MCMC and time-varying volatility models

Petros Dellaportas

ECAS, Lugano, October 2001

The Bayesian integration problem

- Obtain data $\boldsymbol{y}$.
- Assume a model that gives rise to a likelihood function $L(\boldsymbol{y}|\boldsymbol{\theta})$ through which we receive information about the parameters $\boldsymbol{\theta} = (\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(d)})$ from $\boldsymbol{y}$.
- Assign prior distributions $p(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$.
- Obtain the posterior density

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{L(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int L(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

- The posterior density provides the basis for inferences about $\boldsymbol{\theta}$, so all posterior summaries of interest can be derived from it.

What we need?

- Normalisation: To obtain $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ we need to calculate the integral in the denominator:

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{L(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\Re^d} L(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

- Marginalisation: Given the joint posterior of $(\boldsymbol{\psi}, \boldsymbol{\phi}) \in \Re^m \times \Re^n$, we may be interested in the marginal posterior

$$\pi(\boldsymbol{\psi}|\boldsymbol{y}) = \int_{\Re^n} \pi(\boldsymbol{\psi}, \boldsymbol{\phi} \mid \boldsymbol{y})d\boldsymbol{\phi}$$

- Expectation: It is often of interest to obtain

$$E(g(\boldsymbol{\theta})|\boldsymbol{y}) = \int_{\Re^d} g(\boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \boldsymbol{y})d\boldsymbol{\theta}$$

Before MCMC years

- before 80's: Conjugate analysis
- early 80's: Numerical integration (e.g. Gauss rules)
- early 80's: Monte Carlo integration
- mid 80's: Laplace approximations
- late 80's: Gelfand and Smith use Gibbs sampler for simple Normal models

Monte Carlo simulation

The idea is that we can draw an IID sample $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n$ from $\pi(\boldsymbol{\theta})$ and approximate integrals by discrete sums:

$$\bar{g}(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} g(\boldsymbol{\theta}_i) \rightarrow E(g(\boldsymbol{\theta})) = \int g(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \text{ almost surely, as } n \rightarrow \infty$$

.

If the variance σ^2 of $g(\boldsymbol{\theta})$ is finite, then

$$\sqrt{n}\,(\bar{g}(\boldsymbol{\theta}) - E(g(\boldsymbol{\theta})) \Rightarrow \mathcal{N}(0, \sigma^2)$$

NOTE: $\pi(\boldsymbol{\theta})$ could be the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$.

How can we perform this *multivariate* sampling?

It would be nice if we could use direct sampling

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) = \pi(\theta^{(1)} \mid \boldsymbol{y})\pi(\theta^{(2)} \mid \theta^{(1)}, \boldsymbol{y}) \cdots \pi(\theta^{(d)} \mid \theta^{(d-1)}, \theta^{(d-2)}, \ldots, \theta^{(1)}, \boldsymbol{y})$$

but all but the last resulting univariate densities are marginalisations derived via the same integration operations that we are seeking to perform.

Maybe we can exploit the fact that *univariate* sampling is easier!

A simple example

Suppose $Y_i|\mu,\sigma^2 \sim N(\mu,\sigma^2)$ independently, $1 \leq i \leq n$. Suppose also we have prior distributions for μ and σ^2:

$$p(\mu) \sim N(\mu_0, \sigma_0^2)$$

$$p(\sigma^{-2}) \sim \text{Gamma}\,(\alpha_0, \beta_0)$$

where μ and σ are considered to be a priori independent and μ_0, σ_0^2, α_0 and β_0 are considered to be known hyperparameters. Writing $\tau = \sigma^{-2}, \tau_0 = \sigma_0^{-2}$ we can write the posterior distribution of (μ,τ):

$$\pi(\mu,\tau|\boldsymbol{y}) \propto \text{likelihood} \times \text{prior}$$

$$\propto \prod_{i=1}^{n} e^{-\frac{\tau}{2}(y_i-\mu)^2} e^{-\frac{\tau_0}{2}(\mu-\mu_0)^2} \tau^{\alpha_0+\frac{n}{2}-1} e^{-\beta_0\tau}.$$

Conditional conjugacy

We can characterise the posterior distribution in terms of its conditional distributions:

$$\pi(\mu|\tau,\boldsymbol{y}) \propto \prod_{i=1}^{n} e^{-\frac{\tau}{2}(y_i-\mu)^2} e^{-\frac{\tau_0}{2}(\mu-\mu_0)^2}$$

$$\sim N\left(\frac{\tau\Sigma y_i + \mu_0\tau_0}{n\tau+\tau_0}, \frac{1}{n\tau+\tau_0}\right)$$

and

$$\pi(\tau|\mu,\mathbf{x}) \sim \text{Gamma}\left(\alpha_0 + \frac{1}{2}, \beta_0 + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}\right).$$

This phenomenon, where the conditionals have nice simple forms, is called *conditional conjugacy*, and is common, even in extremely complex high-dimensional problems.

The Gibbs sampler

1. Initialise with $\boldsymbol{\theta} = (\theta_0^{(1)}, \ldots, \theta_0^{(d)})$.
2. Simulate $\theta_1^{(1)}$ from the conditional $\theta^{(1)}|(\theta_0^{(2)}, \ldots, \theta_0^{(d)})$.
3. Simulate $\theta_1^{(2)}$ from the conditional $\theta^{(2)}|(\theta_1^{(1)}, \theta_0^{(3)}, \ldots, \theta_0^{(d)})$.
4. ...
5. Simulate $\theta_1^{(d)}$ from the conditional $\theta^{(d)}|(\theta_1^{(1)}, \ldots, \theta_1^{(d-1)})$.
6. Iterate this procedure.

