

Example 4.4

- Data: $X_i|\alpha, \lambda \sim Ga(\alpha, \lambda)$, $i = 1, 2, \dots, n$ (independent), where the index α and scale parameter λ are unknown
- Prior: $\alpha \sim Ga(a, b)$ and $\lambda \sim Ga(c, d)$, independent
- Determine the posterior density for $(\alpha, \lambda)^T$ and hence the posterior conditional densities for $\alpha|\lambda$ and $\lambda|\alpha$

Solution

...

Problem

- Although the FCD for λ is a standard distribution and easy to simulate from, the FCD for α is NOT!
- Therefore we can't use a Gibbs sampler for this analysis
- Fortunately there are other methods we can use ...

4.5 Metropolis–Hastings sampling

- The Gibbs sampler is a very powerful tool
 - Only useful if the full conditional distributions (FCDs) are standard distributions (which are easy to simulate from)
 - Fortunately there is a class of methods which can be used when the FCDs are non-standard
 - These methods are known as Metropolis-Hastings schemes
-
- Want to simulate realisations from the posterior density $\pi(\boldsymbol{\theta}|\mathbf{x})$
 - All of the FCDs are non-standard
 - Choose a **proposal distribution** with density $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$, which is easy to simulate from
 - This distribution gives us a way of proposing new values $\boldsymbol{\theta}^*$ from the current value $\boldsymbol{\theta}$

Metropolis–Hastings algorithm

- 1 Initialise the iteration counter to $j = 1$, and initialise the chain to $\theta^{(0)}$
- 2 Generate a **proposed** value θ^* using the proposal distribution $q(\theta^*|\theta^{(j-1)})$
- 3 Evaluate the **acceptance probability** $\alpha(\theta^{(j-1)}, \theta^*)$ of the proposed move, where

$$\alpha(\theta, \theta^*) = \min \left\{ 1, \frac{\pi(\theta^*|\mathbf{x}) q(\theta|\theta^*)}{\pi(\theta|\mathbf{x}) q(\theta^*|\theta)} \right\}$$

- 4 Set $\theta^{(j)} = \theta^*$ with probability $\alpha(\theta^{(j-1)}, \theta^*)$, and set $\theta^{(j)} = \theta^{(j-1)}$ otherwise; that is:

$\theta^{(j)}$	θ^*	$\theta^{(j-1)}$
Probability	α	$1 - \alpha$

- 5 Change the counter from j to $j + 1$ and return to step 2

Metropolis–Hastings algorithm

- 1 Initialise the iteration counter to $j = 1$, and initialise the chain to $\theta^{(0)}$
- 2 Generate a **proposed** value θ^* using the proposal distribution $q(\theta^*|\theta^{(j-1)})$
- 3 Evaluate the **acceptance probability** $\alpha(\theta^{(j-1)}, \theta^*)$ of the proposed move, where

$$\alpha(\theta, \theta^*) = \min \left\{ 1, \frac{\pi(\theta^*|\mathbf{x}) q(\theta|\theta^*)}{\pi(\theta|\mathbf{x}) q(\theta^*|\theta)} \right\}$$

- 4 Set $\theta^{(j)} = \theta^*$ with probability $\alpha(\theta^{(j-1)}, \theta^*)$, and set $\theta^{(j)} = \theta^{(j-1)}$ otherwise; that is:

$\theta^{(j)}$	θ^*	$\theta^{(j-1)}$
Probability	α	$1 - \alpha$

- 5 Change the counter from j to $j + 1$ and return to step 2

Metropolis–Hastings algorithm

- 1 Initialise the iteration counter to $j = 1$, and initialise the chain to $\theta^{(0)}$
- 2 Generate a **proposed** value θ^* using the proposal distribution $q(\theta^*|\theta^{(j-1)})$
- 3 Evaluate the **acceptance probability** $\alpha(\theta^{(j-1)}, \theta^*)$ of the proposed move, where

$$\alpha(\theta, \theta^*) = \min \left\{ 1, \frac{\pi(\theta^*|\mathbf{x}) q(\theta|\theta^*)}{\pi(\theta|\mathbf{x}) q(\theta^*|\theta)} \right\}$$

