Metropolis-Hastings Algorithm

(K Ch. 5, KPT Ch. 11)

Matlab scripts: `mod7_Poisson_data, mod7_Poisson_rwc_prep, mod7_Poisson_rwc, mod7_Poisson_rwc_chib, mod7_Poisson_ic, mod7_Poisson_rwc_ACplots, mod7_Poisson_ic_ACplots,mod7_Poisson_fishing_rwc_prep, mod7_Poisson_fishin_rwc, mod7_Poisson_fishing_ic, mod7_HP_mudsnail, mod7_HP-mudsnail_parallel, AR1data, AR1ic`

Matlab functions: `gs_Poisson_rwc, gs_Poisson_rwc_chib, gs_Poisson_ic, gs_HP, gs_HPpar, Poisson_beta_mle, gs_mudsnail, HP_beta_mle, HP_gi_mle, gs_AR1_ic, rho_mode`

Generally, whenever even the *conditional posterior kernel* for a parameter or set of parameters is unknown, we can no longer draw from it as would be the case in a "well-behaved" Gibbs Sampler. Instead, we need to approximate the unknown density via tools that have been specifically designed for this purpose, such as Importance Sampling , or, as is becoming increasingly popular given computational advances, a Metropolis-Hastings (MH) algorithm.

Let's assume the *unknown* conditional posterior for some parameter θ (scalar or vector) is generically given by $p(\theta \mid \mathbf{\Gamma}, \mathbf{y}) = \frac{p(\theta, \mathbf{y} \mid \mathbf{\Gamma})}{p(\mathbf{y} \mid \mathbf{\Gamma})} = \frac{p(\theta)\, p(\mathbf{y} \mid \theta, \mathbf{\Gamma})}{p(\mathbf{y} \mid \mathbf{\Gamma})}$, where $\mathbf{\Gamma}$ are other model parameters that enter the conditional posterior. However, we do know its kernel $\tilde{p}(\theta \mid \mathbf{\Gamma}, \mathbf{y}) = p(\theta)\, p(\mathbf{y} \mid \theta, \mathbf{\Gamma})$. Thus, the only unknown element is the normalizing constant $p(\mathbf{y} \mid \mathbf{\Gamma})$. The rationale of the MH algorithm is to use $\tilde{p}(\theta \mid \mathbf{\Gamma}, \mathbf{y})$, plus a *candidate-generating density* (CGD) or *proposal density* $q(\theta)$ to obtain draws from the unknown $p(\theta \mid \mathbf{\Gamma}, \mathbf{y})$. Note that $q(\theta)$ can also be a function of the data in addition to θ.

Suppose that the most recent draw of θ in the GS (which will initially be the starting draw) is θ^a. Now obtain a new "candidate draw" of θ, call it θ^b, from $q(\theta)$. Let's call this CGD generically $q(\theta)$. The new draw of θ is then accepted with probability

$$\alpha(\theta^a, \theta^b) = \min\left(\frac{p(\theta^b \mid \mathbf{\Gamma}, \mathbf{y}) q(\theta^a)}{p(\theta^a \mid \mathbf{\Gamma}, \mathbf{y}) q(\theta^b)}, 1\right) = \min\left(\frac{\tilde{p}(\theta^b \mid \mathbf{\Gamma}, \mathbf{y}) q(\theta^a)}{\tilde{p}(\theta^a \mid \mathbf{\Gamma}, \mathbf{y}) q(\theta^b)}, 1\right), \tag{1}$$

since the normalizing constant $p(\mathbf{y} \mid \mathbf{\Gamma})$ cancels out in the ratio. We usually work with logs:

$$\begin{aligned} &\log\left(\alpha(\theta^a, \theta^b)\right) = \min\left(\log \tilde{p}(\theta^b \mid \mathbf{\Gamma}, \mathbf{y}) + \log q(\theta^a) - \log \tilde{p}(\theta^a \mid \mathbf{\Gamma}, \mathbf{y}) - \log q(\theta^b), 0\right) = \\ &\min\left(\left(\log \tilde{p}(\theta^b \mid \mathbf{\Gamma}, \mathbf{y}) - \log q(\theta^b)\right) - \left(\log \tilde{p}(\theta^a \mid \mathbf{\Gamma}, \mathbf{y}) - \log q(\theta^a)\right), 0\right) \end{aligned} \tag{2}$$

In practice this is implemented by comparing the α value to a random uniform [0,1] draw (or, equivalently, $\log(\alpha)$ to the log of a uniform draw). If α exceeds the random value, the new draw θ^b is accepted, otherwise θ^a remains the most current draw. As discussed in Gelman et al (Ch. 12) , Koop (Ch. 5) and KPT, Ch. 11, after a sufficient number of "burn-ins" the sequence of draws of θ will converge to the desired underlying density.

If the proposal density is symmetric, i.e. $q(\theta^a) = q(\theta^b)$ the acceptance probability reduces to

$$\alpha(\theta^a, \theta^b) = \min\left(\frac{\tilde{p}(\theta^b \mid \boldsymbol{\Gamma}, \mathbf{y})}{\tilde{p}(\theta^a \mid \boldsymbol{\Gamma}, \mathbf{y})}, 1\right) \tag{3}$$

Also note that the generic GS is a special case of the MH with $q(\theta^j) = p(\theta^j \mid \boldsymbol{\Gamma}, \mathbf{y}), \quad j = a, b$ such that the acceptance probability is always one, i.e. every new draw of θ is accepted by default.

There are two basic types of MH algorithms – the *Random Walk Chain MH (RWMH)* and the *Independence Chain MH (IMH).* In the *RWMH* the proposal density is a function of the current draw of θ, i.e. takes the form of $q(\theta^b \mid \theta^a)$. For example, the current draw θ^a might be designated the mean of the proposal density. In this case, the acceptance probability takes the form of

$$\alpha(\theta^a, \theta^b) = \min\left(\frac{\tilde{p}(\theta^b \mid \boldsymbol{\Gamma}, \mathbf{y}) q(\theta^a \mid \theta^b)}{\tilde{p}(\theta^a \mid \boldsymbol{\Gamma}, \mathbf{y}) q(\theta^b \mid \theta^a)}, 1\right) \tag{4}$$

Usually, $q(\theta^b \mid \theta^a)$ is chosen to be symmetric to yield the simplified expression

$$\alpha(\theta^a, \theta^b) = \min\left(\frac{\tilde{p}(\theta^b \mid \boldsymbol{\Gamma}, \mathbf{y})}{\tilde{p}(\theta^a \mid \boldsymbol{\Gamma}, \mathbf{y})}, 1\right) \tag{5}$$

For the *IMH* the proposal density is not a direct function of the most recent draw of θ. A popular choice is a t-density or multivariate t-density with the mean set to the MLE solution (mode) of the posterior kernel $\tilde{p}(\theta \mid \boldsymbol{\Gamma}, \mathbf{y})$ and the variance set to the (possibly scaled) inverted negative Hessian coming out of the MLE sub-routine. Naturally, this is more complex and takes longer than the *RWMH*, but it tends to reduce auto-correlation compared to draws from the *RWMH*.

How can we influence acceptance rates? In the *IMH*, we can choose the scalar for the Inverted Hessian to boost or reduce the variance matrix in the t-density, and also the degrees of freedom for the t density. In the *RWMH*, we usually specify $q(\theta^b \mid \theta^a)$ to be a normal density or multivariate normal density with mean θ^a and an arbitrary variance. This variance is specified at the onset, along with the priors, then refined after a few "practice runs" to achieve optimal acceptance rates.

To assure that the MH algorithm covers the entire posterior kernel (with more weight given to areas with higher density), Gelman et al recommend an optimal acceptance rate of about 40-45% of all draws for a single-valued θ, and about 20-25% if θ is a vector. **However, these target thresholds only apply to the RWMH with a normal CGD**! For any other version of a MH algorithm IEF scores and convergence

diagnostics should be used for guidance on the optimal acceptance rate. For example, for the Poisson fishing application, boosting acceptance rates in the IMH from 35% to 75% doubled efficiency (= halved IEF scores), but in some other cases acceptance rates in the 30-50% may be sufficient.

An informal "proof" of the MH algorithm based on KPT, p. 153, is given at the end of this chapter.

Example 1: Parameterized Poisson Model

Consider a Poisson model with parameterized expectation. These models are useful in contexts such as visitation counts to recreation sites, patient visits to medical facilities, etc. The basic Poisson density has a single parameter, λ, which denotes both the expectation and variance of the distribution. Note that while a Poisson random variable is a non-negative integer, λ can be any positive real number.

