A novel proposal construction for reversible jump MCMC

Michail Papathomas*, Petros Dellaportas† and Vassilis G.S. Vasdekis†

June 2, 2010

Abstract

We propose a novel methodology to construct proposal densities in reversible jump MCMC algorithms so that consistent mappings across competing models are achieved. Unlike nearly all previous approaches our proposals are not restricted to operate to moves between local models, but they are applicable even to models that do not share any common parameters. We focus on linear regression models and produce concrete guidelines on proposal choices for moves between any models. These guidelines can be immediately applied to any regression models after applying some standard data transformations to near-normality. We illustrate our methodology by providing concrete guidelines for model determination problems in logistic regression and log-linear graphical models. Two real data analyses illustrate how our suggested proposal densities together with the resulting freedom to pro-

*Department of Epidemiology and Public Health, Imperial College London, UK
†Department of Statistics, Athens University of Economics and Business, Greece

pose moves between any models improve the mixing of the reversible jump Metropolis algorithm.

Keywords — Bayesian inference; Graphical models; Linear regression; Log-linear models; Logistic regression

1 Introduction

The reversible jump MCMC algorithm was introduced by Green (1995) as an extension to the standard Metropolis-Hastings algorithm to variable dimension spaces; see also Tierney (1998). It is based on creating a Markov chain which can 'jump' between models with parameter spaces of different dimension. In a Bayesian inference framework, its great impact stems from the fact that it allows the calculation of posterior model probabilities for a large number of competing models. Here the key issue is not the calculation of marginal densities per se, but the ability to search, via a Markov chain simulation, in a large space of models in which marginal densities are not available. Although reversible jump has been extensively used in many applied model determination problems, its widespread applicability has been hindered by the difficulty to achieve proposal moves between models that employ some notion of inter-model consistency that facilitates good mixing across models. We provide a methodology that constructs moves between any models in the model space in a general regression setting, and we illustrate its applicability in logistic regression and log-linear graphical models.

The general reversible jump algorithm can be described as follows. Assume that a data vector y is generated by model $i \in \mathcal{M}$, where $\mathcal{M}$ is a set of competing models. Each model specifies a likelihood $f(y|\theta_i, i)$, subject to an unknown parameter vector $\theta_i \in \Theta_i$ of size p_i, where $\Theta_i \subseteq R^{p_i}$ is the parameter space for model i. Let (θ_i, i) be the current state of the Markov chain. Then, the reversible jump algorithm consists of the following steps:

1. Propose a new model j with probability $\pi(i, j)$.

2. Generate u from a proposal density $q(u|\theta_i, i, j, y)$.

3. Set $(\theta_j, u^*) = g_{i,j}(\theta_i, u)$, where the deterministic transformation function $g_{i,j}$ and its inverse are differentiable. Note that $p_j + \dim(u^*) = p_i + \dim(u)$ and that $g_{i,j} = g_{j,i}^{-1}$.

4. Accept the proposed move from model i to model j with a probability $\alpha_{i,j} = \min(1, A)$,

$$A = \frac{f(y|\theta_j, j) f(\theta_j|j) f(j) \pi(j, i) q(u^*|\theta_j, j, i, y)}{f(y|\theta_i, i) f(\theta_i|i) f(i) \pi(i, j) q(u|\theta_i, i, j, y)} \times \left| \frac{\partial(\theta_j, u^*)}{\partial(\theta_i, u)} \right|$$

where $f(i)$ and $f(\theta_i|i)$ denote prior densities for model i and parameter vector θ_i respectively.

Step 1 of the algorithm seems to create a freedom of choice, but unfortunately proposed models should be carefully chosen such that θ_j in step 3 belongs to a relatively high region of the posterior density $f(\theta_j|j, y) \propto$

$f(y|\theta_j, j) \times f(\theta_j|j)$ to increase the probability of acceptance. This, in turn, implies that the functions q and g in steps 2 and 3 are key elements of the successful application of the algorithm.

Brooks *et. al.* (2003) have reviewed and suggested various ways to choose the parameters of q efficiently. However, the requirement for the existence of some consistency in the mapping between models has limited all these methods to operate with 'local' moves in the model space. This means that θ_i and θ_j often have many common elements, and in fact in most cases the one is a subset of the other, resulting in attempted moves between nested models. Richardson and Green (1997) introduce a technique where the desired compatibility between models is retained through moment matching. Ehlers and Brooks (2008) construct moves between non-nested autoregressive models and choose the parameters of the proposal densities by approximating relevant posterior conditional distributions, setting the first and higher order derivatives of the acceptance ratio with respect to u equal to zero. Fan *at al.* (2009) construct proposals with the use of a marginal density estimator. They provide guidelines for the implementation of their method on normal mixtures and autoregressive models, allowing for local moves in the model space; see also Vermaak *at al.* (2004).

An intuitive description of our proposed methodology is based on the following two points. First, when model jumps are proposed, it is desirable that beliefs about the data under the current model (with beliefs described

by the likelihood) should be the same to what we expect to believe if we accept the proposed model. Second, these proposals should be general enough to allow moves even when θ_i and θ_j do not have any common elements.