Under mild regularity conditions, convergence of the Markov chain to the stationary distribution $\pi(\theta^{(1)}, \ldots, \theta^{(d)})$ is guaranteed, so after a burn–in period, the observations $\pi(\theta_k^{(1)}, \ldots, \theta_k^{(d)}), \ldots, \pi(\theta_n^{(1)}, \ldots, \theta_n^{(d)})$ can be regarded as realisations from this distribution.

The Metropolis–Hastings algorithm

Suppose that we wish to simulate from a (multivariate) distribution $\pi(x)$. Let $q(x, y)$ be any arbitrary transistion probability (that is $q(x, y)$ is the probability density of moving to y from x), but from which simulation is straightforward. The Metropolis–Hastings algorithm is

1. Given the current position $X_n = x$, generate a 'candidate value', y^* from $q(x, y)$.
2. Calculate
$$\alpha(x, y) = \min\left\{\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, 1\right\}$$
with $y = y^*$.
3. With probability $\alpha(x, y^*)$ accept the candidate value and set $X_{n+1} = y^*$; otherwise reject and set $X_{n+1} = x$.
4. Iterate

The independence sampler

Choose

$$q(\boldsymbol{x}, \boldsymbol{y}) = q(\boldsymbol{y}) \ .$$

$$\alpha(\boldsymbol{x}, \boldsymbol{y}) = \min \left\{ \frac{w(\boldsymbol{y})}{w(\boldsymbol{x})}, 1 \right\}$$

where

$$w(\boldsymbol{x}) = \frac{\pi(\boldsymbol{x})}{q(\boldsymbol{x})} \ .$$

If $q \propto \pi$, w is constant and the algorithm reduces, as we would expect, to IID sampling from π.

Symmetric random walk Metropolis algorithm

$q(\boldsymbol{x}, \boldsymbol{y}) = q_{rw}(\boldsymbol{y} - \boldsymbol{x})$ where $(q_{rw}(\boldsymbol{x}) = q_{rw}(-\boldsymbol{x})$ for all $\boldsymbol{x})$.

$$\alpha(\boldsymbol{x}, \boldsymbol{y}) = \min\left\{\frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})}, 1\right\} \ .$$

Accept all moves which increase π ("uphill moves"), but reject moves which decrease π ("downhill moves").

Multiplicative random walk algorithms

From current state X_n we propose a move to $X_{n+1} = X_n e^N$ where N is drawn from a symmetric density q_{rw}. By a simple Jacobian transformation of π, the acceptance probability can be easily shown to be

$$\alpha(X_n, X_{n+1}) = \min\left\{1, \frac{\pi(X_{n+1})X_{n+1}}{\pi(X_n)X_n}\right\}.$$

Langevin algorithms

Use as a proposal

$$q(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{(2\pi\sigma)^{d/2}} \exp\left\{ \frac{-(\boldsymbol{y} - \boldsymbol{x} - \sigma^2 \nabla \log \pi(\boldsymbol{x})/2)^2}{2\sigma^2} \right\} ,$$

for some suitable scaling parameter σ. Rather than be centered at the current state, the proposal center is adjusted according to information about where the target density is likely to be larger.

Auxiliary variable methods

Suppose $\boldsymbol{X} \sim \pi$ and $\mathbf{U} \in \Re^l$ is defined in terms as

$$\mathbf{U}|\boldsymbol{X} = \boldsymbol{x} \sim h(\cdot|\boldsymbol{x})$$

$(\boldsymbol{X}, \mathbf{U})$ is now an $\Re^{d+l}$ random variable with joint density

$$\pi_E(\boldsymbol{x}, \boldsymbol{\beta}) \propto \pi(\boldsymbol{x}) h(\boldsymbol{\beta}|\boldsymbol{x}) \ .$$

Suppose that

$$\pi(\boldsymbol{x}) = f_0(\boldsymbol{x}) \prod_{i=1}^{l} f_i(\boldsymbol{x}) \ .$$

(Think as prior $\times$ likelihood). Can we obtain π_E?

Auxiliary variables -continued

Define $\boldsymbol{U}$ as l conditionally independent components with

$$U^{(i)}|\boldsymbol{X} \sim U(0, f_i(\boldsymbol{X})) \ . \tag{1}$$

Letting $S(\boldsymbol{\beta}) = \{\boldsymbol{x};\ \ u^{(i)} \leq f_i(\boldsymbol{x}), 1 \leq i \leq l\}$, then

$$\pi_E(\boldsymbol{x}, \boldsymbol{\beta}) \propto f_0(\boldsymbol{x}) \mathbf{1}_{\boldsymbol{x} \in S(\boldsymbol{\beta})}$$

Often we take f_0 to be constant, and in this case, π_E is the uniform density on $\cap_{i=1}^{l} \{\boldsymbol{x};\ u^{(i)} \leq f_i(\boldsymbol{x})\}$.

The slice sampler carries out a Gibbs sampler on π_E. We set $\boldsymbol{W}^{(1)} = \boldsymbol{U}$ and $\boldsymbol{W}^{(2)} = \boldsymbol{X}$. The conditional distribution of $\boldsymbol{U}|\boldsymbol{X}$ is specified by (1) and the requirement that the $U^{(i)}$s be conditionally independent. Sampling from $\boldsymbol{X}|\boldsymbol{U}$ involves simulating from a density proportional to the truncated function $f_0(\boldsymbol{x}) \mathbf{1}_{\boldsymbol{x} \in S(\boldsymbol{\beta})}$. Even where f_0 is constant, this simulation can be difficult, and as such this limits the applicability of the slice sampler.

More on MCMC

- Metropolis-within-Gibbs
- Blocking
- Marginalisation
- reparameterisation
- Use classical estimates in random walk Metropolis

Implementation and output analysis

- single chain vs several chains
- convergence
- mixing
- diagnostic plots (Autocorrelation, crosscorrelation)
- Monte Carlo error (effective sample size or integrated auto-correlation time)

Data augmentation

Data augmentation is a technique which can be thought of as a special case of the Gibbs sampler. It can be applied equally well to the following two situations.