The explanatory variables enter the scene as linear regressors for the logged expectation. Thus we have:

$$f\left(y_i \mid \lambda_i\right)=\frac{\exp\left(-\lambda_i\right)\lambda_i^{\,y_i}}{y_i!} \quad y_i=0,1,2\ldots, \quad \lambda_i>0 \qquad \text{where}$$
$$\lambda_i=\exp\left(\mathbf{x_i'\beta}\right) \tag{6}$$

The interpretation of the elements of $\boldsymbol{\beta}$ is thus a fractional change in expected outcome due to a unit change in the corresponding explanatory variable. Multiplying by 100 gives the percentage change.

The likelihood function is given as

$$p\left(\mathbf{y}\mid\boldsymbol{\beta}\right)=\prod_{i=1}^{N}\frac{\exp\left(-\lambda_i\right)\lambda_i^{\,y_i}}{y_i!}=\prod_{i=1}^{N}\frac{\exp\left(-\exp\left(x_i'\boldsymbol{\beta}\right)\right)\left(-\exp\left(x_i'\boldsymbol{\beta}\right)\right)^{y_i}}{y_i!} \tag{7}$$

In absence of any theoretical restrictions on marginal effects, we can choose the usual multivariate normal priors for $\boldsymbol{\beta}$:

$$p\left(\boldsymbol{\beta}\right)=\left(2\pi\right)^{-k/2}\left|\mathbf{V}_0\right|^{-1/2}\exp\left(-\tfrac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_0\right)'\mathbf{V}_0^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_0\right)\right) \tag{8}$$

where *k* is the dimension of $\boldsymbol{\beta}$.

The posterior kernel is thus given as

$$p\left(\boldsymbol{\beta}\mid\mathbf{y},\mathbf{X}\right)\propto\tilde{p}\left(\boldsymbol{\beta}\mid\mathbf{y},\mathbf{X}\right)=$$
$$\exp\left(-\tfrac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_0\right)'\mathbf{V}_0^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_0\right)\right)*\prod_{i=1}^{N}\frac{\exp\left(-\exp\left(\mathbf{x_i'\beta}\right)\right)\left(\exp\left(\mathbf{x_i'\beta}\right)\right)^{y_{ij}}}{y_i!} \tag{9}$$

This is not the kernel of a well-understood density, so we cannot obtain direct draws from it or derive analytical results.

Instead, we can apply a M-H procedure where we draw a candidate vector $\boldsymbol{\beta}_c$ from some known candidate-generating function $q(\boldsymbol{\beta}_c)$, and retain it with acceptance probability $\alpha(\boldsymbol{\beta}_o \to \boldsymbol{\beta}_c)$, where $\boldsymbol{\beta}_o$ is the "old" or "current" draw. In the most general sense, this CDG can be a function of $\boldsymbol{\beta}_o$, and / or the data, i.e. $q(\boldsymbol{\beta}_c) = q(\boldsymbol{\beta}_c \mid \boldsymbol{\beta}_o, \mathbf{y}, \mathbf{X})$. Different "flavors" of the M-H step evolve from different choices of $q(.)$.

Version 1: Random Walk Chain MH

For the RWMH we choose a multivariate normal CGD with mean equal to the current draw $\boldsymbol{\beta}_o$ (which, initially will be the starting draw), and variance matrix equal to cV_c, where c is some arbitrary scalar (initially set to something small, like 0.01) and V_c is initially set to an identity matrix.

Therefore, we have $q(\boldsymbol{\beta}_c \mid \boldsymbol{\beta}_o) = n(\boldsymbol{\beta}_o, c\mathbf{V}_c) = q(\boldsymbol{\beta}_o \mid \boldsymbol{\beta}_c)$ due to the symmetry property of the normal density.

The formula for the log (acceptance probability) thus reduces to

$$
\begin{aligned}
&\log\left(\alpha(\boldsymbol{\beta}_o \to \boldsymbol{\beta}_c)\right) = \min\left(\log\left(\frac{\tilde{p}(\boldsymbol{\beta}_c \mid \mathbf{y}, \mathbf{X}) q(\boldsymbol{\beta}_o \mid \boldsymbol{\beta}_c)}{\tilde{p}(\boldsymbol{\beta}_o \mid \mathbf{y}, \mathbf{X}) q(\boldsymbol{\beta}_c \mid \boldsymbol{\beta}_o)}\right), 0\right) = \min\left(\log\left(\frac{\tilde{p}(\boldsymbol{\beta}_c \mid \mathbf{y}, \mathbf{X})}{\tilde{p}(\boldsymbol{\beta}_o \mid \mathbf{y}, \mathbf{X})}\right), 0\right) = \\
&\min\left(\log \tilde{p}(\boldsymbol{\beta}_c \mid \mathbf{y}, \mathbf{X}) - \log \tilde{p}(\boldsymbol{\beta}_o \mid \mathbf{y}, \mathbf{X}), 0\right), \quad \text{where} \\
&\log \tilde{p}(\boldsymbol{\beta}_c \mid \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N}\left(-\exp(\mathbf{x}_i'\boldsymbol{\beta}_c) + y_i(\mathbf{x}_i'\boldsymbol{\beta}_c) - \log y_i!\right) + \log\left((\boldsymbol{\beta}_c - \boldsymbol{\mu}_0)' \mathbf{V}_0^{-1} (\boldsymbol{\beta}_c - \boldsymbol{\mu}_0)\right) \\
&\log \tilde{p}(\boldsymbol{\beta}_o \mid \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N}\left(-\exp(\mathbf{x}_i'\boldsymbol{\beta}_o) + y_i(\mathbf{x}_i'\boldsymbol{\beta}_o) - \log y_i!\right) + \log\left((\boldsymbol{\beta}_o - \boldsymbol{\mu}_0)' \mathbf{V}_0^{-1} (\boldsymbol{\beta}_o - \boldsymbol{\mu}_0)\right)
\end{aligned}
\tag{10}
$$

We first run a “prep” script “ with a generous number of burn-ins that simply aims at getting non-ridiculous acceptance rates (say 10% or more). You may run it a few times, experimenting with the setting for *c*. As a general rule, acceptance rates will increase with smaller settings for *c*.

Matlab script `mod7_Poisson_rwc_prep` and function `gs_Poisson_rwc` illustrate this process.

Then run the main script, this time using the estimated posterior variance matrix for $\boldsymbol{\beta}$ from the prep – results for V_c. Start with $c = 1$ and experiment with different settings for c until you have a reasonable acceptance rate, ideally in the 20-25% range.

Matlab script `mod7_Poisson_rwc` illustrates this process.

You may want to repeat this step a few times –working with an ever more refined estimate of V_c. Also, don’t be shy with your burn-ins – often several 100,000 's may be required until CD diagnostics check out.

Version 2: Independence Chain MH

In the *IMH* the CGD is no longer a function of the current draw $\boldsymbol{\beta}_o$, but rather – at least indirectly – of the data **y**, **X**. A popular version is a multivariate *t* distribution, with mean equal to the mode of the posterior kernel $\tilde{p}(\boldsymbol{\beta}\,|\,\mathbf{y},\mathbf{X})$, call it $\hat{\boldsymbol{\beta}}$, and variance ("scale matrix") equal to inverted negative Hessian flowing from the MLE routine, evaluated at $\hat{\boldsymbol{\beta}}$ and possibly scaled by some arbitrary scalar "*c*" (similar to the scalar in the RWMH version). Also, the degree-of-freedom parameter for the *mvt* (call it *v*) is specified at the onset along with the model priors. A smaller value of *v* (say in the 4-8 range) implies heavier tails, which is generally desirable. However, if *v* is chosen too small, too many candidate draws will be rejected.

Thus, in each round of the sampler, we first have to run a sub-routine MLE procedure to find $\hat{\boldsymbol{\beta}}$ as

$$\hat{\boldsymbol{\beta}} = \max_{\boldsymbol{\beta}} \left[\sum_{i=1}^{N} \left(-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right) + y_i\left(\mathbf{x}_i'\boldsymbol{\beta}\right) - \log y_i! \right) + \log\left(\left(\boldsymbol{\beta}-\boldsymbol{\mu}_0\right)' \mathbf{V}_0^{-1} \left(\boldsymbol{\beta}-\boldsymbol{\mu}_0\right) \right) \right] \tag{11}$$

It usually helps in terms of speed and precision to supply the analytical gradient and Hessian for this step. However, in most cases software-supplied numerical approximations for these constructs will be sufficient to obtain $\hat{\boldsymbol{\beta}}$ in a few iterations.