- 4 Set $\theta^{(j)} = \theta^*$ with probability $\alpha(\theta^{(j-1)}, \theta^*)$, and set $\theta^{(j)} = \theta^{(j-1)}$ otherwise; that is:

$\theta^{(j)}$	θ^*	$\theta^{(j-1)}$
Probability	α	$1 - \alpha$

- 5 Change the counter from j to $j + 1$ and return to step 2

Comments

- At each stage, a new value is generated from the proposal distribution
- This value is either accepted, in which case the chain moves, or rejected, in which case the chain stays where it is
- Whether or not the move is accepted or rejected depends on an acceptance probability which itself depends on the relationship between the density of interest and the proposal distribution
- The posterior density $\pi(\boldsymbol{\theta}|\mathbf{x})$ only enters into the acceptance probability as a ratio, and so the method can be used when the posterior density is only known up to a scaling constant, that is,

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \min \left\{ 1, \frac{\pi(\boldsymbol{\theta}^*) f(\mathbf{x}|\boldsymbol{\theta}^*) q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta}) q(\boldsymbol{\theta}^*|\boldsymbol{\theta})} \right\}$$

- This algorithm defines a Markov chain with $\pi(\boldsymbol{\theta}|\mathbf{x})$ as its stationary distribution
- It holds for all possible proposal distributions (subject to them generating realisations from the full parameter space)

Possible proposal distributions

- Are some choices of proposal distribution better than others?
- We now discuss some commonly used proposal distributions

4.5.1 Symmetric chains (Metropolis method)

- Use a symmetric proposal distribution: $q(\theta^*|\theta) = q(\theta|\theta^*)$, $\forall \theta, \theta^*$
- The acceptance probability simplifies to

$$\alpha(\theta, \theta^*) = \min \left\{ 1, \frac{\pi(\theta^*|\mathbf{x})}{\pi(\theta|\mathbf{x})} \right\} = \min \left\{ 1, \frac{\pi(\theta^*) f(\mathbf{x}|\theta^*)}{\pi(\theta) f(\mathbf{x}|\theta)} \right\}$$

and hence does not involve the proposal density at all

- Proposed moves which will take the chain to a region of higher posterior density are always accepted
- Moves which take the chain to a region of lower posterior density are accepted with probability proportional to the ratio of the two densities
- Moves which will take the chain to a region of very low density will be accepted with very low probability

Possible proposal distributions

- Are some choices of proposal distribution better than others?
- We now discuss some commonly used proposal distributions

4.5.1 Symmetric chains (Metropolis method)

- Use a symmetric proposal distribution: $q(\theta^*|\theta) = q(\theta|\theta^*)$, $\forall \theta, \theta^*$
- The acceptance probability simplifies to

$$\alpha(\theta, \theta^*) = \min \left\{ 1, \frac{\pi(\theta^*|\mathbf{x})}{\pi(\theta|\mathbf{x})} \right\} = \min \left\{ 1, \frac{\pi(\theta^*) f(\mathbf{x}|\theta^*)}{\pi(\theta) f(\mathbf{x}|\theta)} \right\}$$

and hence does not involve the proposal density at all

- Proposed moves which will take the chain to a region of higher posterior density are always accepted
- Moves which take the chain to a region of lower posterior density are accepted with probability proportional to the ratio of the two densities
- Moves which will take the chain to a region of very low density will be accepted with very low probability

Random walk proposals

- Consider the random walk proposal in which the proposed value θ^* at stage j is

$$\theta^* = \theta^{(j-1)} + \mathbf{w}_j$$

where the $\mathbf{w}_j$ are independent and identically distributed random $p \times 1$ vectors (completely independent of the state of the chain)

- Suppose that the $\mathbf{w}_j$ have density $f(\cdot)$, which is easy to simulate from, has mean $\mathbf{0}$ and is symmetric about its mean
- We can then simulate an **innovation** $\mathbf{w}_j$, and set the **proposal** value to $\theta^* = \theta^{(j-1)} + \mathbf{w}_j$
- What distribution should we use for $f(\cdot)$?

