are inconsistent: the DIC prefers the more complex model 1 (DIC=263) to the Poisson regression (DIC=278), while a BIC based on the effective parameter count (Pourahmadi & Daniels, 2002) prefers model 2 (BIC=282) to model 1 (BIC=293). (The BIC definition used here is the deviance at the parameter mean plus $d_e \log(32)$, where effective parameters d_e under model 1 are 21).

In the parallel sampling stage, with equal prior model weights $P(m=j)=0.5$, the u effects in model 1 are integrated out using Simpsons rule applied to the poisson-log normal density product with 50 intervals for u_i between -0.5 and 0.5. Initial analysis indicated that all 95% intervals for u_i were within this range. Initially the cross model priors use a lognormal approximation to the posterior of τ_u and normal approximations for the regression coefficient densities. A two chain run of 10000 iterations shows a slightly higher weight for the random effects model with $\bar{w}_1=0.69$.

Comparing the posterior variances ξ_j of the $H_j^{(t)} = \frac{P(Y|m=j, \theta_j^{(t)})P(\theta_j^{(t)}|m=j)}{g_j^{(t)}}$ for normal/LN against Student t/Log Student t densities for g_{jk} , shows the lowest ξ_j for $\nu=15$, so again a heavier tailed density than the normal/LN is indicated for the cross-model priors. However, markedly heavy tails are not indicated as the ξ_j are higher than for normal/LN g_{jk} when $\nu=2.5$.

As in the previous example, posterior weights and model averages are sensitive to priors, especially for τ_u . As mentioned above, a more informative prior on τ_u might be justified in terms of the historic patterns of relative risk, and taking $\tau_u \sim \text{LN}(0,1)$ rather than $\tau_u \sim \text{LN}(0,2)$ gives a posterior weight of $\bar{w}_1=0.991$.

To illustrate inferences about parameter differences between models which are simple to implement in the methodology suggested in this paper, the differences

$$\Delta_1|Y = (\beta_{12} - \beta_{11})|Y$$

$$\Delta_2|Y = (\beta_{22} - \beta_{21})|Y$$

are monitored under the more informative prior on τ_u , and the probabilities

$$Q_1 = P(\beta_{12} > \beta_{11}|Y)$$

$$Q_2 = P(\beta_{22} > \beta_{21}|Y).$$

obtained. For example, it might be that the effect of the known risk factors is reduced when random effects (proxying unknown risk factors) are added. We find $Q_1=0.6$ and $Q_2=0.46$, so the effect of X1 appears slightly attenuated in the random effects model. However, density plots of Δ_1 and Δ_2 suggest little change in the significance of the risk factors (Figures 2a and 2b).

7. Discussion

The method described builds on existing work on model choice and averaging using sampling based Bayesian estimation. It has affinities to the frequentist multi-model procedures of Burnham and Anderson (2002) but differs in taking account of prior information on parameters, and can thus be squared with a central Bayesian principle enunciated by Kadane & Lazar (2004). The advantages of the method include the straightforward derivation of full density profiles for model averaged parameters or for parameter differences between models. This extends to cross-model tests regarding parameter differences (e.g. are regression effects attenuated or enhanced when random effects are added to a model). These features were illustrated by the profile of the CHD probability for an individual subject in the first worked example, based on averaging over four

models, and by a comparison of regression coefficients in the fourth. As well as this, the method is straightforwardly implemented in the program WINBUGS.

The worked examples confirm what would be expected from the form of $H_j^{(t)} = \frac{P(Y|m=j, \theta_j^{(t)})P(\theta_j^{(t)}|m=j)}{g_j^{(t)}}$ in the weights $w_j^{(t)} = \frac{H_j^{(t)} P(m=j)}{\sum_{k=1}^J [H_k^{(t)} P(m=k)]}$, namely that heavier tailed Student t densities provide lower posterior variances ξ_j for $H_j^{(t)}$. For a parameter θ_j on the real line this would mean a Student t form for g_j , $T(M_j, V_j, \nu)$ (with ν typically low) rather than a $N(M_j, V_j)$ density where M_j and V_j are posterior means and variances of θ_j from a preliminary run. Adopting heavy tailed densities is distinct from downweighting the precisions $F_j=1/V_j$ which leads to inflation of the variances ξ_j .