The MLE routine thus produces $\hat{\boldsymbol{\beta}}$ and $V_c = \left(-H\left(\hat{\boldsymbol{\beta}}\right)\right)^{-1}$ where $H(.)$ is the Hessian matrix evaluated at the MLE solution. We can now draw a candidate vector $\boldsymbol{\beta}_c$ from the *mvt* density, i.e.

$$\boldsymbol{\beta}_c \sim mvt\left(\hat{\boldsymbol{\beta}}, c\mathbf{V}_c, \nu\right) = q\left(\boldsymbol{\beta}_c\,|\,\mathbf{y},\mathbf{X}\right) \tag{12}$$

Since this CDG is no longer symmetric, we need to explicitly consider it in the acceptance probability, i.e.

$$\log\left(\alpha\left(\boldsymbol{\beta}_o \to \boldsymbol{\beta}_c\right)\right) = \min\left(\log\left(\frac{\tilde{p}\left(\boldsymbol{\beta}_c\,|\,\mathbf{y},\mathbf{X}\right) q\left(\boldsymbol{\beta}_o\,|\,\mathbf{y},\mathbf{X}\right)}{\tilde{p}\left(\boldsymbol{\beta}_o\,|\,\mathbf{y},\mathbf{X}\right) q\left(\boldsymbol{\beta}_c\,|\,\mathbf{y},\mathbf{X}\right)} \right), 0 \right) =$$

$$\min\left(\left(\log \tilde{p}\left(\boldsymbol{\beta}_c\,|\,\mathbf{y},\mathbf{X}\right) - \log q\left(\boldsymbol{\beta}_c\,|\,\mathbf{y},\mathbf{X}\right)\right) - \left(\log \tilde{p}\left(\boldsymbol{\beta}_o\,|\,\mathbf{y},\mathbf{X}\right) - \log q\left(\boldsymbol{\beta}_o\,|\,\mathbf{y},\mathbf{X}\right)\right), 0 \right), \quad \text{where}$$

$$\log \tilde{p}\left(\boldsymbol{\beta}_c\,|\,\mathbf{y},\mathbf{X}\right) = \sum_{i=1}^{N} \left(-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_c\right) + y_i\left(\mathbf{x}_i'\boldsymbol{\beta}_c\right) - \log y_i! \right) + \log\left(\left(\boldsymbol{\beta}_c-\boldsymbol{\mu}_0\right)' \mathbf{V}_0^{-1} \left(\boldsymbol{\beta}_c-\boldsymbol{\mu}_0\right) \right)$$

$$\log \tilde{p}\left(\boldsymbol{\beta}_o\,|\,\mathbf{y},\mathbf{X}\right) = \sum_{i=1}^{N} \left(-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_o\right) + y_i\left(\mathbf{x}_i'\boldsymbol{\beta}_o\right) - \log y_i! \right) + \log\left(\left(\boldsymbol{\beta}_o-\boldsymbol{\mu}_0\right)' \mathbf{V}_0^{-1} \left(\boldsymbol{\beta}_o-\boldsymbol{\mu}_0\right) \right) \tag{13}$$

$$\log q\left(\boldsymbol{\beta}_c\,|\,\mathbf{y},\mathbf{X}\right) = \log\left(mvt\left(\boldsymbol{\beta}_c\,|\,\hat{\boldsymbol{\beta}}, c\mathbf{V}_c, \nu\right)\right)$$

$$\log q\left(\boldsymbol{\beta}_o\,|\,\mathbf{y},\mathbf{X}\right) = \log\left(mvt\left(\boldsymbol{\beta}_o\,|\,\hat{\boldsymbol{\beta}}, c\mathbf{V}_c, \nu\right)\right)$$

Matlab script `mod7_Poisson_ic` with functions `gs_Poisson_ic` and `Poisson_beta_mle` show the implementation of this basic IMH.

You can see that compared to the RWC version, the IC version takes much longer per iteration, but converges much faster, such that overall runtime is comparable between the two approaches, at least for this application.

There are clear efficiency gains from using the IC approach, as is evident by inspection of the IEF statistics and the autocorrelation plots. Thus, breaking the direct dependence of the old and new draw via the RWC proposal density has the expected effect of reducing correlation in the Markov Chain.

Example 2: Hierarchical Poisson Model

Knowledge of the MH technique also enables us to tackle a broader variety of hierarchical models. Specifically, we can combine elements of the preceding Poisson model with those of the HNRM of Module 6 to specify and estimate a Hierarchical Poisson . For a recent application of this model see Davis and Moeltner (2010).

Here is the outline of Davis and Moeltner's hierarchical Poisson model with normally distributed random coefficients in the mean function – also called the Poisson-lognormal model. A detailed discussion on the Bayesian estimation of this model is given in Chib et al. (1998).

Consider an individual *i* that takes trips to *j=1…J* sites during a specific time period. Let the density function for trips be Poisson with the usual parameterized mean function. The hierarchical structure is added by letting some parameters in the mean function be "random", i.e. follow a (normal) hierarchical distribution. Thus, the structural model and likelihood are given as:

$$f\left(y_{ij} \mid \lambda_{ij}\right) = \frac{\exp\left(-\lambda_{ij}\right)\lambda_{ij}^{\,y_{ij}}}{y_{ip}!} \qquad \text{where}$$

$$\lambda_{ij} = \exp\left(\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{h}_{ij}'\boldsymbol{\gamma}_{\mathbf{i}}\right) \quad \text{and} \quad \boldsymbol{\gamma}_{\mathbf{i}} \sim mvn\left(\boldsymbol{\gamma}, \boldsymbol{\Sigma}\right). \tag{14}$$

$$p\left(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}\right) = \prod_{i=1}^{N}\left(\int_{\boldsymbol{\gamma}_{\mathbf{i}}}\left(\left(\prod_{j=1}^{J}\frac{\exp\left(-\lambda_{ij}\right)\lambda_{ij}^{\,y_{ij}}}{y_{ij}!}\right) f\left(\boldsymbol{\gamma}_{\mathbf{i}} \mid \boldsymbol{\gamma}, \boldsymbol{\Sigma}\right)\right) d\,\boldsymbol{\gamma}_{\mathbf{i}}\right)$$

We choose the usual multivariate normal priors for the fixed effects ($\boldsymbol{\beta}$) and the hierarchical expectation of the random effects ($\boldsymbol{\gamma}$), and an IW prior for the variance-covariance matrix of the random effects ($\boldsymbol{\Sigma}$).

$$p\left(\boldsymbol{\beta}\right) = \left(2\pi\right)^{-k_f/2}\left|\mathbf{V}_{\boldsymbol{\beta}}\right|^{-1/2}\exp\left(-\tfrac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}}\right)'\mathbf{V}_{\boldsymbol{\beta}}^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}}\right)\right)$$

$$p\left(\boldsymbol{\gamma}\right) = \left(2\pi\right)^{-k_r/2}\left|\mathbf{V}_{\boldsymbol{\gamma}}\right|^{-1/2}\exp\left(-\tfrac{1}{2}\left(\boldsymbol{\gamma}-\boldsymbol{\mu}_{\boldsymbol{\gamma}}\right)'\mathbf{V}_{\boldsymbol{\gamma}}^{-1}\left(\boldsymbol{\gamma}-\boldsymbol{\mu}_{\boldsymbol{\gamma}}\right)\right) \tag{15}$$

$$p\left(\boldsymbol{\Sigma}\right) = \left(2^{\nu_0 k_r/2}\pi^{k_r(k_r-1)/4}\prod_{i=1}^{k_r}\Gamma\left(\frac{\nu_0+1-i}{2}\right)\right)^{-1} * \left|\mathbf{S}_{\mathbf{0}}\right|^{\nu_0/2}\left|\boldsymbol{\Sigma}\right|^{-(\nu_0+k_r+1)/2}\exp\left(-\frac{1}{2}tr\left(\mathbf{S}_{\mathbf{0}}\cdot\boldsymbol{\Sigma}^{-1}\right)\right)$$

It would be difficult to break the resulting posterior into easy-to-draw-from conditional components, due to the multi-dimensional integral over the random effects in the likelihood function. We can again facilitate posterior simulation by augmenting the parameter space with the individual-level random effects, $\boldsymbol{\gamma}_\mathbf{i}$, i=1…N. The augmented posterior takes the following generic form:

$$\begin{aligned} &p\left(\boldsymbol{\beta},\boldsymbol{\Sigma},\boldsymbol{\gamma},\{\boldsymbol{\gamma}_\mathbf{i}\}_{i=1}^{N}\,|\,\mathbf{y},\mathbf{X}\right)\propto \\ &p(\boldsymbol{\beta})\,p\left(\{\boldsymbol{\gamma}_\mathbf{i}\}_{i=1}^{N},\boldsymbol{\gamma},\boldsymbol{\Sigma}\right)p(\mathbf{y}\,|\,\boldsymbol{\beta},\boldsymbol{\Sigma},\boldsymbol{\gamma},\boldsymbol{\gamma}_\mathbf{i})= \\ &p(\boldsymbol{\beta})\,p(\boldsymbol{\gamma})\,p(\boldsymbol{\Sigma})\prod_{i=1}^{N}p(\boldsymbol{\gamma}_\mathbf{i}\,|\,\boldsymbol{\gamma},\boldsymbol{\Sigma})p(\mathbf{y}\,|\,\boldsymbol{\beta},\boldsymbol{\gamma}_\mathbf{i}) \end{aligned} \tag{16}$$

The last component of (32) (I call it the conditional likelihood, as it conditions on the random effects) has a well-known closed form (the basic Poisson density). This circumvents the approximation of the multi-dimensional integral in the original likelihood function (as would be necessary in a classical estimation routine, e.g. via MLE).