After specifying the mathematical formulation that satisfies the two key points above, we assume that q is a multivariate normal density and we derive exact solutions for its mean vector and covariance matrix in the case of linear regression models. We then investigate the applicability of our method to some binomial and contingency table data in which the data are transformed to approximate normality. Although this approximation might not be accurate, the derived proposal densities are still appealing and in fact provide an impressive improvement over the currently available reversible jump proposed algorithm of Dellaportas and Forster (1999).

Many currently available ways to choose q and g are described in great detail in the paper by Brooks *et. al.* (2003) and the accompanying discussion. See also Sisson (2005), Ehlers and Brooks (2008) and Fan *at al.* (2009). As pointed out earlier, the majority of them refers to 'local' moves in $\mathcal{M}$. An interesting different approach, that is a 'global' method and in very close line with our suggested proposal densities, is given by Green (2003), who develops a method for constructing proposal distributions that is similar in spirit to the random walk Metropolis sampler of Roberts (2003). He considers normal proposal densities and suggests that their mean and variances should be functions of the mean and variances of the target density, which

can be estimated with a pilot run. This requirement reduces the appeal of this method when the number of models is large. Hastie (2004) extends this approach, considering a proposal that is a mixture of normal distributions and adaptive sampler for the specification of relevant parameters.

In an unpublished report, Green (2000) produces similar results with those presented here. In fact, our empirical findings show that, in some instances, the resulting reversible jump efficiency between the two approaches is comparable. Therefore, one can view our work as an effort to provide theoretical justifications to the approach of Green (2000). In this report, two competing general linear models, M_X and M_Z are considered with independent homoscedastic normal errors and known variance. This stylised situation allows to visualise the structure of the problem and utilize the residuals of the two models to construct efficient proposals. The author uses the random u vector to perturb the starting point orthogonally away from the hyperspace defined by the M_X design matrix, before it is projected onto the M_Z hyperspace. This leads to a normal proposal density with mean $\mu = (X_j'X_j)^{-1}X_j'y + (X_j'X_j)^{-1}X_j'(X_i\theta_i - P_iy)$. Green suggested the variance of this proposal to be, $\Sigma = (X_j'X_j)^{-1}X_j'\left(I_n - P_i\right)X_j(X_j'X_j)^{-1}$. The rank of this covariance matrix is $p_j - t$. We implement these proposals, along with the ones we derive in this manuscript, in the real data illustrations presented in Section 3.

The rest of the paper proceeds as follows. Section 2 gives the mathe-

matical derivation of our proposed methodology. In Section 3, we search the space of graphical models for a large contingency table using reversible jump MCMC. Also, we search through a series of logistic regression models for a data set that contains binomial observations. In Section 4 we conclude with some discussion.

2 The proposed approach

We consider an n-dimensional vector y of normal observations and competing linear models $N(\eta_i, V_i)$, $i \in \mathcal{M}$, where $\eta_i = X_i\theta_i$, X_i is the design matrix of model i and θ_i is of dimension p_i. Assume that the reversible jump MCMC algorithm has a current state (θ_i, V_i, i) and that a move is proposed to (θ_j, V_j, j). The variances V_i and V_j are considered known and are not the subject of inference. Then, our key idea is that the proposal density for θ_j, $q(u|\theta_i, i, j, y)$, should satisfy the relationship

$$f(y|\theta_i, V_i, i) = c_{i,j} E_u \left\{ f(y|u, V_j, j) \right\} \tag{1}$$

where $f(y|\theta_i, V_i, i)$ denotes the likelihood of the data under the current state of the chain, and $f(y|u, V_j, j)$ denotes the likelihood of the data under model j with parameters u. E_u denotes expectation with respect to the proposal density $q(u|\theta_i, i, j, y)$. Note that we denote with u the argument of the proposal density to simplify the notation, although u is just the random component in the construction of θ_j.

Equation (1) expresses the desire to propose θ_j that should, on average with respect to the proposal density, obtain $f(y|\theta_i, V_i, i) = f(y|\theta_j, V_j, j)$. In Section 4, we discuss in detail the simple coherence argument this requirement is based on. The constant $c_{i,j}$ is introduced to take into account the possibility that two competing models may be inherently different, so that obtaining similar likelihoods is not possible. The role of $c_{i,j}$ is to weight down or up the likelihood of the destination model so that equation (1) can always be realized. We set that $c_{i,j}$ takes the form $f(y|\hat{\theta}_i, V_i, i)/f(y|\hat{\theta}_j, V_j, j)$, where $\hat{\theta}_i$ and $\hat{\theta}_j$ are the maximum likelihood estimates of the model parameters.

We attack (1) by assuming that $q(u|\theta_i, i, j, y)$ is a Normal density $N(\mu, \Sigma)$. There are clearly many values of μ and Σ that satisfy (1), and consequently many proposal densities $q(u|\theta_i, i, j, y)$ with that property. This fact is taken care of in our theoretical development below. When these solutions are available, they provide a yardstick to construct proposal densities for other linear regression models with non-normal responses; we provide such examples in Section 3.