1. One of the main problems that afflicts statistical computing is that in the real world, some of the data could be missing. In theory the distribution of the observed data can be obtained by integrating out the missing data. However this is frequently difficult if not impossible to do.
2. The likelihood of the data might not tractable for some reason (for instance a simple conjugate prior might not be available), but conditional on a collection of unobserved data the likelihood becomes easy to handle.

Missing data

Let $\boldsymbol{y}_{obs}$ denote the observed data, $\boldsymbol{y}_{mis}$ denote the missing data, and θ be the unknown parameter with prior $p(\theta)$. It is often the case that $\pi(\boldsymbol{y}_{obs} \mid \boldsymbol{y}_{mis}, \theta)$ is available but $\pi(\boldsymbol{y}_{obs}|\theta)$ is not. As a result of this, we consider $\boldsymbol{y}_{mis}$ a further set of parameters by noting that

$$\pi(\theta, \boldsymbol{y}_{mis}|\boldsymbol{y}_{obs}) \propto \pi(\boldsymbol{y}_{obs} \mid \boldsymbol{y}_{mis}, \theta)\pi(\theta).$$

Data augmentation proceeds by carrying out Gibbs sampling to successively sample from θ and $\boldsymbol{Y}_{mis}$ to produce a sample from this joint distribution. The marginal distribution of θ is therefore the posterior distribution of interest.

Prediction

The predictive distribution of a future observation z given past observations $\boldsymbol{y}$ is defined as

$$p(z|\boldsymbol{y}) = \int p(z|\theta)p(\theta|\boldsymbol{y})d\theta$$

which can be viewed as the expected likelihood, $p(z|\theta)$, under the uncertainty in θ contained in the posterior distribution $p(\theta|\boldsymbol{y})$. Hence, given a sampled sequence of realisations $\theta_1, \ldots, \theta_n$ from this posterior, possibly obtained by Gibbs sampling, we can estimate

$$p(z|\boldsymbol{y}) \approx \frac{1}{n}\sum_{i=1}^{n} p(z|\theta_i).$$

Bayesian model choice

$$\pi(m|\boldsymbol{y}) = \frac{p(m)L(\boldsymbol{y}|m)}{\sum_{m\in M} p(m)L(\boldsymbol{y}|m)}, \qquad m \in M$$

$$L(\boldsymbol{y}|m) = \int L(\boldsymbol{y}|m,\boldsymbol{\theta}_m)p(\boldsymbol{\theta}_m|m)d\boldsymbol{\theta}_m$$

Interested in $\pi(m,\boldsymbol{\theta}_m|\boldsymbol{y})$.

The natural parameter space for $(m,\boldsymbol{\theta}_m)$ is

$$\Theta = \bigcup_{m\in M} \{m\} \times \Theta_m.$$

Random walk and independent Metropolis on Θ

Current value $(m, \boldsymbol{\theta}_m)$, proposal $(m', \boldsymbol{\theta}'_{m'})$ from $q(m', \boldsymbol{\theta}'_{m'}|m, \boldsymbol{\theta}_m)$.

$$\alpha = \min\left(1, \frac{L(\boldsymbol{y}|m', \boldsymbol{\theta}'_{m'})\pi(\boldsymbol{\theta}'_{m'}|m')p(m')q(m, \boldsymbol{\theta}_m|m', \boldsymbol{\theta}'_{m'})}{L(\boldsymbol{y}|m, \boldsymbol{\theta}_m)\pi(\boldsymbol{\theta}_m|m)p(m)q(m', \boldsymbol{\theta}'_{m'}|m, \boldsymbol{\theta}_m)}\right).$$

$$q(m', \boldsymbol{\theta}'_{m'}|m, \boldsymbol{\theta}_m) = q(m'|m, \boldsymbol{\theta}_m)q(\boldsymbol{\theta}'_{m'}|m', m, \boldsymbol{\theta}_m)$$

Reversible jump

- Current state is $(m, \boldsymbol{\theta}_m)$, where $\boldsymbol{\theta}_m$ has dimension $d(\boldsymbol{\theta}_m)$.
- Propose a new model m' with probability $j(m, m')$.
- Generate $\boldsymbol{u}$ (which can be of lower dimension than $\boldsymbol{\theta}_{m'}$) from a specified proposal density $q(\boldsymbol{u}|\boldsymbol{\theta}_m, m, m')$.
- Set $(\boldsymbol{\theta}'_{m'}, \boldsymbol{u}') = g_{m,m'}(\boldsymbol{\theta}_m, \boldsymbol{u})$ where $g_{m,m'}$ is a specified invertible function. Hence $d(\boldsymbol{\theta}_m) + d(\boldsymbol{u}) = d(\boldsymbol{\theta}_{m'}) + d(\boldsymbol{u}')$. Note that $g_{m',m} = g^{-1}_{m,m'}$.
- Accept the proposed move to model m' with probability

$$\alpha = \min\left(1, \frac{L(\boldsymbol{y}|m', \boldsymbol{\theta}'_{m'})p(\boldsymbol{\theta}'_{m'}|m')p(m')j(m', m), q(\boldsymbol{u}'|\boldsymbol{\theta}_m, m', m)}{L(\boldsymbol{y}|m, \boldsymbol{\theta}_m)p(\boldsymbol{\theta}_m|m)p(m)j(m, m')q(\boldsymbol{u}|\boldsymbol{\theta}_{m'}, m, m')} \left|\frac{\partial g_{m,m'}(\boldsymbol{\theta}_m, \boldsymbol{u})}{\partial(\boldsymbol{\theta}_m, \boldsymbol{u})}\right|\right)$$