A drawback to the method is that when there are a large number of potential models (e.g. linear regressions with a large number of possible predictor combinations) it would be cumbersome to sample from all models in parallel. Probably the best application of the method suggested in this paper is when a relatively small subset of models has been selected by another procedure as having a relatively high posterior probability or providing a similarly good fit, and it is then required to average parameters over that small subset, especially if relatively complex questions require to be answered about model averaged parameters. A preliminary sifting from a large number of models might for instance be made using stochastic variable search selection (George & McCulloch, 1997) or the deviance information criterion (Spiegelhalter et al, 2002).

The analysis rests on pooled posterior output $\{\theta_j^{(t)}, t = 1, \dots, T; j = 1, \dots, J\}$ from all models under consideration, possibly but not necessarily obtained by sampling the models in parallel. This approach provides Monte Carlo estimates of probabilities and model averaged quantities based on averaging at each iteration. Such multi-model sampling might be used in other ways. The above examples demonstrate the sensitivity of posterior model probabilities to priors on hyperparameters governing random effects. One might therefore average both over models and over priors for hyperparameters of varying but still plausible levels of informativeness. Another possibility is averaging over ‘sceptical’ and informative priors, providing the sceptical priors are not diffuse (Higgins & Spiegelhalter, 2002).

8. References

- Aitkin, M, Alfò, M (2003) Longitudinal analysis of repeated binary data using autoregressive and random effect modelling, *Stat. Modelling*, 3, 191-203
- Albert, J (1999) Criticism of a hierarchical model using Bayes factors, *Stat.in Med.* 18,287-305
- Brooks, S (2002) Discussion to Spiegelhalter et al. *J. Roy. Stat. Soc. B*,64,616-618
- Burnham K, Anderson, D (2002) *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach*. New York: Springer-Verlag, 2nd

Edition

Carlin, B, Chib, S (1995) Bayesian model choice via the Markov Chain Monte Carlo methods, *J. Roy. Stat. Soc. B*, 57, 473-484

Chen, M (2005) Bayesian computation: from posterior densities to bayes factors, marginal likelihoods, and posterior model probabilities. In “Bayesian Thinking, Modeling and Computation”, eds D Dey & C Rao. Elsevier

Chib, S (1995) Marginal likelihood from the Gibbs output. *J. Amer. Stat. Assn.*, 90, 1313-1321

Congdon, P (2004) A multivariate model for spatio-temporal health outcomes with an application to suicide mortality, *Geog. Ana.*, 36, 234-258

Congdon, P (2006) Bayesian model choice based on Monte Carlo estimates of posterior model probabilities, *Comp. Stat.& Data Ana.*, 50, 346-357

Dellaportas, P, Forster, J, Ntzoufras, I (2002) On Bayesian model and variable selection using MCMC. *Statistics and Computing*, 12, 27–36.

Fitzmaurice G, Laird N, Ware J (2004) *Applied Longitudinal Analysis*. New York: John Wiley

Friel, N, Pettitt, A (2006) Marginal likelihood estimation via power posteriors. Technical Report, University of Glasgow (<http://www.statslab.cam.ac.uk/~mcmc/>)

Fruhwirth-Schnatter, S (2004) Estimating marginal likelihoods for mixture and Markov switching models using bridge-sampling techniques. *Econometrics J*, 7, 143-167

Gelfand, A, Dey, D (1994) Bayesian model choice: asymptotics and exact calculations. *J. Roy. Stat. Soc. B*, 56, 501-514.