Choice of innovation distribution

- A distribution which is simple and easy to simulate from is always a good idea; for example, the uniform or normal
Normal is generally better, but is a bit more expensive to simulate
- What variance should we choose for the innovation distribution?
- This choice will affect the acceptance probability, and hence the overall proportion of accepted moves
- If the variance of the innovation is too low, then most proposed values will be accepted, but the chain will move very slowly around the space — the chain is said to be too “cold”
- If the variance of the innovation is too large, very few proposed values will be accepted, but when they are, they will often correspond to quite large moves — the chain is said to be too “hot”
- Theoretically it has been shown that the optimal acceptance rate is around 0.234 — this is an asymptotic result (for large samples of data) — but experience suggests that an acceptance rate of around 20–30% is okay
- Thus, the variance of the innovation distribution should be “tuned” to get an acceptance rate of around this level

Example

- Asked by the UK Met Office to analyse annual maximum wind speed observations (mph) at 206 locations across the UK
- Suppose $X_i, i = 1, \dots, 50$ represent these annual maximum wind observations for a particular location in the Peak District
- An appropriate model is the *Generalised Extreme Value* distribution, that is, $X_i | \mu, \sigma, \xi \sim \text{GEV}(\mu, \sigma, \xi)$
- After discussions with Simon Brown, a climate scientist at the Met Office, we use:

$$\mu \sim N(55, 2)$$

$$\sigma \sim \text{Ga}(20, 10) \quad \text{and}$$

$$\xi \sim N(0, 0.5)$$

Example

- Asked by the UK Met Office to analyse annual maximum wind speed observations (mph) at 206 locations across the UK
- Suppose $X_i, i = 1, \dots, 50$ represent these annual maximum wind observations for a particular location in the Peak District
- An appropriate model is the *Generalised Extreme Value* distribution, that is, $X_i | \mu, \sigma, \xi \sim \text{GEV}(\mu, \sigma, \xi)$
- After discussions with Simon Brown, a climate scientist at the Met Office, we use:

$$\mu \sim N(55, 2)$$

$$\sigma \sim \text{Ga}(20, 10) \quad \text{and}$$

$$\xi \sim N(0, 0.5)$$

Example

- Asked by the UK Met Office to analyse annual maximum wind speed observations (mph) at 206 locations across the UK
- Suppose $X_i, i = 1, \dots, 50$ represent these annual maximum wind observations for a particular location in the Peak District
- An appropriate model is the *Generalised Extreme Value* distribution, that is, $X_i | \mu, \sigma, \xi \sim \text{GEV}(\mu, \sigma, \xi)$
- After discussions with Simon Brown, a climate scientist at the Met Office, we use:

$$\mu \sim N(55, 2)$$

$$\sigma \sim \text{Ga}(20, 10) \quad \text{and}$$

$$\xi \sim N(0, 0.5)$$

Example

- Asked by the UK Met Office to analyse annual maximum wind speed observations (mph) at 206 locations across the UK
- Suppose $X_i, i = 1, \dots, 50$ represent these annual maximum wind observations for a particular location in the Peak District
- An appropriate model is the *Generalised Extreme Value* distribution, that is, $X_i | \mu, \sigma, \xi \sim \text{GEV}(\mu, \sigma, \xi)$
- After discussions with Simon Brown, a climate scientist at the Met Office, we use:

$$\mu \sim N(55, 2)$$

$$\sigma \sim \text{Ga}(20, 10) \quad \text{and}$$

$$\xi \sim N(0, 0.5)$$

Example

- No conjugate prior distributions exist for this model, and the full conditionals are not of standard form
- We need to use Metropolis-Hastings; we use Normal random walk updates:

$$\begin{aligned}\mu^* &= \mu^{(j-1)} + w_{\mu_j}, \\ \log \sigma^* &= \log \sigma^{(j-1)} + w_{\sigma_j} \quad \text{and} \\ \xi^* &= \xi^{(j-1)} + w_{\xi_j},\end{aligned}$$

where

$$\begin{aligned}w_{\mu_j} &\sim N(0, \epsilon_{\mu}), \\ w_{\sigma_j} &\sim N(0, \epsilon_{\sigma}) \quad \text{and} \\ w_{\xi} &\sim N(0, \epsilon_{\xi})\end{aligned}$$