A simple example

- Switch between $(1, \theta)$ and $(2, \theta_1, \theta_2)$
- To go from $(2, \theta_1, \theta_2)$ to $(1, \theta)$, set $\theta = (\theta_1 + \theta_2)/2)$
- To go from $(1, \theta)$ to $(2, \theta_1, \theta_2)$ generate a r.v. u and set $\theta_1 = \theta - u$ and $\theta_2 = \theta + u$.
- $g_{1,2}(\theta, u) = (\theta - u, \theta + u)$
- $g_{2,1}(\theta_1, \theta_2) = ((\theta_1 + \theta_2)/2, \theta_2)$
- Jacobians are

$$\frac{\partial g_{1,2}(\theta, u)}{\partial(\theta, u)} = 2, \qquad \frac{\partial g_{2,1}(\theta_1, \theta_2)}{\partial(\theta_1, \theta_2)} = \frac{1}{2}$$

Reversible Jump Variations

- If all parameters of the proposed model are generated directly from a proposal distribution, then $(\boldsymbol{\theta}'_{m'}, \boldsymbol{u}') = (\boldsymbol{u}, \boldsymbol{\theta}_m)$ with $d(\boldsymbol{\theta}_m) = d(\boldsymbol{u}')$ and $d(\boldsymbol{\theta}_{m'}) = d(\boldsymbol{u})$, and the jacobian is one (Independence sampler)
- With the same proposals, but where the function $(\boldsymbol{\theta}'_{m'}, \boldsymbol{u}') = g_{m,m'}(\boldsymbol{u}, \boldsymbol{\theta}_m)$ is not the identity then we have a more general Metropolis-Hastings algorithm where $\boldsymbol{\theta}'_{m'}$ is allowed to depend on $\boldsymbol{\theta}_m$. If $m' = m$, then the move is a standard Metropolis-Hastings step.
- Use proposal distributions of lower dimension than $d(\boldsymbol{\theta}'_{m'})$: If model m is nested in m', then $g_{m,m'}$ may be the identity function such that $d(\boldsymbol{u}') = 0$ and $\boldsymbol{\theta}'_{m'} = g_{m,m'}(\boldsymbol{\theta}_m, \boldsymbol{u})$. In the reverse move the model parameters are proposed deterministically.

Carlin and Chib's Gibbs sampler

The parameter space for $(m, \boldsymbol{\theta}_k : k \in M)$ is $M \times \prod_{m \in M} \Theta_m$. Therefore, a prior distribution for $(m, \boldsymbol{\theta}_k : k \in M)$ is no longer completely specified by $p(m)$ and $p(\boldsymbol{\theta}_m|m)$, so Carlin and Chib proposed the use of pseudopriors or linking densities $p(\boldsymbol{\theta}_k|m \neq k)$, $k \in M$.

$$\pi(\boldsymbol{\theta}_k|\boldsymbol{y},\ \{\boldsymbol{\theta}_l : l \neq k\}, m) \propto \begin{cases} L(\boldsymbol{y}|\boldsymbol{\theta}_m, m)p(\boldsymbol{\theta}_m|m) & k = m \\ p(\boldsymbol{\theta}_k|k \neq m) & k \neq m \end{cases}$$

and

$$\pi(m|\{\boldsymbol{\theta}_k : k \in M\}, \boldsymbol{y}) = \frac{A_m}{\sum_{k \in M}\ A_k}$$

where

$$A_m = L(\boldsymbol{y}|\boldsymbol{\theta}_m, m) \prod_{l \in M} [p(\boldsymbol{\theta}_l|m)]p(m).$$

For $k \neq m$ optimise with $f(\boldsymbol{\theta}_k|m) \approx f(\boldsymbol{\theta}_k|k, \boldsymbol{y})$.

Metropolised Carlin and Chib

$$\begin{aligned}\alpha &= \min\left(1, \frac{A_{m'}j(m',m)}{A_m j(m,m')}\right) \\ &= \min\left(1, \frac{L(\boldsymbol{y}|\boldsymbol{\theta}_{m'},m')p(\boldsymbol{\theta}_{m'}|m')p(\boldsymbol{\theta}_m|m')p(m')j(m',m)}{L(\boldsymbol{y}|\boldsymbol{\theta}_m,m)p(\boldsymbol{\theta}_m|m)p(\boldsymbol{\theta}_{m'}|m)p(m)j(m,m')}\right) \qquad (2)\end{aligned}$$

as all other pseudopriors cancel.

- Propose a new model m' with probability $j(m,m')$.
- Generate $\boldsymbol{\theta}_m$ from the posterior $\pi(\boldsymbol{\theta}_m|\boldsymbol{y},m)$.
- Generate $\boldsymbol{\theta}_{m'}$ from the pseudoprior $p(\boldsymbol{\theta}_{m'}|m \neq m')$.
- Accept the proposed move to model m' with probability α given by (2).

Comment

All these approaches are less flexible than general reversible jump, as they all require a proposal distribution of full dimension $d(\boldsymbol{\theta}'_{m'})$ when model m' is proposed. Reversible jump on the other hand allows the model parameters of the proposed model to depend on the parameters of the current model in a totally general way through the function $g_{m',m}$. In particular, the dimension of the proposal distribution may be much less than the dimension of the proposed model.

Example

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2), \quad i = 1, \ldots, n, \quad m = 1$$

$$Y_i \sim N(\gamma + \delta z_i, \tau^2), \quad i = 1, \ldots, n \quad , m = 2$$

Independence sampler requires a proposal for (γ, δ, τ) independent of the current values of (α, β, σ).

How can we apply Carlin and Chib's approach? reversible jump?