- Gelman, A, Carlin, J, Stern, H, Rubin, D (1995) *Bayesian Data Analysis*. London: Chapman and Hall.
- George, E, McCulloch, R (1997) Approaches for Bayesian variable selection. *Statistica Sinica* 7, 339-374
- Geweke, J (1989) Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57, 1317-1340
- Godsill, S (2001) On the relationship between Markov chain Monte Carlo for model uncertainty. *J. Comp. Graph. Stat.*, 10, 230-248
- Green, P (2003) Trans-dimensional Markov chain Monte Carlo. In: Green, P, Hjort, N, Richardson, S (Eds.), *Highly Structured Stochastic Systems*, Oxford University Press, Oxford, pp 179-198
- Green, P, O'Hagan, A (1998) Model choice with MCMC on product spaces without using pseudo priors. Technical Report, Department of Statistics, University of Nottingham
- Henderson, R, Oman, P (1999) Effect of frailty on marginal regression estimates in survival analysis. *J. Roy Stat Soc B*, 61, 367-379
- Higgins J, Spiegelhalter D (2002) Being sceptical about meta-analyses: a Bayesian perspective on magnesium trials in myocardial infarction. *Int. J. Epid.*, 31, 96-104
- Hoeting, J, Madigan, D, Raftery, A, Volinsky, C (1999) Bayesian model averaging: a tutorial. *Stat. Sci.*, 14, 382-401
- Kadane, J, Lazar, N (2004) Methods and criteria for model selection. *J. Amer. Stat. Assn.*, 99, 279-290

- Kahn, M, Raftery, A (1996) Discharge rates of Medicare stroke patients to skilled nursing facilities: Bayesian logistic regression with unobserved heterogeneity. *J. Amer. Stat. Assn.*, 91, 29-41.
- Knorr-Held, L, Rainer, E (2001) Projections of lung cancer mortality in West Germany: a case study in Bayesian prediction. *Biostatistics*, 2, 109-129
- Lewis, S, Raftery, A (1997) Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *J. Amer. Stat. Assn.*, 92, 648-655
- Meyer, M, Laud, P (2002) Predictive variable selection in generalized linear models. *Journal of the American Statistical Association*, 97, 859-871
- O'Hagan, A.(2003) HSSS Model Criticism. In: Green, P, Hjort, N, Richardson, S (Eds.), *Highly Structured Stochastic Systems*, Oxford University Press, Oxford, pp 179-198
- Pourahmadi, M, Daniels (2002) Dynamic conditionally linear mixed models for longitudinal data. *Biometrics*, 58, 225-231
- Raftery, A (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*, 83, 251-266
- Rossi, P, Allenby, G, McCulloch, R (2005) *Bayesian Statistics and Marketing*, 1st edition. Wiley: New York
- Selvin, S (1998), *Modern Applied Biostatistical Methods Using S-Plus*, 1st Edition, Oxford University Press
- Sinharay, S, Stern, H (2005) An empirical comparison of methods for computing Bayes factors in generalized linear mixed models, *J. Comp. Graph. Statist.*, 14, 415-435

- Song, X, Lee, S (2002). Bayesian model selection method with applications. *Comput. Stat.& Data Ana.*, 40, 539-557
- Spiegelhalter, D, Best, N, Carlin, B, van der Linde, A (2002) Bayesian measures of model complexity and fit. *J. Roy. Statist. Soc.* 64B, 583-639.
- Wasserman, L (2000) Bayesian model selection and model averaging. *J. Math. Psych.*, 44, 92-107
- Whitley E, Gunnell D, Dorling D, Smith G (1999) Ecological study of social fragmentation, poverty, and suicide. *Br Med J*, 319, 1034-1037
- Yuan, C, Druzdzel, M (2005) How heavy should the tails be? In *Proceedings of the Eighteenth International FLAIRS Conference (FLAIRS-05)* Russell, I, Markov, Z (eds), AAAI Press/The MIT Press, Menlo Park , CA,799-804.