Example

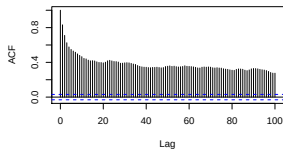
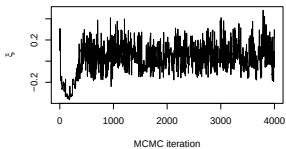
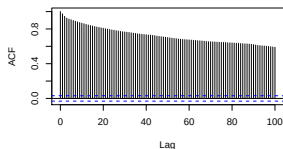
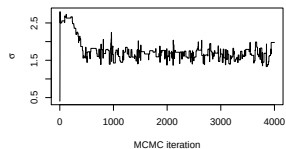
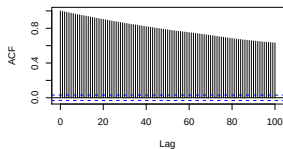
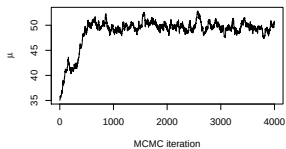
- No conjugate prior distributions exist for this model, and the full conditionals are not of standard form
- We need to use Metropolis-Hastings; we use Normal random walk updates:

$$\begin{aligned}\mu^* &= \mu^{(j-1)} + w_{\mu_j}, \\ \log \sigma^* &= \log \sigma^{(j-1)} + w_{\sigma_j} \quad \text{and} \\ \xi^* &= \xi^{(j-1)} + w_{\xi_j},\end{aligned}$$

where

$$\begin{aligned}w_{\mu_j} &\sim N(0, \epsilon_\mu), \\ w_{\sigma_j} &\sim N(0, \epsilon_\sigma) \quad \text{and} \\ w_\xi &\sim N(0, \epsilon_\xi)\end{aligned}$$

Example



4.6 Hybrid methods

- The Metropolis-Hastings method in last section samples θ in a single block, i.e. all parameters are updated at the same time.
- When the number of parameter is large, it is difficult to choose the proposal distribution. For example, when there are 5 parameters, the proposal distribution $N(0, \Sigma)$ has $5 \times 5 = 25$ tuning parameters in Σ .
- Recall the Gibbs sampler for how the componentwise update works.

4.6.1 Componentwise transitions

- Given a posterior distribution with FCDs that are awkward to sample from directly, we can define a Metropolis-Hastings scheme for each full conditional distribution, and apply them to each component in turn for each iteration
- This is like the Gibbs sampler, but each component update is a Metropolis-Hastings update, rather than a direct simulation from the FCD
- Each of these steps will require its own proposal distribution

The algorithm

- 1 Initialise the iteration counter to $j = 1$
Initialise the state of the chain to $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \dots, \theta_p^{(0)})^T$
- 2 Let $\boldsymbol{\theta}_{-i}^{(j)} = (\theta_1^{(j)}, \dots, \theta_{i-1}^{(j)}, \theta_{i+1}^{(j-1)}, \dots, \theta_p^{(j-1)})^T$, $i = 1, 2, \dots, p$
Obtain a new value $\boldsymbol{\theta}^{(j)}$ from $\boldsymbol{\theta}^{(j-1)}$ by successive generation of values
 - $\theta_1^{(j)} \sim \pi(\theta_1 | \boldsymbol{\theta}_{-1}^{(j)}, \mathbf{x})$ using a Metropolis–Hastings step with proposal distribution $q_1(\theta_1 | \theta_1^{(j-1)}, \boldsymbol{\theta}_{-1}^{(j)})$
 - $\theta_2^{(j)} \sim \pi(\theta_2 | \boldsymbol{\theta}_{-2}^{(j)}, \mathbf{x})$ using a Metropolis–Hastings step with proposal distribution $q_2(\theta_2 | \theta_2^{(j-1)}, \boldsymbol{\theta}_{-2}^{(j)})$
 - $\vdots$
 - $\theta_p^{(j)} \sim \pi(\theta_p | \boldsymbol{\theta}_{-p}^{(j)}, \mathbf{x})$ using a Metropolis–Hastings step with proposal distribution $q_p(\theta_p | \theta_p^{(j-1)}, \boldsymbol{\theta}_{-p}^{(j)})$
- 3 Change counter j to $j + 1$, and return to step 2