Here is the explicit form of the augmented posterior:

$$\begin{aligned} &p\left(\boldsymbol{\beta},\boldsymbol{\Sigma},\boldsymbol{\gamma},\boldsymbol{\gamma}_\mathbf{i}\,|\,\mathbf{y},\mathbf{X}\right)\propto \\ &\exp\left(-\tfrac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}}\right)'\mathbf{V}_{\boldsymbol{\beta}}^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}}\right)\right)*\exp\left(-\tfrac{1}{2}\left(\boldsymbol{\gamma}-\boldsymbol{\mu}_{\boldsymbol{\gamma}}\right)'\mathbf{V}_{\boldsymbol{\gamma}}^{-1}\left(\boldsymbol{\gamma}-\boldsymbol{\mu}_{\boldsymbol{\gamma}}\right)\right)* \\ &|\boldsymbol{\Sigma}|^{-(v_0+k_r+1)/2}\exp\left(-\frac{1}{2}tr\left(\mathbf{S}_\mathbf{0}\cdot\boldsymbol{\Sigma}^{-1}\right)\right)* \\ &\prod_{i=1}^{N}|\boldsymbol{\Sigma}|^{-1/2}\exp\left(-\tfrac{1}{2}(\boldsymbol{\gamma}_\mathbf{i}-\boldsymbol{\gamma})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\gamma}_\mathbf{i}-\boldsymbol{\gamma})\right)*\prod_{i=1}^{N}\prod_{j=1}^{J}\frac{\exp\left(-\lambda_{ij}\right)\lambda_{ij}^{\,y_{ij}}}{y_{ij}!}\quad\text{with} \\ &\lambda_{ij}=\exp\left(\mathbf{x}_{\mathbf{ij}}'\boldsymbol{\beta}+\mathbf{h}_{\mathbf{ij}}'\boldsymbol{\gamma}_\mathbf{i}\right) \end{aligned} \tag{17}$$

The Gibbs Sampler proceeds in four blocks, two of which require a built-in Metropolis-Hastings (MH) routine. Details and options for the MH steps can also be found in Chib et al. (1998).

Step 1: Draw $\boldsymbol{\beta}$:

$$\begin{aligned} &p\left(\boldsymbol{\beta}\,|\,\mathbf{y},\mathbf{X},\boldsymbol{\Sigma},\boldsymbol{\gamma}_\mathbf{i}\right)\propto \\ &\exp\left(-\tfrac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}}\right)'\mathbf{V}_{\boldsymbol{\beta}}^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}}\right)\right)*\prod_{i=1}^{N}\prod_{j=1}^{J}\frac{\exp\left(-\lambda_{ij}\right)\lambda_{ij}^{\,y_{ij}}}{y_{ij}!}\quad\text{with} \\ &\lambda_{ij}=\exp\left(\mathbf{x}_{\mathbf{ij}}'\boldsymbol{\beta}+\mathbf{h}_{\mathbf{ij}}'\boldsymbol{\gamma}_\mathbf{i}\right) \end{aligned} \tag{18}$$

➔ MH (Davis & Moeltner use a Independence-Chain MH with a *mvt* proposal function)

Step 2: Draw the individual vectors of random effects (repeat N times)

$$p\left(\boldsymbol{\gamma}_\mathbf{i} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \mathbf{y}_\mathbf{i}, \mathbf{X}_\mathbf{i}\right) \propto$$
$$\exp\left(-\tfrac{1}{2}\left(\boldsymbol{\gamma}_\mathbf{i}-\boldsymbol{\gamma}\right)' \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\gamma}_\mathbf{i}-\boldsymbol{\gamma}\right)\right)\prod_{i=1}^{N}\prod_{j=1}^{J}\frac{\exp\left(-\lambda_{ij}\right)\lambda_{ij}^{\,y_{ij}}}{y_{ij}!} \tag{19}$$

➔ MH (Davis & Moeltner use a Independence-Chain MH with a *mvt* proposal function)

Step 3: Draw the vector of random effect expectations

$$p\left(\boldsymbol{\gamma} \mid \boldsymbol{\gamma}_\mathbf{i}, \mathbf{y}, \mathbf{X}\right) \propto$$
$$\exp\left(-\tfrac{1}{2}\left(\boldsymbol{\gamma}-\boldsymbol{\mu}_\gamma\right)' \mathbf{V}_\gamma^{-1}\left(\boldsymbol{\gamma}-\boldsymbol{\mu}_\gamma\right)\right) * \prod_{i=1}^{N}\exp\left(-\tfrac{1}{2}\left(\boldsymbol{\gamma}_\mathbf{i}-\boldsymbol{\gamma}\right)' \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\gamma}_\mathbf{i}-\boldsymbol{\gamma}\right)\right) =$$
$$\exp\left(-\tfrac{1}{2}\left(\left(\boldsymbol{\gamma}-\boldsymbol{\mu}_\gamma\right)' \mathbf{V}_\gamma^{-1}\left(\boldsymbol{\gamma}-\boldsymbol{\mu}_\gamma\right)+\sum_{i=1}^{N}\left(\boldsymbol{\gamma}_\mathbf{i}-\boldsymbol{\gamma}\right)' \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\gamma}_\mathbf{i}-\boldsymbol{\gamma}\right)\right)\right) \tag{20}$$

Thus:

$$\boldsymbol{\gamma} \mid \boldsymbol{\gamma}_\mathbf{i}, \mathbf{y}, \mathbf{X} \sim n\left(\boldsymbol{\mu}_\mathbf{1}, \mathbf{V}_\mathbf{1}\right)$$
$$\mathbf{V}_\mathbf{1} = \left(\mathbf{V}_\gamma^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1} \qquad \boldsymbol{\mu}_\mathbf{1} = \mathbf{V}_\mathbf{1}\left(\mathbf{V}_\gamma^{-1}\boldsymbol{\mu}_\gamma + \boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N}\boldsymbol{\gamma}_\mathbf{i}\right)$$

Step 4: Draw the random effect variance matrix

$$p\left(\boldsymbol{\Sigma} \mid \boldsymbol{\gamma}, \boldsymbol{\gamma}_\mathbf{i}\right) \propto$$
$$\left|\boldsymbol{\Sigma}\right|^{-\left(v_0+k_r+1+N\right)/2}\exp\left(-\frac{1}{2}\left(tr\left(\mathbf{S}_\mathbf{0}\cdot\boldsymbol{\Sigma}^{-1}\right)+\sum_{i=1}^{N}\left(\boldsymbol{\gamma}_\mathbf{i}-\boldsymbol{\gamma}\right)' \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\gamma}_\mathbf{i}-\boldsymbol{\gamma}\right)\right)\right)$$

Thus: (21)

$$\boldsymbol{\Sigma} \mid \boldsymbol{\gamma}, \boldsymbol{\gamma}_\mathbf{i} \sim IW\left(v_1, \mathbf{S}_\mathbf{1}\right)$$
$$v_1 = v_0 + N \qquad \mathbf{S}_\mathbf{1} = \mathbf{S}_\mathbf{0} + \sum_{i=1}^{N}\left(\boldsymbol{\gamma}_\mathbf{i}-\boldsymbol{\gamma}\right)\left(\boldsymbol{\gamma}_\mathbf{i}-\boldsymbol{\gamma}\right)'$$

As can be seen from the output file, the IEF scores are uncomfortable high for some parameters, and longer run-time plus thinning would have likely been a good idea in this case. However, posterior results for the key parameters are plausible, as are those for the posterior predictive densities (see paper).

Example 3: AR-1 Model

(see separate pdf file)

Some caveats when working with MH:

- Any *MH* step in your GS will generally hurt the efficiency of your posterior sampler, so expect higher IEF scores and lower m*'s. This effect is especially pronounced if you use an *RWMH* algorithm.

- IEF scores in the 10-30 range may often be as good as it gets, especially for the *RWMH* . As a general rule you can expect higher autocorrelation for the *RWMH* compared to the *IMH*. This trade-off comes with a price tag: The *IMH* usually takes much, much, longer to run. However, you may get away with a substantially fewer burn-ins! (see Matlab application).

- Try running the *IMH* without analytical gradient and Hessian in the MLE sub-routine. If the Sampler has trouble converging or is simply moving too slowly, you may need to spend some extra "pencil & paper" time to derive the analytical gradient and Hessian.

- Save time and avoid frustration by first running the *IMH* with just a few hundred iterations, then check your acceptance rates. There is no "fast & steady" rule on *acceptance rates for the IMH*. In theory, since the chain is independent, you want all draws to be accepted. However, this may still lead to high autocorrelation in the sequence even in the *IMH*. Let the IEF scores guide you in this respect. I have found that for most applications acceptance rates between 0.5 and 0.8 lead to best efficiency.