Figure 1 Kernel Plot of Model Averaged Probability

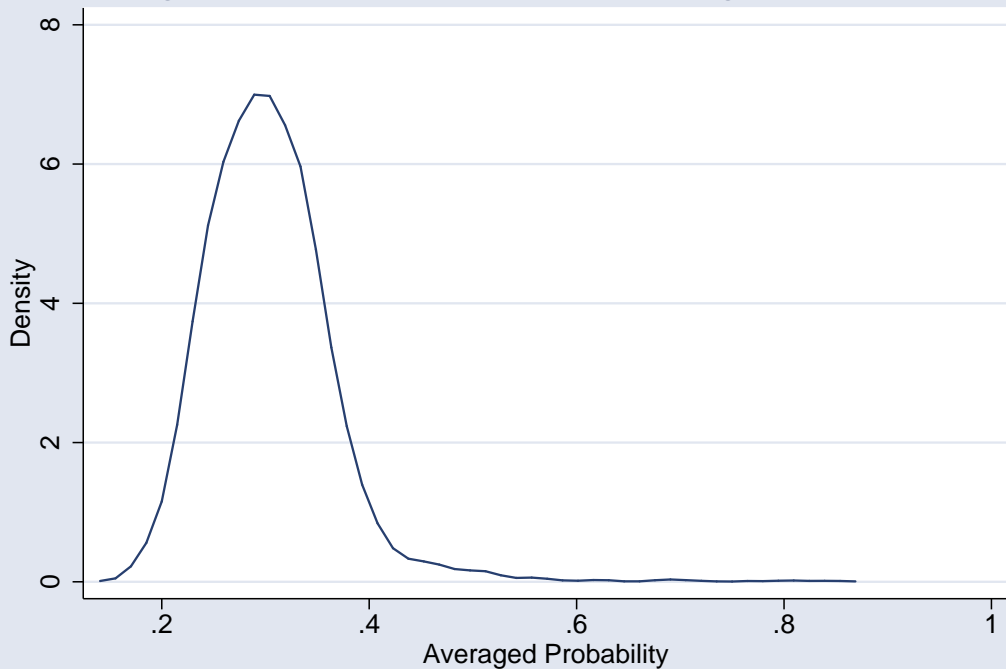


Figure 2a Change in beta1 between Models

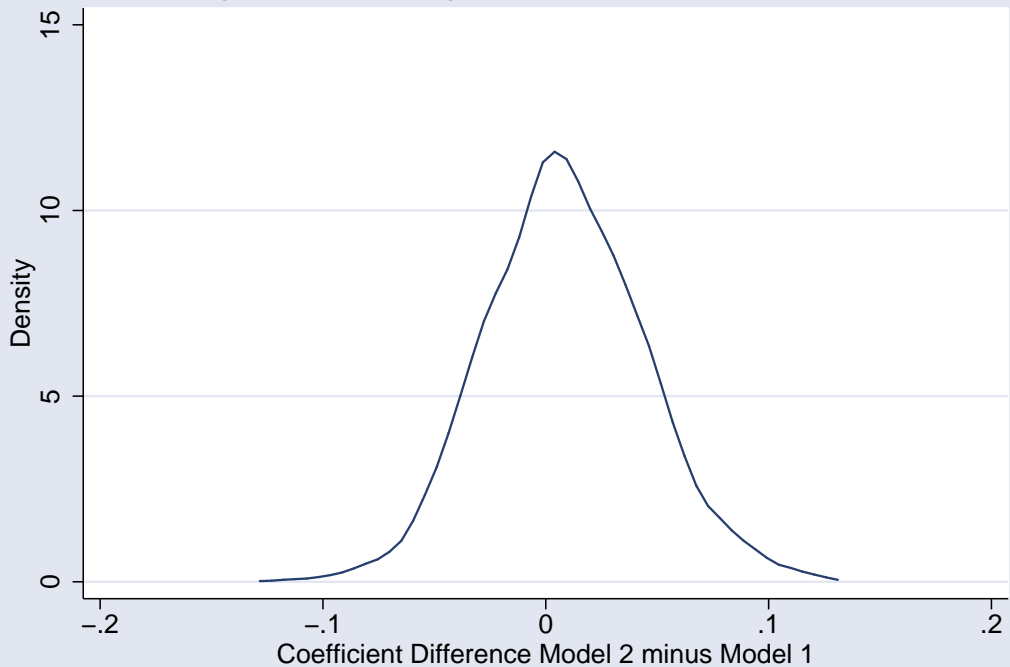


Figure 2b Change in beta2 between Models