Using Posterior Distributions as Proposals

Suppose that, for each m, the posterior density $f(\boldsymbol{\theta}_m|m,\boldsymbol{y})$ is available, including the normalising constant which is the marginal likelihood $f(\boldsymbol{y}|m)$. If this distribution is used as a pseudoprior then the acceptance probability in the Metropolised Carlin+Chib algorithm is given by

$$\begin{aligned}\alpha &= min\left(1, \frac{f(\boldsymbol{y}|m',\boldsymbol{\theta}'_{m'})f(\boldsymbol{\theta}'_{m'}|m')f(m')j(m',m)f(\boldsymbol{\theta}_m|m,\boldsymbol{y})}{f(\boldsymbol{y}|m,\boldsymbol{\theta}_m)f(\boldsymbol{\theta}_m|m)f(m)j(m,m')f(\boldsymbol{\theta}'_{m'}|m',\boldsymbol{y})}\right)\\ &= min\left(1, B_{m'm}\frac{f(m')j(m',m)}{f(m)j(m,m')}\right)\end{aligned}$$

where $B_{m'm}$ is the Bayes factor of model m' against model m. In practice, we cannot usually calculate $B_{m'm}$. In the special case where models are decomposable graphical models, Madigan and York (1995) used exactly this approach, which they called MC^3. Here there is no need to generate the model parameters $\boldsymbol{\theta}_m$ as part of the Markov chain. These can be generated separately from the known posterior distributions $f(\boldsymbol{\theta}_m|m,\boldsymbol{y})$ if required.

Variable Selection

$$\boldsymbol{\eta} = \sum_{i=1}^{p} \gamma_i \boldsymbol{X}_i \boldsymbol{\theta}_i$$

where $\boldsymbol{X}_i$ is the design matrix and $\boldsymbol{\theta}_i$ the parameter vector related to the ith term.

Substitute $\boldsymbol{\gamma}$ for model indicator m.

In some cases, it is sensible to set $f(\gamma_i|\boldsymbol{\gamma}_{\setminus i}) = f(\gamma_i)$ (where the subscript $\setminus i$ denotes all elements of a vector except the ith) , whereas in other cases (e.g. hierarchical or graphical log-linear models) it is required that $f(\gamma_i|\boldsymbol{\gamma}_{\setminus i})$ depends on $\gamma_{\setminus i}$.

Notes

In Carlin and Chib's method, at each iteration of the Gibbs sampler, all parameters of all models are generated from either posterior distribution or pseudoprior, and the model selection step allows a simultaneous change of all γ_is. In variable selection, it is possible to generate an observation of $\boldsymbol{\gamma}$ followed by a generation of all $\boldsymbol{\theta}_k$ from posterior distributions or pseudopriors (smaller computational burden).

Moves between models m ($\boldsymbol{\gamma}$) and m' ($\boldsymbol{\gamma'}$) may be based on a Metropolis step, as in Metropolised Carlin and Chib's method. Then the pseudopriors may be thought of as part of the proposal density for parameters which are present in one model but not in the other.

A drawback with the variable selection approaches is that parameters which are 'common' to both models remain unchanged, and therefore the procedure will not be efficient unless posterior distributions for such parameters are similar under both models.

Choice of proposals

- Local moves: We hope to make use of the current parameter values to propose plausible values for the parameters of the proposed model (Gibbs variable selection, hierarchical loglinear models). Use a pilot chain for the saturated model.
- Global moves: pilot chain for each model (expensive), normal distribution centred at the maximum likelihood estimate for $\boldsymbol{\theta}_m$ with variance equal to the asymptotic variance of the mle (expensive). See other talks later this week.

Chib's marginal likelihood

$$L(\boldsymbol{y}|m) = \frac{L(\boldsymbol{y}|\boldsymbol{\theta}, m)p(\boldsymbol{\theta}|m)}{\pi(\boldsymbol{\theta}|\boldsymbol{y}, m)}$$

Assume $\boldsymbol{\theta} = (\theta^{(1)}, \theta^{(2)}, \theta^{(3)})$ and a sample from Gibbs sampling output $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n$ is available. Then the joint posterior density at the point $(\theta^{(1)*}, \theta^{(2)*}, \theta^{(3)*})$ is given by

$$\pi(\theta^{(1)*}, \theta^{(2)*}, \theta^{(3)*} \mid \boldsymbol{y}) = \pi(\theta^{(1)*} \mid \boldsymbol{y})\pi(\theta^{(2)*} \mid \theta^{(1)*}, \boldsymbol{y})\pi(\theta^{(3)*} \mid \theta^{(1)*}, \theta^{(2)*}, \boldsymbol{y}).$$

$$\pi(\theta^{(1)*} \mid \boldsymbol{y}) \simeq n^{-1} \sum_{i=1}^{n} \pi(\theta^{(1)*} \mid \theta_i^{(2)}, \theta_i^{(3)}, y)$$

$$\pi(\theta^{(2)*} \mid \theta^{(1)*}, \boldsymbol{y}) \simeq n^{-1} \sum_{i=1}^{n} \pi(\theta^{(2)*} \mid \theta^{(1)*}, \theta_i^{(3)}, y)$$

with $\theta_i^{(3)}$ drawn from $\pi(\theta^{(3)} \mid \theta^{(1)*}, \boldsymbol{y})$.