Our approach has a similarity with the centering functions approach suggested by Brooks *et al.* (2003), but the two methods are inherently different. The centering functions approach imposes exact equality between the likelihood functions of models i and j so that a deterministic mapping can be constructed. The function $g_{i,j}$ is predetermined, defined for the case where moves are attempted between nested models and common parameters

are kept fixed. In contrast, we aim to explore (1) and construct proposals for complex moves between models that do not necessarily share parameters, with proposed values that change adaptively in accordance with the current state of the chain and no parameters are kept fixed. The following theorem provides the required solution to (1):

Theorem 1: *Under the model determination setup defined above, one solution for the mean μ of the proposal distribution $N(\mu, \Sigma)$ is given by*

$$\mu = (X_j' V_j^{-1} X_j)^{-1} X_j' V_j^{-1} \left\{ y + B^{-1} V_i^{-1/2} (X_i \theta_i - P_i y) \right\}. \qquad (2)$$

where $B = (V_j + X_j \Sigma X_j')^{-1/2}$ and $P_i = X_i (X_i' V_i^{-1} X_i)^{-1} X_i' V_i^{-1}$ is the projection matrix to the space generated by the columns of X_i, weighted by V_i^{-1}. The proof of Theorem 1 is given in the Appendix.

This result has an interesting interpretation. The mean of the proposal density is the maximum likelihood estimate of the new model plus a correction term that depends upon the difference between the fitted values under the maximum likelihood estimate for model i, $P_i y$, and the fitted values under the currently accepted θ_i. Intuitively, the difference $X_i \theta_i - P_i y$ determines a distance between the current value θ_i from the mode of its posterior density, so the proposed value of θ_j lies, in expectation, in a relatively equally high posterior region in model j. Note that, throughout the paper, we assume that prior densities of parameters within each model are non-informative in the sense that they are constant in the important region

of the likelihood function.

We now turn to the determination of Σ. Note that Σ appears in Theorem 1 through matrix B in such a way that any choice for Σ would make B invertible. However, it should be recognized that when jumping from model i to model j some elements of θ_i and θ_j may be common to both models, so it would be desirable to propose a move with reduced variability to these elements. Assume that the last t parameters in θ_j are common to both models. There are at least two possible choices for the form of Σ. Setting $Q_{ij} = (X_i' V_i^{-1/2} V_j^{-1/2} X_j)$, the first choice involves the matrix Q_{jj}^{-1} which is the covariance matrix associated with $f(\theta_j|j, V_j, y)$. Σ can be formed by that part of rows and columns of Q_{jj}^{-1} which correspond to the $p_j - t$ uncommon parameters between models i and j, whilst all other elements of Q_{jj}^{-1} are replaced by zero. The second choice involves the matrix $Q_{jj}^{-1} - Q_{jj}^{-1} Q_{ji} Q_{ii}^{-1} Q_{ij} Q_{jj}^{-1}$, of which a simplified version was proposed by Green (2000) in the unpublished report discussed in Section 1. This suggestion has two advantages; the first is that it is smaller than Q_{jj}^{-1} in the Löwner sense (Harville, 1997) providing small variances for our proposals. The second is that the rank of this matrix is $p_j - t$ and this matches the idea of using the already gathered information about the t common parameters. Therefore, we suggest that a reasonable choice for Σ is

$$\Sigma = Q_{jj}^{-1} - Q_{jj}^{-1} Q_{ji} Q_{ii}^{-1} Q_{ij} Q_{jj}^{-1} + c I_{p_j} \tag{3}$$

with any scalar $c > 0$ which makes Σ invertible. Thus, the proposed θ_j

is constructed as $\theta_j = (X_j' V_j^{-1} X_j)^{-1} X_j' V_i^{-1} \left\{ y + B^{-1} V_i^{-1/2} (X_i \theta_i - P_i y) \right\} + \Sigma^{1/2} u$ where $u \sim N(0, I_{p_j})$.

The constant c is clearly a tuning parameter that determines the variability of the proposals for the common parameters of models i and j. If $c > 0$, then $\dim(u) = p_j$ and $\dim(u^*) = p_i$, even if some of the parameters of the two models are common. In all analyses we have performed, mixing performance was very robust to small values of c, but of course some tuning is required as it is usually the case in reversible jump algorithms. For log-linear models, where scaling the design matrix has smaller, comparatively to linear models, effect on the parameter values, we have found that any value of c below 10^{-3} gives very similar results. On the other extreme, the case $c = 0$ acts as in the usual nested models case in which there is zero variability in the proposals of common parameters, although in our case, values of common parameters also change when moving to a different model. Then, $\dim(u) = p_j - t$ and $\dim(u^*) = p_i - t$.

Finally, one last issue in implementing the reversible jump MCMC algorithm is the derivation of the jacobian term in step 4. Since our proposals for θ_j are constructed using both θ_i and u, the jacobian of the transformation seems to have a complex form. But, interestingly, after some algebra, this Jacobian (for $c > 0$) simplifies to,

$$\left| \frac{\partial(\theta_j, u^*)}{\partial(\theta_i, u)} \right| = \left| \Sigma_{i,j}^{1/2} \right| \left| \Sigma_{j,i}^{-1/2} \right|,$$

where $\Sigma_{i,j}$ denotes the proposal density covariance matrix Σ when a move

is attempted from model i to model j.

3 Implementation of proposals

We provide two real data examples in log-linear graphical and logistic regression models. The key idea is that we can exploit the results of Section 2 by first applying a data transformation to responses so that they approximate normality; see, for example, Clyde (1999). We chose these examples to point out that our results are immediately applicable to any regression models since this approximation is not needed to be adequately precise: we just use it to approximate some good proposal densities for the MCMC algorithm. All acceptance probabilities are based on the original data and models. Any departure from (1) would not affect the correctness of our sampler, whilst small departures should still provide good proposal densities.