- Other "remedies" include experimenting with different parameter blockings and re-parameterization of the structural model.

Appendix A: Why / How MH works and the Equality of Joint Probabilities

(KPT p. 153)

Suppose the discrete RV θ has pmf $p(\theta)$. Let θ^{t-1} be the current value of the MH chain and assume that we can be certain that $\theta^{t-1} \sim p$. Now assume the next draw produced by the MH algorithm is θ^t. Show that $\theta^t \sim p$.[1]

Let θ^a and θ^b be two distinct points in the support of θ. Then the MH acceptance probability of *moving from* θ^b *to* θ^a is given as

$$\alpha(\theta^a \mid \theta^b) = \min\left\{1, \frac{p(\theta^a)q(\theta^b \mid \theta^a)}{p(\theta^b)q(\theta^a \mid \theta^b)}\right\} \tag{22}$$

[1] Naturally, this would be trivial if we had $\theta^t = \theta^{t-1}$.

Without loss in generality assume that $\theta^t \neq \theta^{t-1}$ and that $\alpha\left(\theta^a \mid \theta^b\right) > 1$ such that its reciprocal $\alpha\left(\theta^b \mid \theta^a\right) < 1$. Now consider the joint probability of observing the sequence θ^b, θ^a in the MH algorithm, i.e.

$$\Pr\left(\theta^{t-1} = \theta^b, \theta^t = \theta^a\right) = \Pr\left(\theta^t = \theta^a \mid \theta^{t-1} = \theta^b\right) \Pr\left(\theta^{t-1} = \theta^b\right) \tag{23}$$

The first term on the right hand side is the conditional probability of arriving at θ^a when the current draw is θ^b. This conditional can be expressed as the probability of *drawing* θ^a when the current draw is θ^b times the probability of accepting θ^a once it's drawn, i.e.

$$\Pr\left(\theta^t = \theta^a \mid \theta^{t-1} = \theta^b\right) = q\left(\theta^a \mid \theta^b\right) \alpha\left(\theta^a \mid \theta^b\right) = q\left(\theta^a \mid \theta^b\right) \tag{24}$$

where the last equality follows from our assumption that $\alpha\left(\theta^a \mid \theta^b\right) > 1$. Thus we get

$$\Pr\left(\theta^{t-1} = \theta^b, \theta^t = \theta^a\right) = q\left(\theta^a \mid \theta^b\right) \Pr\left(\theta^{t-1} = \theta^b\right) = q\left(\theta^a \mid \theta^b\right) p\left(\theta^b\right) \tag{25}$$

Now consider a different joint probability

$$\begin{aligned} &\Pr\left(\theta^{t-1} = \theta^a, \theta^t = \theta^b\right) = \Pr\left(\theta^t = \theta^b \mid \theta^{t-1} = \theta^a\right) \Pr\left(\theta^{t-1} = \theta^a\right) = \\ &q\left(\theta^b \mid \theta^a\right) \alpha\left(\theta^b \mid \theta^a\right) p\left(\theta^a\right) = \\ &q\left(\theta^b \mid \theta^a\right) \frac{p\left(\theta^b\right) q\left(\theta^a \mid \theta^b\right)}{p\left(\theta^a\right) q\left(\theta^b \mid \theta^a\right)} p\left(\theta^a\right) = p\left(\theta^b\right) q\left(\theta^a \mid \theta^b\right) \end{aligned} \tag{26}$$

Thus the two joint probabilities are indeed equal. This *equality of joint probabilities* implies that

$$\Pr\left(\theta^t = \theta^a \mid \theta^{t-1} = \theta^b\right) \Pr\left(\theta^{t-1} = \theta^b\right) = \Pr\left(\theta^t = \theta^b \mid \theta^{t-1} = \theta^a\right) \Pr\left(\theta^{t-1} = \theta^a\right) \tag{27}$$

Now summing over θ^a (i.e. marginalizing the joint pmf over θ^a) yields

$$\Pr\left(\theta^{t-1} = \theta^b\right) \sum_{\theta^a = \theta_{\min}}^{\theta_{\max}} \Pr\left(\theta^t = \theta^a \mid \theta^{t-1} = \theta^b\right) = \sum_{\theta^a = \theta_{\min}}^{\theta_{\max}} \Pr\left(\theta^t = \theta^b \mid \theta^{t-1} = \theta^a\right) \Pr\left(\theta^{t-1} = \theta^a\right) \tag{28}$$

The second term on the left hand side is just the full *cdf* of a conditional density for a discrete variable (summing to 1), whereas the right hand side is the definition for a marginal density for discrete variables. Thus we get:

$$\Pr\left(\theta^{t-1} = \theta^b\right) = \Pr\left(\theta^t = \theta^b\right) \tag{29}$$

Since $\theta^{t-1} \sim p(\theta)$ it follows that $\theta^t \sim p(\theta)$ as well.

Appendix B: Marginal Likelihood Estimation for Models with MH Components

Chib and Jeliazkov (2001)

To compare models that use MH in their posterior simulator we need to derive their marginal likelihood. This requires a modification of the Chib (1995) method, as illustrated in Chib and Jeliazkov (2001), henceforth C/J.

As before, we can write the log *mL*, evaluated at point $\overline{\boldsymbol{\theta}}$ (e.g. the posterior mean) as

$$\log p(\mathbf{y}) = \log p(\mathbf{y} \mid \overline{\boldsymbol{\theta}}) + \log p(\overline{\boldsymbol{\theta}}) - \log p(\overline{\boldsymbol{\theta}} \mid \mathbf{y}) \tag{1}$$

The trick is to find an estimate of the posterior ordinate $p(\overline{\boldsymbol{\theta}} \mid \mathbf{y})$. Assume that one or more of the normalizing constants of full conditional densities are not known (that's why we're using an *MH* routine in the posterior sampler). This is where the C/J method comes in.

One Block Sampling

Suppose the posterior density $p(\boldsymbol{\theta} \mid \mathbf{y}) \propto p(\boldsymbol{\theta})\, p(\mathbf{y} \mid \boldsymbol{\theta})$ is sampled in one block via *MH*[2]. Our goal is again to estimate the posterior ordinate at the posterior mean, i.e. $p(\overline{\boldsymbol{\theta}} \mid \mathbf{y})$.

Let $q(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b \mid \mathbf{y})$ be the proposal density for the transition from $\boldsymbol{\theta}^a$ to $\boldsymbol{\theta}^b$ (the dependence of q on $\mathbf{y}$ denotes the most general case and may not always apply). Then, as discussed before, the probability of move (of accepting $\boldsymbol{\theta}^b$) is given by

$$\alpha(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b \mid \mathbf{y}) = \min\left\{1, \frac{p(\mathbf{y} \mid \boldsymbol{\theta}^b)\, p(\boldsymbol{\theta}^b)\, q(\boldsymbol{\theta}^b, \boldsymbol{\theta}^a \mid \mathbf{y})}{p(\mathbf{y} \mid \boldsymbol{\theta}^a)\, p(\boldsymbol{\theta}^a)\, q(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b \mid \mathbf{y})}\right\} \tag{2}$$

For any two points $\boldsymbol{\theta}^0$ and $\overline{\boldsymbol{\theta}}$ denote the joint probability of observing the sequence $\boldsymbol{\theta}^0, \overline{\boldsymbol{\theta}}$ as

$$\Pr(\boldsymbol{\theta}^0, \overline{\boldsymbol{\theta}} \mid \mathbf{y}) = \Pr(\overline{\boldsymbol{\theta}} \mid \mathbf{y}, \boldsymbol{\theta}^0) \Pr(\boldsymbol{\theta}^0 \mid \mathbf{y}) = q(\boldsymbol{\theta}^0, \overline{\boldsymbol{\theta}} \mid \mathbf{y}) \alpha(\boldsymbol{\theta}^0, \overline{\boldsymbol{\theta}} \mid \mathbf{y})\, p(\boldsymbol{\theta}^0 \mid \mathbf{y}) =$$

$$\frac{p(\mathbf{y} \mid \overline{\boldsymbol{\theta}})\, p(\overline{\boldsymbol{\theta}})\, q(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta}^0 \mid \mathbf{y})}{p(\mathbf{y} \mid \boldsymbol{\theta}^0)\, p(\boldsymbol{\theta}^0)}\, p(\boldsymbol{\theta}^0 \mid \mathbf{y}) \tag{3}$$

where it is assumed that $\boldsymbol{\theta}^0$ is a draw from the posterior density $p(\boldsymbol{\theta}^0 \mid \mathbf{y})$. From KPT, p. 153ff (see appendix Lecture notes 8) we know that this joint density is reversible, i.e. $\Pr(\boldsymbol{\theta}^0, \overline{\boldsymbol{\theta}} \mid \mathbf{y}) = \Pr(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta}^0 \mid \mathbf{y})$, or

$$q(\boldsymbol{\theta}^0, \overline{\boldsymbol{\theta}} \mid \mathbf{y}) \alpha(\boldsymbol{\theta}^0, \overline{\boldsymbol{\theta}} \mid \mathbf{y})\, p(\boldsymbol{\theta}^0 \mid \mathbf{y}) = q(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta}^0 \mid \mathbf{y}) \alpha(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta}^0 \mid \mathbf{y})\, p(\overline{\boldsymbol{\theta}} \mid \mathbf{y}) \tag{4}$$