The distributions $\pi(\theta_i | \boldsymbol{\theta}_{-i}^{(j)}, \mathbf{x})$ are just the FCDs

Hybrid methods

- The Gibbs sampler can be used to sample from multivariate posterior distributions provided that we can simulate from the full conditional distributions (FCDs). The acceptance probability is 1.
- The Metropolis-Hastings method can be used to sample from awkward FCDs. The acceptance probability is less than 1.
- We can combine these, when some FCDs can be simulated from directly and some can not, to increase the acceptance probability.

4.6.2 Metropolis within Gibbs

- Given a posterior distribution with full conditional distributions,
 - some of which may be simulated from directly,
 - and others of which have Metropolis-Hastings updating schemes,the Metropolis within Gibbs algorithm goes through each in turn, and simulates directly from the full conditional, or carries out a Metropolis-Hastings update as necessary
- This algorithm is, in fact, just the above algorithm but uses the full conditional distributions as the proposal distributions when they are easy to simulate from
- To see this, suppose that we can simulate from the FCD $\pi(\theta_i | \boldsymbol{\theta}_{-i}^{(j)}, \mathbf{x})$ and use this as the proposal distribution, that is, take
$$\theta_i^* \sim \pi(\theta_i | \boldsymbol{\theta}_{-i}^{(j)}, \mathbf{x})$$

- Then the acceptance probability for this step is

$$\begin{aligned}\alpha(\theta_i, \theta_i^*) &= \min \left\{ 1, \frac{\pi(\theta_i^* | \boldsymbol{\theta}_{-i}^{(j)}, \mathbf{x})}{\pi(\theta_i | \boldsymbol{\theta}_{-i}^{(j)}, \mathbf{x})} \frac{q(\theta_i | \theta_i^*, \boldsymbol{\theta}_{-i}^{(j)})}{q(\theta_i^* | \theta_i, \boldsymbol{\theta}_{-i}^{(j)})} \right\} \\ &= \min \left\{ 1, \frac{\pi(\theta_i^* | \boldsymbol{\theta}_{-i}^{(j)}, \mathbf{x})}{\pi(\theta_i | \boldsymbol{\theta}_{-i}^{(j)}, \mathbf{x})} \frac{\pi(\theta_i | \boldsymbol{\theta}_{-i}^{(j)}, \mathbf{x})}{\pi(\theta_i^* | \boldsymbol{\theta}_{-i}^{(j)}, \mathbf{x})} \right\} \\ &= \min\{1, 1\} \\ &= 1,\end{aligned}$$

that is, we always accept the proposal from the FCD

Example 4.5

- Construct an MCMC scheme for the problem in Example 4.4 where we had a random sample of size n from a gamma $Ga(\alpha, \lambda)$ distribution and independent gamma $Ga(a, b)$ and $Ga(c, d)$ prior distributions for α and λ respectively
- Recall that the FCDs were

$$\pi(\alpha|\lambda, \mathbf{x}) \propto \frac{\alpha^{a-1} e^{(-b+n \log \bar{x}_g + n \log \lambda)\alpha}}{\Gamma(\alpha)^n}, \quad \alpha > 0$$

and

$$\pi(\lambda|\alpha, \mathbf{x}) \propto \lambda^{c+n\alpha-1} e^{-(d+n\bar{x})\lambda}, \quad \lambda > 0$$

Solution

...