Bridge sampling (Meng and Wong)

Ratio of two marginal likelihoods $L(\boldsymbol{y}|m)$ and $L(\boldsymbol{y}|m')$.

$$c_m(\boldsymbol{\theta}) = L(\boldsymbol{y}|\boldsymbol{\theta}_m, m)p(\boldsymbol{\theta}_m|m), c_{m'}(\boldsymbol{\theta}) = L(\boldsymbol{y}|\boldsymbol{\theta}_{m'}, m')p(\boldsymbol{\theta}_{m'}|m')$$

$$\frac{L(\boldsymbol{y}|m)}{L(\boldsymbol{y}|m')} = E_{m'}\left\{\frac{c_m(\boldsymbol{\theta})}{c_{m'}(\boldsymbol{\theta})}\right\} \quad \boldsymbol{\Theta}_m \subset \boldsymbol{\Theta}_{m'}$$

$$\frac{L(\boldsymbol{y}|m)}{L(\boldsymbol{y}|m')} = \frac{E_{m'}\left\{c_m(\boldsymbol{\theta})g(\boldsymbol{\theta})\right\}}{E_m\left\{c_{m'}(\boldsymbol{\theta})g(\boldsymbol{\theta})\right\}}$$

where $g(\boldsymbol{\theta})$ is a function defined on $\boldsymbol{\Theta}_m \bigcap \boldsymbol{\Theta}_{m'}$ such that

$$0 < \left| \int_{\boldsymbol{\Theta}_m \bigcap \boldsymbol{\Theta}_{m'}} c_m(\boldsymbol{\theta})c_{m'}(\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta} \right| < \infty.$$

bridge sampling (continued)

Assume that a "bridge" density $\pi(\boldsymbol{\theta})$ lies 'between' $c_m(\boldsymbol{\theta})$ and $c_{m'}(\boldsymbol{\theta})$. Then $g(\boldsymbol{\theta})$ can be written as

$$g(\boldsymbol{\theta}) = \pi_b(\boldsymbol{\theta})/[c_m(\boldsymbol{\theta})c_{m'}(\boldsymbol{\theta})]$$

so that

$$\frac{L(\boldsymbol{y}|m)}{L(\boldsymbol{y}|m')} = \frac{E_{m'}\left\{\pi_b(\boldsymbol{\theta})/c_{m'}(\boldsymbol{\theta})\right\}}{E_m\left\{\pi_b(\boldsymbol{\theta})/c_m(\boldsymbol{\theta})\right\}}.$$

Message: the ratio of the two marginal likelihoods is estimated by using samples from both posterior densities of the two models with the density $\pi_b(\boldsymbol{\theta})$ served as a 'bridge' between them. Since the idea is based on the fact that this bridge 'lives' in $\boldsymbol{\Theta}_m \bigcap \boldsymbol{\Theta}_{m'}$ this method would not work if $\boldsymbol{\Theta}_m \bigcap \boldsymbol{\theta}_{m'} = \emptyset$ and it would work poorly if this intersection is 'small'.

Path sampling

Use a series of bridges to compare models without any restrictions on $\boldsymbol{\Theta}_m$ and $\boldsymbol{\Theta}_{m'}$.

Consider the family of posterior densities indexed by the parameter $z \in [0,1]$:

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}, z) = \frac{\pi(\boldsymbol{\theta}, \boldsymbol{y} \mid z)}{L(\boldsymbol{y} \mid z)}$$

and think of it as a density producing $L(\boldsymbol{y} \mid m)$ if $z = 1$ and $L(\boldsymbol{y} \mid m')$ if $z = 0$. First note that by assuming interchange of differentiation and integration, since $L(\boldsymbol{y} \mid z)$ is just the marginal density of $\pi(\boldsymbol{\theta}, \boldsymbol{y} \mid z)$, we obtain

$$\frac{d}{dz} \log L(\boldsymbol{y} \mid z) = \int L^{-1}(\boldsymbol{y} \mid m) \frac{d}{dz} \pi(\boldsymbol{\theta}, \boldsymbol{y} \mid z) d\boldsymbol{\theta} = E_{\boldsymbol{\theta} \mid \boldsymbol{y}, z} \left[\frac{d}{dz} \log \pi(\boldsymbol{\theta}, \boldsymbol{y} \mid z) \right].$$

If we treat z as a parameter with prior density $p(z)$ independent of $\boldsymbol{\theta}$, then

$\frac{d}{dz}\log\pi(\boldsymbol{\theta},\boldsymbol{y}\mid z)=\frac{d}{dz}\log L(\boldsymbol{y}\mid\boldsymbol{\theta},z)$ and

$$\log\left(\frac{L(\boldsymbol{y}\mid z=1)}{L(\boldsymbol{y}\mid z=0)}\right)=\int_0^1 E_{\boldsymbol{\theta}|\boldsymbol{y},z}\left[\frac{d}{dz}\log L(\boldsymbol{y}\mid\boldsymbol{\theta},z)\right]dz=E_{\boldsymbol{\theta},z|\boldsymbol{y}}\left[\frac{\frac{d}{dz}\log L(\boldsymbol{y}\mid\boldsymbol{\theta},z)}{p(z)}\right].$$

Calculate the integral over z with a trapezoidal rule based on a equidistant set of points $z_0=0,z_1,z_2,\ldots,z_k=1$ over $[0,1]$. The resulting estimator for samples $\boldsymbol{\theta}_{1,z_i},\boldsymbol{\theta}_{2,z_i},\ldots,\boldsymbol{\theta}_{n,z_i}$ from $\pi(\boldsymbol{\theta}\mid\boldsymbol{y},z_i)$ for $i=1,\ldots,k$ is

$$\log\left(\frac{L(\boldsymbol{y}\mid z=1)}{L(\boldsymbol{y}\mid z=0)}\right)=$$

$$\frac{1}{2}\sum_{i=0}^{k-1}(z_{i+1}-z_i)\left[\frac{1}{n}\sum_{t=1}^{n}\left(\left[\frac{d}{dz}\log\pi(\boldsymbol{y}\mid\boldsymbol{\theta}_t,z)\right]_{z_{i+1}}+\left[\frac{d}{dz}\log\pi(\boldsymbol{y}\mid\boldsymbol{\theta}_t,z)\right]_{z_i}\right)\right].$$

Model choice concluding remarks

- Reversible jump (or, indeed, a more general Metropolis-Hastings algorithm) seems to us the most promising algorithm to search in
  $\Theta = \bigcup_{m \in M} \{m\} \times \Theta_m$ for high regions of $\pi(m, \boldsymbol{\theta}_m)$
- Good proposals is the key element for constructing efficient algorithms; for recent advances in this direction see the technical report of Brooks, Giudici and Roberts in the *MCMC preprint service*.