Thus, for a *specific* draw of $\boldsymbol{\theta}^0$ we have

[2] We encountered this case in the generic Poisson model.

$$p\left(\overline{\boldsymbol{\theta}} \mid \mathbf{y}, \boldsymbol{\theta}^0\right) = \frac{q\left(\boldsymbol{\theta}^0, \overline{\boldsymbol{\theta}} \mid \mathbf{y}\right) \alpha\left(\boldsymbol{\theta}^0, \overline{\boldsymbol{\theta}} \mid \mathbf{y}\right) p\left(\boldsymbol{\theta}^0 \mid \mathbf{y}\right)}{q\left(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta}^0 \mid \mathbf{y}\right) \alpha\left(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta}^0 \mid \mathbf{y}\right)} \tag{5}$$

Marginalizing over all possible values of $\boldsymbol{\theta}^0$, i.e. over $\boldsymbol{\theta}$, this yields

$$p\left(\overline{\boldsymbol{\theta}} \mid \mathbf{y}\right) = \frac{\int q\left(\boldsymbol{\theta}, \overline{\boldsymbol{\theta}} \mid \mathbf{y}\right) \alpha\left(\boldsymbol{\theta}, \overline{\boldsymbol{\theta}} \mid \mathbf{y}\right) p\left(\boldsymbol{\theta} \mid \mathbf{y}\right) d\boldsymbol{\theta}}{\int \alpha\left(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta} \mid \mathbf{y}\right) q\left(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta} \mid \mathbf{y}\right) d\boldsymbol{\theta}} \tag{6}$$

To highlight the estimation strategy, we can write this in terms of expectations as

$$p\left(\overline{\boldsymbol{\theta}} \mid \mathbf{y}\right) = \frac{E_{p(\boldsymbol{\theta}|\mathbf{y})}\left\{q\left(\boldsymbol{\theta}, \overline{\boldsymbol{\theta}} \mid \mathbf{y}\right) \alpha\left(\boldsymbol{\theta}, \overline{\boldsymbol{\theta}} \mid \mathbf{y}\right)\right\}}{E_{q\left(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta} \mid \mathbf{y}\right)}\left\{\alpha\left(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta} \mid \mathbf{y}\right)\right\}} \tag{7}$$

This yields a simulation-consistent estimate of

$$\hat{p}\left(\overline{\boldsymbol{\theta}} \mid \mathbf{y}\right) = \frac{M^{-1} \sum_{g=1}^{M} q\left(\boldsymbol{\theta}^g, \overline{\boldsymbol{\theta}} \mid \mathbf{y}\right) \alpha\left(\boldsymbol{\theta}^g, \overline{\boldsymbol{\theta}} \mid \mathbf{y}\right)}{J^{-1} \sum_{j=1}^{J} \alpha\left(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta}^j \mid \mathbf{y}\right)} \tag{8}$$

where $\boldsymbol{\theta}^g$ are the sampled draws from the posterior distribution (from the original sampler run) and $\boldsymbol{\theta}^j$ are draws from $q\left(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta} \mid \mathbf{y}\right)$ given the fixed value $\overline{\boldsymbol{\theta}}$, using a follow-up routine to the main sampler (since we won't know $\overline{\boldsymbol{\theta}}$ until we're done with the main run). Note that in practice we usually set *M* = *J*.

Note: If *S* (the area over which the posterior of $\boldsymbol{\theta}$ is defined) is a proper *subset* of $\Re^d$ (the dimension of $\boldsymbol{\theta}$) then all $\boldsymbol{\theta}^j$'s from $q\left(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta} \mid \mathbf{y}\right)$ that fall outside of *S* are included in the average of the denominator of (8), with $\alpha(.) = 0$. (Example: some of the elements of $\boldsymbol{\theta}$ might be truncated).

For an example of this single-block application see `mod7_Poisson_rwc_chib`.

Two Parameter Blocks and Multiple Latent Variable Blocks

For simplicity assume the normalizing constant of only the *first* full conditional is not known, and that this density is sampled by *MH*.

The log-ML is given by

$$\log m(\mathbf{y}) = \log p\left(\mathbf{y} \mid \overline{\boldsymbol{\theta}}_1, \overline{\boldsymbol{\theta}}_2\right) + \log p\left(\overline{\boldsymbol{\theta}}_1, \overline{\boldsymbol{\theta}}_2\right) - \log p\left(\overline{\boldsymbol{\theta}}_1, \overline{\boldsymbol{\theta}}_2 \mid \mathbf{y}\right) \tag{9}$$

and the objective is to estimate $p\left(\overline{\boldsymbol{\theta}}_1, \overline{\boldsymbol{\theta}}_2 \mid \mathbf{y}\right)$. We assume that the likelihood ordinate is readily available either by direct computation or by a Monte Carlo Integration method. As in Chib (1995) we decompose the posterior ordinate as

$$p\left(\overline{\boldsymbol{\theta}}_1, \overline{\boldsymbol{\theta}}_2 \mid \mathbf{y}\right) = p\left(\overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}\right) p\left(\overline{\boldsymbol{\theta}}_2 \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_1\right) \tag{10}$$

where the marginal density ordinate $p\left(\overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}\right)$ can*not* be estimated as usual via Monte Carlo integration since, by assumption, the normalizing constant of $p\left(\boldsymbol{\theta}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right)$ is not known.

Let $q\left(\boldsymbol{\theta}_1^a, \boldsymbol{\theta}_1^b \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right)$ be the proposal density for transition from $\boldsymbol{\theta}_1^a$ to $\boldsymbol{\theta}_1^b$ (where for generality *q* is allowed to depend on data and the two remaining blocks). The probability of move is

$$\alpha\left(\boldsymbol{\theta}_1^a, \boldsymbol{\theta}_1^b \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) = \min\left\{1, \frac{p\left(\mathbf{y} \mid \boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2, \mathbf{z}\right) p\left(\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2\right) q\left(\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_1^a \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right)}{p\left(\mathbf{y} \mid \boldsymbol{\theta}_1^a, \boldsymbol{\theta}_2, \mathbf{z}\right) p\left(\boldsymbol{\theta}_1^a, \boldsymbol{\theta}_2\right) q\left(\boldsymbol{\theta}_1^a, \boldsymbol{\theta}_1^b \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right)}\right\} \tag{11}$$

Consider again the joint probability

$$\Pr\left(\boldsymbol{\theta}_1^0, \overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) = q\left(\boldsymbol{\theta}_1^0, \overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) \alpha\left(\boldsymbol{\theta}_1^0, \overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) p\left(\boldsymbol{\theta}_1^0 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) \tag{12}$$

Using again the rule of reversibility (or "local reversibility" in this case) we get

$$\begin{aligned} &q\left(\boldsymbol{\theta}_1^0, \overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) \alpha\left(\boldsymbol{\theta}_1^0, \overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) p\left(\boldsymbol{\theta}_1^0 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) = \\ &q\left(\overline{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1^0 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) \alpha\left(\overline{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1^0 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) p\left(\overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) \end{aligned} \tag{13}$$

Now multiply both sides by $p\left(\boldsymbol{\theta}_2, \mathbf{z} \mid \mathbf{y}\right)$ to get the full augmented posterior $p\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{z} \mid \mathbf{y}\right)$

$$\begin{aligned} &q\left(\boldsymbol{\theta}_1^0, \overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) \alpha\left(\boldsymbol{\theta}_1^0, \overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) p\left(\boldsymbol{\theta}_1^0 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) p\left(\boldsymbol{\theta}_2, \mathbf{z} \mid \mathbf{y}\right) = \\ &q\left(\overline{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1^0 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) \alpha\left(\overline{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1^0 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) p\left(\overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) p\left(\boldsymbol{\theta}_2, \mathbf{z} \mid \mathbf{y}\right) \quad \text{or} \end{aligned} \tag{14}$$

$$\begin{aligned} &q\left(\boldsymbol{\theta}_1^0, \overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) \alpha\left(\boldsymbol{\theta}_1^0, \overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) p\left(\boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2, \mathbf{z} \mid \mathbf{y}\right) = \\ &q\left(\overline{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1^0 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) \alpha\left(\overline{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1^0 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) p\left(\overline{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2, \mathbf{z} \mid \mathbf{y}\right) \end{aligned}$$