3.1 Graphical log-linear model determination

Edwards and Havranek (1985) presented a 2^6 table in which 1841 men were cross-classified by six risk factors for coronary heart disease. The factors were smoking (A), strenuous mental work (B), strenuous physical work (C), systolic blood pressure (D), ratio of lipoproteins (E) and history of coronary heart disease in family (F). We assume main effects are always present, and consider for comparison the 32768 possible graphical log-linear models. Equal prior model probabilities were assigned to the models, and the vague

priors suggested in Dellaportas and Forster (1999) were used. As pointed out by these authors, for such a large number of competing models a simple search algorithm consisting of equally likely moves to any of the models is not appropriate, so a model search algorithm that moves locally in model space is necessary for exploring efficiently the model space. Dellaportas and Forster suggested a reversible jump algorithm adopting the common approach of allowing moves that involve only the addition or deletion of an edge; indeed this is used extensively in the literature, see Jones *et al.* (2005). Since jumps between models take place only between nested models, Dellaportas and Forster suggested adopting proposal densities for the uncommon parameters obtained through an initial pilot run on the saturated model.

Compared to the existing algorithm of Dellaportas and Forster, our contribution consists of two parts. First, the initial pilot run that is needed to obtain proposal densities is not needed. Second, since we have no restrictions for moves to only nested models we propose less 'local' moves that allow much better mixing in the large model space. These two advantages become more important as the dimension of the model space increases.

Let w_k be a cell count, distributed as a Poisson random variable and $\eta_{ik} = (X_i\theta_i)_k$ be the linear predictor so that for model i, $E(w_k|M = i) = \exp\{\eta_{ik}\}$, $k = 1, \ldots, n$. To apply Theorem 1 in log-linear and graphical models we can proceed by first transforming the data to near-normality as follows. For adequately large w_k (say greater than five), a Poisson variable

with mean $\exp\{\eta_{ik}\}$ is approximated by a normal density with mean and variance equal to $\exp\{\eta_{ik}\}$. Thus, by applying the Delta method we obtain that for model i, $\sqrt{w_k}$ is approximately distributed as a $N(\exp\{0.5\eta_{ik}\}, 0.25)$ random variable. A further Taylor expansion to $\exp\{0.5\eta_{ik}\}$ around $\log \bar{w}$, where $\bar{w}$ is the observations sample mean, results to the approximation

$$E\left(\sqrt{w_k}\right) \simeq \sqrt{\bar{w}} + \frac{\sqrt{\bar{w}}}{2}\left(\eta_{ik} - \log \bar{w}\right).$$

which can be used to produce the modified random variable

$$y_k = \frac{2}{\sqrt{\bar{w}}}(\sqrt{w_k} - \sqrt{\bar{w}}) + \log(\bar{w})$$

which is approximately distributed as $N(\eta_{ik}, 1/\bar{w})$.

Assuming that all main effects are always present, we adopt a search approach that allows for three possible moves: the removal of an edge from the graph, the addition of an edge, or the replacement of one edge by another. We used (3) for Σ, and $c = 10^{-5}$. For this data set, the MCMC mixing is similar for any value of c within $(10^{-9}, 10^{-3})$. Positive values of c smaller than 10^{-9} are causing numerical instability, whereas the results are virtually identical when $c = 0$.

We obtained results derived from 3×10^6 iterations, after 8×10^5 burn-in iterations were discarded. Results regarding posterior model probabilities are identical to Dellaportas and Forster (1999) and are not reproduced here. The reversible jump MCMC accepted model moves with average probability 5.12%. Setting $c = 0$ gives, effectively, the same acceptance rate (5.15%) but

this choice implies a slightly more complex algorithm from a computational perspective, requiring two matrix decompositions, one for the variance matrix associated with the proposed jump and one for the variance matrix associated with the reverse move. With the proposals derived following Green (2000), the percentage of accepted moves is 5.1%, similar to the acceptance rate with the theoretically derived proposals. Notice that to implement proposals following Green (2000) two variance matrix decompositions are also required.

We compared the performance of the derived proposals with the reversible jump algorithm described by Dellaportas and Forster (1999). This procedure only allows moves between nested models, after an edge has been either removed or added. New values are proposed for the additional parameters when a jump to a larger model is attempted. Values of common parameters remain the same. The proposal distributions are multivariate normal densities fitted around the posterior moments of the saturated model parameters. The posterior moments are calculated with a pilot chain. The proportion of accepted moves with the Dellaportas and Forster algorithm was 2.4%. To evaluate the quality of our proposals, irrespectively of the type of moves that are allowed, we considered an identical sampler, and replaced the Dellaportas and Forster proposals with the ones derived in this manuscript. Under this set up, the acceptance rate is 7.1%, a considerable increase compared to 2.4% that is only attributed to better parameter

proposals.