ARCH and SV models

Data $y_t,\ t = 1, \ldots, T$

$$y_t = \mu + \epsilon_t, \qquad \epsilon_t = z_t \sigma_t$$

GARCH(1,1) models

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \quad t = 1, ..., T$$

Stochastic volatility models

$$\log \sigma_t = \alpha_0 + \alpha_1 \log \sigma_{t-1} + \alpha_2 \eta_t, \qquad t = 1, ..., T$$

Exercise: MCMC GARCH? MCMC SV? Model choice?

Multivariate models

$$\mathbf{y}_t = (y_{1,t}, ..., y_{N,t}), \quad \mathbf{y}_t = \mu + \varepsilon_t, \quad \varepsilon_t | \Phi_{t-1} \sim N_N(\mathbf{0}, \mathbf{\Sigma}_t), \quad t = 1, ..., T.$$

Kraft and Engle (1982)

$$vech(\mathbf{\Sigma}_t) = \mathbf{C} + \mathbf{A} vech\left(\varepsilon_{t-1}\varepsilon_{t-1}^{'}\right)$$

Bollerslev, Engle and Wooldridge (1988)

$$vech(\mathbf{\Sigma}_t) = \mathbf{C} + \mathbf{A} vech\left(\varepsilon_{t-1}\varepsilon_{t-1}^{'}\right) + \mathbf{B} vech(\mathbf{\Sigma}_{t-1})$$

Engle and Kroner (1995) (BEKK representation)

$$\mathbf{\Sigma}_t = \mathbf{C}'\mathbf{C} + \mathbf{A}'\varepsilon_{t-1}\varepsilon_{t-1}^{'}\mathbf{A} + \mathbf{B}'\mathbf{\Sigma}_{t-1}\mathbf{B}$$

Bollerslev(1990), Jeantheau (1998): constant conditional correlation.

$$\sigma_{i,t}^2 = a_{0i} + a_{i1}\varepsilon_{i,t-1}^2 + \beta_{i1}\sigma_{i,t-1}^2, \ i = 1, \ldots, N$$

$$\sigma_{ij,t} = \rho_{ij}\left(\sigma_{i,t}^2\sigma_{j,t}^2\right)^{1/2}, \ -1 \leq \rho_{ij} \leq 1,$$

Factor models

Diebold and Nerlove (1989), King, Sentana and Wadhwani (1994),...

Current state of the art: Chib, Nardari and Shephard (2001), Aquilar and West (2000)

$$\mathbf{y}_t = \lambda f_t + \epsilon_t$$

$(\mathbf{f}_t, \epsilon_t)^T \sim N_{k+N}(0, V_t)$ with $V_t = diag\{\exp(v_{1t}), \exp(v_{2t}), \ldots, \exp(v_{k+N,t})\}$ and

$$v_{it} = \alpha_0 + \alpha_1 v_{i,t-1} + \alpha_2 \eta_t, \quad t = 1, ..., T$$

MCMC applications up to 40 stocks: 16 hours including model choice (number of factors).

Latent factor GARCH

King, Sentana & Wadhwani 1994

$(\mathbf{f}_t, \epsilon_t)^T \sim N_{k+N}(0, V_t)$ with

$$V_t = diag\{\sigma_{1t}^2, \sigma_{2t}^2, \ldots, \sigma_{kt}^2, \gamma_1, \gamma_2, \ldots, \gamma_N\}$$

$$\sigma_{1t}^2 = 1 - a - b + a f_{t-1}^2 + b\sigma_t^2$$

Closely connecting to APT.

Slice Gibbs: Steps for latent factor GARCH

1. $\gamma_i|\cdot \sim \mathbf{IG}(\cdot,\cdot)$, for all $i = 1, ..., 4$.
2. $\lambda_i|\cdot \sim \mathbf{N}(\cdot,\cdot)$, for all $i = 1, ..., 4$.
3. $\sigma_0^2|\cdot \sim \mathbf{U}(\cdot,\cdot)$, for all $i = 1, ..., 4$.
4. $a|\cdot \sim \mathbf{U}(\cdot,\cdot)$.
5. $b|\cdot \sim \mathbf{U}(\cdot,\cdot)$.
6. $\zeta_t|\cdot \sim \mathbf{U}(\cdot,\cdot)$, for all $t = 1, ..., T$.
7. $z_t|\cdot \sim \mathbf{U}(\cdot,\cdot)$, for all $t = 1, ..., T$.
8. $f_0|\cdot \sim N(\cdot,\cdot)\,\mathbf{I}(\cdot,\cdot)$.
9. $f_t|\cdot \sim N(\cdot,\cdot)\,\mathbf{I}(\cdot,\cdot)$, for $t = 1, ..., T-1$.
10. $f_T|\cdot \sim N(\cdot,\cdot)\,\mathbf{I}(\cdot,\cdot)$.

A new multivariate GARCH model

$$\mathbf{y}_t = \mu + \varepsilon_t, \quad \varepsilon_t = W\mathbf{X}_t$$

$$\mathbf{X}_t \mid \Phi_{t-1} \sim N_N(\mathbf{0}, \Sigma_t)$$

$$\Sigma_t = diag\left(\sigma_{1t}^2, ..., \sigma_{Nt}^2\right)$$

with

$$\sigma_{i,t}^2 = a_i + b_i x_{i,t-1}^2 + c_i \sigma_{i,t-1}^2, \quad i = 1, ..., N, \quad t = 1, ..., T$$

Simple case: $b_i = b$ and $c_i = c$.

$$\begin{aligned} Var(y_t|\Phi_{t-1}) = H_t &= W\Sigma_t W' = W\Sigma_t^{1/2}\Sigma_t^{1/2}W' \\ &= \left(W\Sigma_t^{1/2}\right)\left(W\Sigma_t^{1/2}\right)' = LL' \end{aligned}$$

This decomposition is unique if W is lower triangular with positive diagonal elements. We take all diagonal elements equal to one.