Then integrate over $\boldsymbol{\psi} = \left[\boldsymbol{\theta}_1' \quad \boldsymbol{\theta}_2' \quad \mathbf{z}'\right]'$, i.e. all parameters and the augmented data (since we're now integrating over all possible values of $\boldsymbol{\theta}_1^0$ we're switching notation from $\boldsymbol{\theta}_1^0$ to $\boldsymbol{\theta}_1$):

$$\begin{aligned} &\int q\left(\boldsymbol{\theta}_1, \overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) \alpha\left(\boldsymbol{\theta}_1, \overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) p\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{z} \mid \mathbf{y},\right) d\boldsymbol{\psi} = \\ &\int q\left(\overline{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) \alpha\left(\overline{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) p\left(\overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}\right) p\left(\boldsymbol{\theta}_2, \mathbf{z} \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_1\right) d\boldsymbol{\psi} \end{aligned} \tag{15}$$

since

$$p\left(\overline{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2, \mathbf{z} \mid \mathbf{y}\right) = p\left(\overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) p\left(\boldsymbol{\theta}_2, \mathbf{z} \mid \mathbf{y}\right)$$

It follows immediately that

$$\begin{aligned} p\left(\overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}\right) = \quad & \frac{\int \alpha\left(\boldsymbol{\theta}_1, \overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) q\left(\boldsymbol{\theta}_1, \overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) p\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{z} \mid \mathbf{y}\right) d\boldsymbol{\psi}}{\int \alpha\left(\overline{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) q\left(\overline{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) p\left(\boldsymbol{\theta}_2, \mathbf{z} \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_1\right) d\boldsymbol{\psi}} = \\ & \frac{E_{p\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{z} \mid \mathbf{y}\right)}\left\{\alpha\left(\boldsymbol{\theta}_1, \overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) q\left(\boldsymbol{\theta}_1, \overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right)\right\}}{E_{q\left(\overline{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) p\left(\boldsymbol{\theta}_2, \mathbf{z} \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_1\right)}\left\{\alpha\left(\overline{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right)\right\}} \end{aligned} \tag{16}$$

To estimate the numerator by MC take the draws $\left\{\boldsymbol{\theta}_1^{\mathbf{g}}, \boldsymbol{\theta}_2^{\mathbf{g}}, \mathbf{z}^{\mathbf{g}}\right\}_{g=1}^{M}$ from the full (original) run and average the quantity $\alpha\left(\boldsymbol{\theta}_1^g, \overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2^g, \mathbf{z}^g\right) q\left(\overline{\boldsymbol{\theta}}_1 \mid \boldsymbol{\theta}_1^g, \mathbf{y}, \boldsymbol{\theta}_2^g, \mathbf{z}^g\right)$. The expectation in the denominator is conditional on $\overline{\boldsymbol{\theta}}_1$ and the original GS does (obviously) not generate draws conditional on $\overline{\boldsymbol{\theta}}_1$. So *continue* the MCMC simulation for an additional *J* iterations with the two full conditional densities $p\left(\boldsymbol{\theta}_2 \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_1, \mathbf{z}\right)$ and $p\left(\mathbf{z} \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2\right)$. At each iteration, given $\boldsymbol{\theta}_2^{\mathbf{j}}$, $\mathbf{z}^{\mathbf{j}}$ we generate a variate

$$\boldsymbol{\theta}_1^{\mathbf{j}} \sim q\left(\overline{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2^{\mathbf{j}}, \mathbf{z}^{\mathbf{j}}\right) \tag{17}$$

This leads to the triple $\boldsymbol{\theta}_1^{\mathbf{j}}, \boldsymbol{\theta}_2^{\mathbf{j}}, \mathbf{z}^j$ that is a draw from $q\left(\overline{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}\right) p\left(\boldsymbol{\theta}_2, \mathbf{z} \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_1\right)$. The marginal ordinate can now be estimated as

$$\hat{p}\left(\overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}\right) = \frac{M^{-1} \sum_{g=1}^{M} \alpha\left(\boldsymbol{\theta}_1^g, \overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2^g, \mathbf{z}^g\right) q\left(\boldsymbol{\theta}_1^g, \overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2^g, \mathbf{z}^g\right)}{J^{-1} \sum_{j=1}^{J} \alpha\left(\overline{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1^j \mid \mathbf{y}, \boldsymbol{\theta}_2^j, \mathbf{z}^j\right)} \tag{18}$$

Next, the values $\mathbf{z}^{\mathbf{j}}$ from the preceding reduced run, which are marginally from $p\left(\mathbf{z} \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_1\right)$ are used to form the average

$$\hat{p}\left(\overline{\boldsymbol{\theta}}_2 \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_1\right) = J^{-1} \sum_{j=1}^{J} p\left(\overline{\boldsymbol{\theta}}_2 \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_1, \mathbf{z}^{\mathbf{j}}\right) \tag{19}$$

which is a simulation-consistent estimate of $p\left(\overline{\boldsymbol{\theta}}_2 \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_1\right)$. Thus, at the conclusion of the reduced run both ordinates are available, and we get

$$\log \hat{p}\left(\mathbf{y}\right) = \log p\left(\mathbf{y} \mid \overline{\boldsymbol{\theta}}_1, \overline{\boldsymbol{\theta}}_2\right) + \log p\left(\overline{\boldsymbol{\theta}}_1, \overline{\boldsymbol{\theta}}_2\right) - \left(\log \hat{p}\left(\overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}\right) + \log \hat{p}\left(\overline{\boldsymbol{\theta}}_2 \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_1\right)\right) \tag{20}$$

Three remarks: (i) A single reduced run, augmented to sample $\boldsymbol{\theta}_1$ from the proposal density, delivers the variates that are used in the calculation of the ordinates $p\left(\overline{\boldsymbol{\theta}}_1 \mid \mathbf{y}\right)$ and $p\left(\overline{\boldsymbol{\theta}}_2 \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_1\right)$. (ii) If one places the recalcitrant ordinate first in the decomposition of the posterior ordinate, the reduced run does not involve any *MH* steps. (iii) This same approach can also be applied when the full conditional density of **z** is sampled by a sequence of *MH* steps.

Multiple Parameter blocks

We will abstract from latent data in this section, as it can be handled as outlined in the previous section. Suppose we have *B* blocks, with $\boldsymbol{\theta}_k \in S_k \subseteq \Re^{d_k}$. Write the posterior ordinate at $\overline{\boldsymbol{\theta}}$ as $p\left(\overline{\boldsymbol{\theta}}_1, \overline{\boldsymbol{\theta}}_2, \cdots, \overline{\boldsymbol{\theta}}_B \mid \mathbf{y}\right) = \prod_{i=1}^{B} p\left(\overline{\boldsymbol{\theta}}_i \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_1, \cdots, \overline{\boldsymbol{\theta}}_{i-1}\right)$ and consider the estimation of the reduced ordinate $p\left(\overline{\boldsymbol{\theta}}_i \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_1, \cdots, \overline{\boldsymbol{\theta}}_{i-1}\right)$. Let $\boldsymbol{\psi}_{i-1}$ denote the blocks up to *i* and $\boldsymbol{\psi}_{i+1}$ the blocks beyond *i.* (So we have $\boldsymbol{\theta}_{-\mathbf{i}} = \left(\boldsymbol{\psi}_{-\mathbf{i}}' \quad \boldsymbol{\psi}_{+\mathbf{i}}'\right)'$).

Suppose that the full conditional density $p(\boldsymbol{\theta}_i \mid \mathbf{y}, \boldsymbol{\theta}_{-i}) \propto p(\boldsymbol{\theta})\, p(\mathbf{y} \mid \boldsymbol{\theta}), i = 1 \ldots B$ is sampled by the *MH* algorithm with proposal density $q\left(\boldsymbol{\theta}_i^a, \boldsymbol{\theta}_i^b \mid \mathbf{y}, \boldsymbol{\psi}_{i-1}, \boldsymbol{\psi}_{i+1}\right)$ and probability of move

$$\alpha\left(\boldsymbol{\theta}_i^a, \boldsymbol{\theta}_i^b \mid \mathbf{y}, \boldsymbol{\psi}_{i-1}, \boldsymbol{\psi}_{i+1}\right) = \min\left\{1, \frac{p\left(\mathbf{y} \mid \boldsymbol{\theta}_i^b, \boldsymbol{\psi}_{i-1}, \boldsymbol{\psi}_{i+1}\right) p\left(\boldsymbol{\theta}_i^b, \boldsymbol{\theta}_{-i}\right) q\left(\boldsymbol{\theta}_i^b, \boldsymbol{\theta}_i^a \mid \mathbf{y}, \boldsymbol{\psi}_{i-1}, \boldsymbol{\psi}_{i+1}\right)}{p\left(\mathbf{y} \mid \boldsymbol{\theta}_i^a, \boldsymbol{\psi}_{i-1}, \boldsymbol{\psi}_{i+1}\right) p\left(\boldsymbol{\theta}_i^a, \boldsymbol{\theta}_{-i}\right) q\left(\boldsymbol{\theta}_i^a, \boldsymbol{\theta}_i^b \mid \mathbf{y}, \boldsymbol{\psi}_{i-1}, \boldsymbol{\psi}_{i+1}\right)}\right\} \tag{21}$$