Note that, interestingly, allowing only for nested moves favors a higher rather than a lower acceptance rate. For instance, with our proposals, when only nested moves were attempted the acceptance rate increased from 5.12% to 7.1%. (There is a similar increase with the Green (2000) proposals too.) This is because the sampler moves with smaller steps within the model space, accepting plenty of moves close to a local mode. One of the usual worries of practitioners that adopt the reversible jump MCMC algorithm is that the sampler may be trapped at a 'sticky patch' in the model space, which results to very low probability of leaving a model which has locally high probability but it is not the mode of the posterior density. The main advantage of our method is that it allows for moves between non-nested models, potentially increasing the mobility of the chain between local modes in the model space. To obtain an indication of how likely the MCMC chain is trapped in such a local mode, we ran the algorithm 200 times and recorded the number of iterations before the highest posterior density model is first visited. For consistent results we started all chains from the model that only contains main effects. Our strategy required an average of 447 iterations to reach the best model with a standard deviation of 455 iterations, whereas the Dellaportas and Forster (1999) algorithm required 8454 iterations with a standard deviation of 17645. Very similar results were obtained for $c = 0$ (483, s.d. 479) and the Green (2000) method (493, s.d. 554). It is clear

that allowing for non-nested moves between models increases significantly the mobility of the chain, allowing to move faster between local modes in the model space.

Finally, we checked whether the choice of Σ is important, and considered an alternative naive choice of $\Sigma = \bar{w}^{-1}(X_j'X_j)^{-1}$ or other choices such as $\Sigma = r\bar{w}^{-1}(X_j'X_j)^{-1}$ for various values $0 < r < 1$. In all cases the average acceptance rate never exceeded 1%, indicating that choosing Σ as in (3) offers a considerable comparative advantage.

3.2 Competing logistic regression models for binomial data

Our suggested data transformation for logistic regression model determination problems proceeds as follows. Let z_k, $k = 1, \ldots, n$, be the number of successes in a series of n binomial experiments with corresponding n_k trials and probabilities of success p_k. Define $w_k = z_k/n_k$ and let $\eta_{ik} = (X_i\theta_i)_k$ be the linear predictor so that for model $i \in \mathcal{M}$, $E(w_k|i) = \exp\{\eta_{ik}\}/(1+\exp\{\eta_{ik}\})$. Since w_k is approximately normal with mean p_k and variance equal to $p_k(1-p_k)/n_k$, application of the Delta method gives that, for model i, $\arcsin(\sqrt{w_k})$, is approximately normal with mean $\arcsin(\sqrt{p_k})$ and variance $0.25n_k^{-1}$. A further Taylor expansion to $\arcsin(\sqrt{p_k})$ around $\log(\bar{w}/(1-\bar{w}))$, where $\bar{w}$ denotes the sample mean of w_k, results to the approximation

$$E\left(\arcsin(\sqrt{w_k})\right) \simeq \arcsin(\sqrt{\bar{w}_k}) + \frac{\sqrt{\bar{w}(1-\bar{w})}}{2}\left(\eta_{ik} - \log(\bar{w}/(1-\bar{w}))\right).$$

Therefore, the modified random variable

$$y_k = 2(\bar{w}(1-\bar{w}))^{-1/2}(\arcsin(\sqrt{w_k}) - \arcsin(\sqrt{\bar{w}_k})) + \log(\bar{w}/(1-\bar{w}))$$

is approximately distributed as $N(\eta_{ik}, (n_k\bar{w}(1-\bar{w}))^{-1})$ and Theorem 1 can be applied.

We consider a well known data set analysed in Fowlkes *et al.* (1988). The response is the number of subjects satisfied with their employment. The four explanatory factors are Race (A, two levels), Age (B, two levels), Sex (C, two levels) and Region (D, seven levels). Assuming main effects are always present, we compare and assess the 64 models that contain all possible combinations of two-way interactions, from no interactions at all to all six two-way interactions. Equal prior model probabilities are assigned to the models and the unit information priors suggested in Ntzoufras *et al.* (2003) are adopted. Our proposals allow for direct moves between non-nested models and, therefore, we adopt a flexible model search strategy that consists of the following three possible moves: remove an interaction, add an interaction or replace one interaction term with another.

We obtained results derived from 3×10^5 iterations, after 5×10^4 burn-in iterations were discarded. With our proposed strategy, with Σ as in (3), 8.7% of the proposed model jumps are accepted. The algorithm was implemented for $c = 10^{-5}$ and turned out to be very robust for a series of values such that $c < 10^{-4}$. Results were virtually identical when $c = 0$, with an acceptance rate of 8.55%. With the proposals derived in Green

(2000), where the variance matrix is not weighed by V_i, the percentage of accepted moves was 0.5%, considerably smaller than the acceptance rate of our proposals. This illustrates that, for binomial data, it is important to allow for the variance of each y_k to be weighted by the number of trials n_k in cell k.

Finally, proposing from multivariate normal densities fitted around the posterior moments of the saturated model parameters along the lines of Dellaportas and Forster (1999), and allowing for only the addition or removal of an interaction, leads to an acceptance rate of 2.3%. An identical sampler that uses the proposals derived in this manuscript has an acceptance rate of 8.7%, about four times higher compared to Dellaportas and Forster.

4 Discussion

We have presented a novel idea for the construction of general proposal distributions for the reversible jump Markov chain Monte Carlo algorithms and we have provided guidelines to apply this method to general regression models, possibly after some data transformation of the responses. Our proposed strategy can be seen as an effort to attack the most interesting problem emanated by the work of Green (1995), namely the construction of efficient proposal densities for jumps between any (and not just 'local' or nested) models in model space.