H_t is always positive definite, and the number of parameters is $2N + 2 + \frac{N(N-1)}{2}$.

Conditional covariance matrix

$$H_t = \begin{bmatrix} \sigma_{1,t}^2 & w_{21}\sigma_{1,t}^2 & \cdots & w_{N1}\sigma_{1,t}^2 \\ w_{21}\sigma_{1,t}^2 & w_{21}^2\sigma_{1,t}^2 + \sigma_{2,t}^2 & \cdots & \sum_{i=1}^{2} w_{2i}w_{Ni}\sigma_{i,t}^2 \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1}\sigma_{1,t}^2 & \sum_{i=1}^{2} w_{Ni}w_{2i}\sigma_{i,t}^2 & \cdots & \sum_{i=1}^{N-1} w_{N-1,i}^2\sigma_{i,t}^2 + \sigma_{N,t}^2 \end{bmatrix}$$

- Unconditional covariance matrix

$$E(H) = \begin{bmatrix} \frac{\alpha_1}{1-b-g} & \frac{w_{21}\alpha_1}{1-b-g} & \frac{w_{31}\alpha_1}{1-b-g} & \cdots & \frac{w_{N1}\alpha_1}{1-b-g} \\ \frac{w_{21}\alpha_1}{1-b-g} & \frac{\sum_{i=1}^{2} w_{2i}^2\alpha_i}{1-b-g} & \frac{\sum_{i=1}^{2} w_{3i}w_{2i}\alpha_i}{1-b-g} & \cdots & \frac{\sum_{i=1}^{2} w_{Ni}w_{2i}\alpha_i}{1-b-g} \\ \frac{w_{31}\alpha_1}{1-b-g} & \frac{\sum_{i=1}^{2} w_{3i}w_{2i}\alpha_i}{1-b-g} & \frac{\sum_{i=1}^{3} w_{3i}^2\alpha_i}{1-b-g} & \cdots & \frac{\sum_{i=1}^{3} w_{Ni}w_{3i}\alpha_i}{1-b-g} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{w_{N1}\alpha_1}{1-b-g} & \frac{\sum_{i=1}^{2} w_{Ni}w_{2i}\alpha_i}{1-b-g} & \frac{\sum_{i=1}^{3} w_{Ni}w_{3i}\alpha_i}{1-b-g} & \cdots & \frac{\sum_{i=1}^{N} w_{Ni}^2\alpha_i}{1-b-g} \end{bmatrix}$$

- Stationarity Conditions: $b + c < 1$

Inference

- Expected Fisher information matrix available analytically.
- Use 3 blocks of parameters: means, W matrix, GARCH parameters.
- Maximum likelihood estimates available easily (Fiorentini, Galzolari and Panattoni, 1996: Analytic derivatives and the computation of GARCH estimates).
- For MCMC use a blocking sampling scheme.
- convert all parameters to "near" normality
- Multivariate Normal proposal densities $N\left(\theta_i^{t-1}, \ \ c\widehat{\Sigma}_{\theta_i}\right)$ for i block

Ordering of vectors of observations

- Estimate the marginal likelihood by using Laplace Approximation

- Estimate the posterior model probabilities by using Markov Chain Monte Carlo Composition (MC^3). Define proper model neighborhoods and use delayed rejection algorithm (Tierney and Mira, 1999) to deal with multimodalities
- model averaging

Delayed rejection

Suppose that the current state of the chain is model m. Then, at the first stage, model $m^{'}$ is drawn from $q\left(m \to m^{'}\right)$ and accepted with probability

$$\min\left\{1, \frac{\pi\left(\mathbf{y}|m^{'}\right)}{\pi\left(\mathbf{y}|m\right)}\right\}.$$

If the candidate model $m^{'}$ is rejected, a new candidate model $m^{''}$ is proposed from $q\left(m^{'} \to m^{''}\right)$ at the second stage.

The probability of acceptance at this stage is given by

$$\min\left\{1, \frac{\max\left\{0, \left[\pi\left(\mathbf{y}|m^{''}\right) - \pi\left(\mathbf{y}|m^{'}\right)\right]\right\}}{\pi\left(\mathbf{y}|m\right) - \pi\left(\mathbf{y}|m^{'}\right)}\right\}.$$

Model determination

Predictive density: $[Y_{T+1} \mid \mathbf{y}] = \int [Y_{T+1} \mid \theta]\,[\theta \mid \mathbf{y}]\,d\theta$

- Check $[Y_{T+1} \mid \mathbf{y}]$ against the real value y_{T+1}

(Gelfand et al 1992, Shephard and Pitt 1999)

1. Obtain (through MCMC)observations from $[\theta \mid \mathbf{y}]$: θ_s, $s = 1, ..., B$
2. Estimate $[Y_{T+1} \mid \mathbf{y}]$ by

$$\widehat{[Y_{T+1} \mid \mathbf{y}]} = B^{-1} \sum_{s=1}^{B} [y_{T+1} \mid \theta_s, \mathbf{y}]$$

3. Use M one-step-ahead predictions and compare models M_i and M_j via

$$\log \left\{ \frac{\prod^{(M_i)} \widehat{[Y_{t+1} \mid \mathbf{y}]}}{\prod^{(M_j)} \widehat{[Y_{t+1} \mid \mathbf{y}]}} \right\}$$

where the product is over $t = T, ..., T + M - 1$