By completely analogous arguments from above we get

$$\begin{aligned} &q\left(\boldsymbol{\theta}_i^0, \overline{\boldsymbol{\theta}}_i \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}\right) \alpha\left(\boldsymbol{\theta}_i^0, \overline{\boldsymbol{\theta}}_i \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}\right) p\left(\boldsymbol{\theta}_i^0 \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}\right) = \\ &q\left(\overline{\boldsymbol{\theta}}_i, \boldsymbol{\theta}_i^0 \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}\right) \alpha\left(\overline{\boldsymbol{\theta}}_i, \boldsymbol{\theta}_i^0 \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}\right) p\left(\overline{\boldsymbol{\theta}}_i \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}\right) \end{aligned} \tag{22}$$

Where the use of $\overline{\boldsymbol{\psi}}_{i-1}$ indicates that the elements in $\boldsymbol{\psi}_{i-1}$ have been set to their posterior mean. Now multiply both sides by $p\left(\boldsymbol{\psi}_{i+1} \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}\right)$ and integrate over $\left(\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i+1}\right)$ to obtain

$$\begin{aligned} &p\left(\overline{\boldsymbol{\theta}}_i \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_1, \cdots, \overline{\boldsymbol{\theta}}_{i-1}\right) = \frac{\int \alpha\left(\boldsymbol{\theta}_i, \overline{\boldsymbol{\theta}}_i \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}\right) q\left(\boldsymbol{\theta}_i, \overline{\boldsymbol{\theta}}_i \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}\right) p\left(\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i+1} \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}\right) d\left(\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i+1}\right)}{\int \alpha\left(\overline{\boldsymbol{\theta}}_i, \boldsymbol{\theta}_i \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}\right) q\left(\overline{\boldsymbol{\theta}}_i, \boldsymbol{\theta}_i \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}\right) p\left(\boldsymbol{\psi}_{i+1} \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_i, \overline{\boldsymbol{\psi}}_{i-1}\right) d\left(\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i+1}\right)} = \\ &\frac{E_{p(\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i+1} \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1})}\left\{\alpha\left(\boldsymbol{\theta}_i, \overline{\boldsymbol{\theta}}_i \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}\right) q\left(\boldsymbol{\theta}_i, \overline{\boldsymbol{\theta}}_i \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}\right)\right\}}{E_{q(\overline{\boldsymbol{\theta}}_i, \boldsymbol{\theta}_i \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1})\, p(\boldsymbol{\psi}_{i+1} \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_i, \overline{\boldsymbol{\psi}}_{i-1})}\left\{\alpha\left(\overline{\boldsymbol{\theta}}_i, \boldsymbol{\theta}_i \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}\right)\right\}} \end{aligned} \tag{23}$$

These two integrals can be estimated as before from the output of reduced MCMC runs as follows:

Step 1(Numerator):
Set $\boldsymbol{\psi}_{i-1} = \overline{\boldsymbol{\psi}}_{i-1}$ and sample the reduced set of full conditional distributions $p\left(\boldsymbol{\theta}_k \mid \mathbf{y}, \boldsymbol{\theta}_{-k}, \overline{\boldsymbol{\psi}}_{i-1}\right), k = i, \ldots, B$. Let the generated draws be $\left\{\boldsymbol{\theta}_i^g, \ldots, \boldsymbol{\theta}_B^g\right\}, g = 1 \ldots M$. We can now approximate the numerator in (23) by

$$\begin{aligned} &E_{p(\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i+1} \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1})}\left(\alpha\left(\boldsymbol{\theta}_i, \overline{\boldsymbol{\theta}}_i \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}\right) q\left(\boldsymbol{\theta}_i, \overline{\boldsymbol{\theta}}_i \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}\right)\right) = \\ &M^{-1} \sum_{i=1}^{M} \alpha\left(\boldsymbol{\theta}_i^g, \overline{\boldsymbol{\theta}}_i \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}^g\right) q\left(\boldsymbol{\theta}_i^g, \overline{\boldsymbol{\theta}}_i \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}^g\right) \end{aligned} \tag{24}$$

Step 2 (Denominator):
In contrast to the numerator, the expectation in the denominator is conditioned on $\overline{\boldsymbol{\theta}}_i$. Thus, we need to go into another reduced run, drawing $p\left(\boldsymbol{\theta}_k \mid \mathbf{y}, \boldsymbol{\theta}_{-k}, \overline{\boldsymbol{\psi}}_{i-1}, \overline{\boldsymbol{\theta}}_i\right), k = i+1, \ldots, B$. This yields $\left\{\boldsymbol{\theta}_{i+1}^j, \ldots, \boldsymbol{\theta}_B^j\right\}, j = 1 \ldots J$. At each step also draw $\boldsymbol{\theta}_i^j \sim q\left(\overline{\boldsymbol{\theta}}_i, \boldsymbol{\theta}_i \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}^j\right)$. Together, the draws of $\left\{\boldsymbol{\theta}_i^j, \boldsymbol{\theta}_{i+1}^j, \ldots, \boldsymbol{\theta}_B^j\right\}$ are draws from $q\left(\overline{\boldsymbol{\theta}}_i, \boldsymbol{\theta}_i \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}\right) p\left(\boldsymbol{\psi}_{i+1} \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_i, \overline{\boldsymbol{\psi}}_{i-1}\right)$. We can now approximate the denominator as

$$\hat{E}_{q(\overline{\boldsymbol{\theta}}_i, \boldsymbol{\theta}_i \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1})\, p(\boldsymbol{\psi}_{i+1} \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_i, \overline{\boldsymbol{\psi}}_{i-1})}\left(\alpha\left(\overline{\boldsymbol{\theta}}_i, \boldsymbol{\theta}_i \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}\right)\right) = J^{-1} \sum_{j=1}^{J} \alpha\left(\overline{\boldsymbol{\theta}}_i, \boldsymbol{\theta}_i^j \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}^j\right) \tag{25}$$

where the average in the denominator may include zeros if there are values of $\boldsymbol{\theta}_i^j$ that lie outside the support of the posterior *S*. Thus, the reduced ordinate in question can be estimated as

$$\hat{p}\left(\overline{\boldsymbol{\theta}}_i \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_1, \cdots, \overline{\boldsymbol{\theta}}_{i-1}\right) = \frac{M^{-1}\sum_{i=1}^{M} \alpha\left(\boldsymbol{\theta}_i^g, \overline{\boldsymbol{\theta}}_i \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}^g\right) q\left(\boldsymbol{\theta}_i^g, \overline{\boldsymbol{\theta}}_i \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}^g\right)}{J^{-1}\sum_{j=1}^{J} \alpha\left(\overline{\boldsymbol{\theta}}_i, \boldsymbol{\theta}_i^j \mid \mathbf{y}, \overline{\boldsymbol{\psi}}_{i-1}, \boldsymbol{\psi}_{i+1}^j\right)} \tag{26}$$

Note: The draws of $\left\{\boldsymbol{\theta}_{i+1}^j, \ldots, \boldsymbol{\theta}_B^j\right\}, j = 1 \ldots J$ from Step 2 are also used to estimate the *numerator* of the next reduced posterior ordinate $p\left(\overline{\boldsymbol{\theta}}_{i+1} \mid \mathbf{y}, \overline{\boldsymbol{\theta}}_1, \cdots, \overline{\boldsymbol{\theta}}_i\right)$.

References:

Chib, S. 2001. Markov Chain Monte Carlo methods: Computation and inference. In *Handbook of Econometrics*, J. J. Heckman and E. Leamer (eds.). Elsevier Science, p. 3569 – 3649

Chib, S. and E. Greenberg. 1995. Understanding the Metropolis-Hastings algorithm. *The American Statistician* **49:** 327-335.

Chib, S., E. Greenberg and R. Winkelmann. 1998. "Posterior simulation and Bayes factors in panel count data models." *Journal of Econometrics* 86, 33-54.

Chib, Siddhartha and Bradley P. Carlin. 1999. "On MCMC sampling in hierarchical longitudinal models." *Statistics and Computing* 9, 17-26.

Chib, S. and I. Jeliazkov. 2001. Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association* **95:** 270-281.

Davis, A. and K. Moeltner. 2010. "Valuing the Prevention of an Infestation: The Threat of the New Zealand Mud Snail in Northern Nevada. ." *Agricultural and Resource Economics Review*.

Kass, R. E. and A. E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90, 773-795.