The results in this manuscript are based on (1), which follows from the

simple coherence requirement that what we believe about the data under the current model (with beliefs about the data described by the likelihood) should be the same to what we expect to believe about the data if we accept the proposed model. The reasoning behind this requirement becomes clear in the case of discrete data. Assume that for two binary variables y_1 and y_2 we have accepted at the current state of the chain that $P(y_1 = 1, y_2 = 0|\theta_1, 1) = 0.7$. Then, before we accept any change in the state of the chain, it would be incoherent to expect that when we move to, say, state $(\theta_2, 2)$, the joint probability $P(y_1 = 1, y_2 = 0)$ will be different, say 0.9. This line of thought is loosely based on the fundamental requirement that pre-posterior expectations should be equal to prior expectations, viewing $P(y_1 = 1, y_2 = 0)$ as an expectation. In the reversible jump setting, we should not a priori expect that our beliefs for the data will change just because a different model will be adopted. The constant $c_{i,j}$ takes into account generic differences between two competing models, e.g. differences in dimensionality.

Although this recipe has only intuitive appeal and its optimality properties are hard to investigate, it provides a fail-safe strategy in the sense that such moves can be on the one hand very general and on the other will never be dramatically poor: to see this, notice that while the parameters are updated within each model they are sampled with high probability from the higher posterior regions, and therefore proposed moves to other models

will often propose parameters in the higher posterior regions too.

In general, in model comparison applications with a relatively large number of models, it is very rare that strong prior information is considered for the model parameters. In the rare event that, for some models, the prior information for their parameters is at odds with the likelihood, the mode of the posterior density of θ_i and/or θ_j may not be close to the maximum likelihood estimates $\hat{\theta}_i$ and $\hat{\theta}_j$. In that case, it is possible that a θ_i close to the mode of its posterior density will lead to a proposed θ_j that is close to the tail of its posterior density and vice versa.

Our strategy is applicable to any model determination problem in the generalized linear model specification. We have chosen to illustrate it with the most popular Poisson and binomial responses, but the transformation to normality is available to any densities from the exponential family. Thus, we feel that the applicability of our method is very broad and will spread the applications of reversible jump MCMC in many interesting model determination problems.

Acknowledgements

The project is co-funded by the European Social Fund and National Resources (Ministry of Education) Pythagoras II - EPEAEK, and by a MRC grant (G0600609). We would like to thank professor Peter Green for making known to us the unpublished draft and for useful comments related to our

work.

Appendix

First we prove the following Lemma.

Lemma: Consider two quadratic forms $(z-\mu)'\Sigma^{-1}(z-\mu)$ and $(y-Xz)'V^{-1}(y-Xz)$. Then,

$$(z-\mu)'\Sigma^{-1}(z-\mu)+(y-Xz)'V^{-1}(y-Xz)=(z-m)'A(z-m)+K$$

where,

$$A=\Sigma^{-1}+X'V^{-1}X, \qquad m=A^{-1}(\Sigma^{-1}\mu+X'V^{-1}y)$$

and

$$\begin{aligned} K= & \left(\mu-(X'V^{-1}X)^{-1}X'V^{-1}y\right)'\left(\Sigma+(X'V^{-1}X)^{-1}\right)^{-1}\left(\mu-(X'V^{-1}X)^{-1}X'V^{-1}y\right) \\ & +y'V^{-1}(I_n-P_X)y, \end{aligned}$$

where $P_X=X(X'V^{-1}X)^{-1}X'V^{-1}$ is the projection matrix to the space generated by the columns of X weighted by V^{-1}.

Proof of Lemma: After we complete the square using standard linear algebra we obtain that,

$$(z-\mu)'\Sigma^{-1}(z-\mu)+(y-Xz)'V^{-1}(y-Xz)=(z-m)'A(z-m)+K$$

where,

$$A=\Sigma^{-1}+X'V^{-1}X, \qquad m=A^{-1}(\Sigma^{-1}\mu+X'V^{-1}y)$$

and

$$K=-(\Sigma^{-1}\mu-X'V^{-1}y)'(\Sigma^{-1}+X'V^{-1}X)^{-1}(\Sigma^{-1}\mu-X'V^{-1}y)+\mu'\Sigma^{-1}\mu+y'V^{-1}y.$$

To simplify the expression for K we complete the square so that,

$$(\Sigma^{-1}\mu-X'V^{-1}y)'(\Sigma^{-1}+X'V^{-1}X)^{-1}(\Sigma^{-1}\mu-X'V^{-1}y)-\mu'\Sigma^{-1}\mu$$

$$= (\mu - m_1)' A_1 (\mu - m_1) + K_1,$$

where,

$$A_1 = \Sigma^{-1}(\Sigma^{-1} + X'V^{-1}X)^{-1}\Sigma^{-1} - \Sigma^{-1} = -[\Sigma + (X'V^{-1}X)^{-1}]^{-1}$$

$$m_1 = -A_1^{-1}\Sigma^{-1}(\Sigma^{-1} + X'V^{-1}X)^{-1}\Sigma^{-1}\Sigma X'V^{-1}y = (X'V^{-1}X)^{-1}X'V^{-1}y$$

and

$$K_1 = y'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}y.$$

Therefore,

$$(z-\mu)'\Sigma^{-1}(z-\mu) + (y - Xz)'V^{-1}(y - Xz)$$

$$= (z-m)'A(z-m) - (\mu - m_1)'A_1(\mu - m_1) - K_1 + y'V^{-1}y,$$

$$= (z-m)'A(z-m) + (\mu - (X'V^{-1}X)^{-1}X'V^{-1}y)'(\Sigma + (X'V^{-1}X)^{-1})^{-1}(\mu - (X'V^{-1}X)^{-1}X'V^{-1}y)$$

$$+ y'V^{-1}(I - X(X'V^{-1}X)^{-1}X'V^{-1})y.$$

Proof of Theorem 1:

$$E_u\{f(y|u, V_j, j)\} = \int \det(2\pi\Sigma)^{-1/2} \det(2\pi V_j)^{-1/2} \exp\left\{-\frac{1}{2}(u-\mu)'\Sigma^{-1}(u-\mu)\right.$$

$$\left. -\frac{1}{2}(X_j u - y)V_j^{-1}(X_j u - y)\right\} du$$

We apply the previous Lemma to the sum which is inside the exponential, replacing z with u. We obtain that this sum becomes,

$(u-m)'A(u-m) + y'V_j^{-1}(I_n - P_j)y$

$+(\mu - (X_j'V_j^{-1}X_j)^{-1}X_j'V_j^{-1}y)'\left(\Sigma + (X_j'V_j^{-1}X_j)^{-1}\right)^{-1}(\mu - (X_j'V_j^{-1}X_j)^{-1}X_j'V_j^{-1}y)$

where $P_j = X_j(X_j'V^{-1}X_j)^{-1}X_j'V^{-1}$, $A = \Sigma^{-1} + X_j'V_j^{-1}X_j$ and $m = A^{-1}(\Sigma^{-1}\mu + X_j'V_j^{-1}y)$. Only the first part of this sum depends on u, and the integral becomes,

$$\det(2\pi\Sigma)^{-1/2} \det(2\pi V_j)^{-1/2} \det(2\pi(\Sigma^{-1} + X_j V_j^{-1} X_j)^{-1})^{1/2} \times$$

$$\exp\left\{-\frac{1}{2}(\mu-(X_j'V_j^{-1}X_j)^{-1}X_j'V_j^{-1}y)'(\Sigma+(X_j'V_j^{-1}X_j)^{-1})^{-1}(\mu-(X_j'V_j^{-1}X_j)^{-1}X_j'V_j^{-1}y)\right\}$$
$$\times\exp\left\{-\frac{1}{2}y'V_j^{-1}(I_n-P_j)y\right\}$$

The product of the first and the third determinant can be simplified using (1.5) of Harville (1997, p.417) to $\det\left\{V_j^{-1}\left(V_j+X_j\Sigma X_j'\right)\right\}^{-1/2}$ By applying twice (2.2) of Harville (1997, p.424) we can also show that $\left(\Sigma+(X_j'V_j^{-1}X_j)^{-1}\right)^{-1}=X_j'\left(V_j+X_j\Sigma X_j'\right)^{-1}X_j$. Consequently, we obtain that,

$$E_u\left\{f(y|u,V_j,j)\right\}=\det(2\pi V_j)^{-1/2}\det\left\{V_j^{-1}\left(V_j+X_j\Sigma X_j'\right)\right\}^{-1/2}$$
$$\exp\left\{-\frac{1}{2}(X_j\mu-P_jy)'\left(V_j+X_j\Sigma X_j'\right)^{-1}(X_j\mu-P_jy)\right\}\times\exp\left\{-\frac{1}{2}y'V_j^{-1}(I_n-P_j)y\right\}$$

Note that for the MLE $\hat{\theta}_i=(X_i'V_i^{-1}X_i)^{-1}X_i'V_i^{-1}y$ the likelihood of the data becomes

$$f(y|\hat{\theta}_i,V_i,i)=\det(2\pi V_i)^{-1/2}\exp\left\{-\frac{1}{2}y'V_i^{-1}(I_n-P_i)y\right\}$$

and the ratio $f(y|\theta_i,V_i,i)/f(y|\hat{\theta}_i,V_i,i)$ reduces to

$$\exp\left\{-\frac{1}{2}(X_i\theta_i-y)'V_i^{-1}P_i(X_i\theta_i-y)\right\}=\exp\left\{-\frac{1}{2}(X_i\theta_i-P_iy)'V_i^{-1}(X_i\theta_i-P_iy)\right\}$$

Now, for $c_{i,j}=f(y|\hat{\theta}_i,V_i,i)/f(y|\hat{\theta}_j,V_j,j)$ condition (1) becomes,

$$\frac{f(y|\theta_i,V_i,i)}{f(y|\hat{\theta}_i,V_i,i)}=\frac{E_u\left\{f(y|u,V_j,j)\right\}}{f(y|\hat{\theta}_j,V_j,j)}$$

or

$$\exp\left\{-\frac{1}{2}(X_i\theta_i-P_iy)'V_i^{-1}(X_i\theta_i-P_iy)\right\}=\det(2\pi V_j)^{-1/2}\det\left\{V_j^{-1}\left(V_j+X_j\Sigma X_j'\right)\right\}^{-1/2}$$
$$\exp\left\{-\frac{1}{2}(X_j\mu-P_jy)'\left(V_j+X_j\Sigma X_j'\right)^{-1}(X_j\mu-P_jy)\right\}\times\exp\left\{-\frac{1}{2}y'V_j^{-1}(I_n-P_j)y\right\}$$
$$\det(2\pi V_j)^{1/2}\exp\left\{\frac{1}{2}y'V_i^{-1}(I_n-P_i)y\right\}$$

or

$$\exp\left\{-\frac{1}{2}(X_i\theta_i-P_iy)'V_i^{-1}(X_i\theta_i-P_iy)\right\}$$

$$= \kappa^{-1/2} \exp\left\{-\frac{1}{2}(X_j\mu - P_j y)' \left(V_j + X_j \Sigma X_j'\right)^{-1} (X_j\mu - P_j y)\right\}$$

where $\kappa = \det\left\{V_j^{-1}\left(V_j + X_j\Sigma X_j'\right)\right\} > 1$ (Rao and Toutenburg, 1995, p.299). By taking logarithms, the last equation becomes,

$$(X_i\theta_i - P_i y)' V_i^{-1} (X_i\theta_i - P_i y) = \log\kappa + (X_j\mu - P_j y)' B^2 (X_j\mu - P_j y)$$

where $B^2 = (V_j + X_j\Sigma X_j')^{-1}$. Setting $\tau = (\log\kappa)^{1/2}(\alpha' B^2 \alpha)^{-1/2}$, where $\alpha = (I_n - P_j)v$ and v any n-dimensional vector, we can see that the previous equation can be written as,

$$(X_i\theta_i - P_i y)' V_i^{-1} (X_i\theta_i - P_i y) = (X_j\mu - P_j y - \tau\alpha)' B^2 (X_j\mu - P_j y - \tau\alpha).$$

Therefore, a μ which satisfies the above equality is given by

$$B(X_j\mu - P_j y - \tau\alpha) = V_i^{-1/2}(X_i\theta_i - P_i y)$$

Solving with respect to μ, we obtain

$$\mu = (X_j' V_j^{-1} X_j)^{-1} X_j' V_j^{-1} y + (X_j' V_j^{-1} X_j)^{-1} X_j' V_j^{-1} B^{-1} V_i^{-1/2}(X_i\theta_i - P_i y)$$

which proves Theorem 1.

References

Brooks, S.P., Giudici, P. and Roberts, G.O. (2003): Efficient construction of Markov chain Monte Carlo proposal distributions. *J. R. Statist. Soc. B*, **65**, 3-55.

Clyde, M.A. (1999): Bayesian model averaging and model search strategies. *Bayesian Statistics 6*, Oxford University Press

Dellaportas, P. and Forster, J.J. (1999): Markov chain Monte Carlo model determination for hierarhical and graphical log-linear models. *Biometrika*, **86**, 615-633.

Edwards, D. and Havranek, T. (1985): A fast procedure for model search in multidimensional contingency tables. *Biometrika*, **72**, 339-351.

Ehlers, R.S. and Brooks, S.P. (2008): Adaptive proposal construction for Reversible jump MCMC. *Scandinavian Journal of Statistics*, **35**, 677-690.

Fan, Y., Peters, G.W. and Sisson, S.A. (2009): Automating and evaluating reversible jump MCMC proposal distributions. *Stat. Comput.*, **19**, 409-421.

Fowlkes, E.B., Freeny, A.E. and Landwehr, J.M. (1988): Evaluating logistic models for large contingency tables. *J. Amer. Stat. Assoc.*, **83**, 611-622.

Green, P.J. (1995): Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, **82**, 711-732.

Green, P.J. (2000): Trans-dimensional Markov chain Monte Carlo. *unpub-*

lished draft

Green P.J. (2003): Trans-dimensional Markov chain Monte Carlo. In *Highly structured stochastic systems.* Green, P.J., Hjort, N.L. and Richardson, S. (Eds) Oxford University Press.

Harville, D.A. (1997): *Matrix algebra from a statistician's perspective.* Springer-Verlag, Berlin

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carte, C. and West, M. (2005): Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, **20**, 388-400.

Ntzoufras, I., Dellaportas, P. and Forster, J.J. (2003): Bayesian variable and link determination for generalized linear models. *Journal of Statistical Planning and Inference*, **111**, 165-180.

Rao, C.R. and Toutenburg, H. (1995): *Linear models.* Springer, New York

Richardson, S. and Green, P.J. (1997): On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Stat. Soc. B*, **59**, 731-758.

Roberts, G.O. (2003): Linking theory and practice of MCMC. In *Highly structured stochastic systems.* Green, P.J., Hjort, N.L. and Richardson, S. (Eds) Oxford University Press.

Sisson, S.A. (2005): Trans-dimensional Markov chains: A decade of progress and future perspectives. *J. Amer. Stat. Assoc.*, **100**, 1077-1089.

Tierney, L. (1998): A note on Metropolis-Hastings kernels for general state

spaces. *Annals of Applied Probability*, **8**, 1-9.

Vermaak, J., Andrieu, C., Doucet, A. and Godsill, S.J. (2004): Reversible jump markov chain Monte Carlo strategies for Bayesian model selection in autoregressive processes. *J. Time S. Anal.*, **25**, 785-809.