R code

- Use the function `mwgGamma` in the `nclbayes` library
- Data: $n = 50$, $\bar{x} = 0.62$, $\bar{x}_g = 0.46$ and $s = 0.4$
- Prior: $a = 2$, $b = 1$, $c = 3$, $d = 1$
- Use a normal random walk proposal with variance $\Sigma_{\alpha} = 0.9^2$
- This gives a reasonable acceptance probability of 0.237

```
library(nclbayes)
posterior=mwgGamma(N=20000,initial=(xbar/s)^2,innov=0.9,
  priorparam=c(2,1,3,1),n=50,xbar=0.62,xgbar=0.46,show=TRUE)
mcmcAnalysis(posterior,rows=2,show=F)
posterior2=mcmcProcess(input=posterior,burnin=10,thin=20)
mcmcAnalysis(posterior2,rows=2)
```

- Using `burnin = 10`, `thin = 20` produces (almost) un-autocorrelated posterior output

Analysis of MCMC output (all realisations)

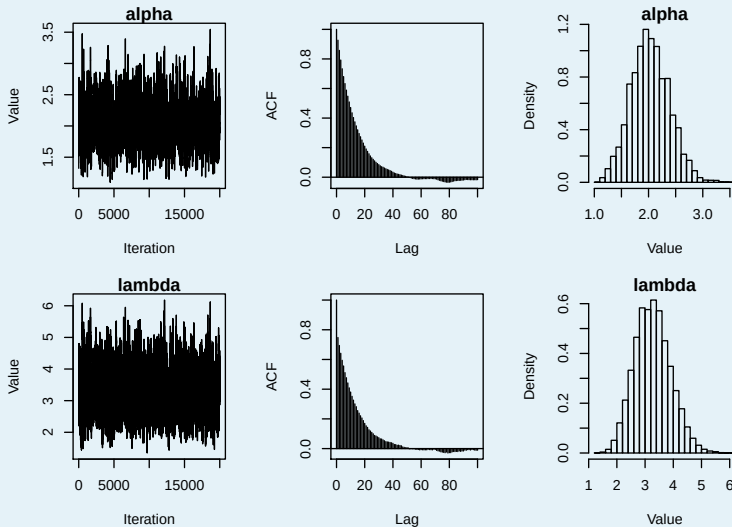


Figure: Trace plots, autocorrelation plots and histograms of the Metropolis within Gibbs sampler output

Analysis of MCMC output (with burnin = 10, thin = 20)

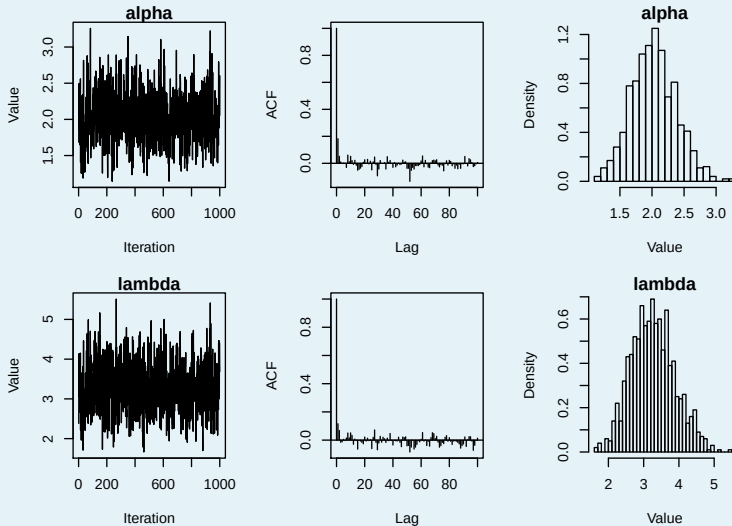


Figure: Trace plots, autocorrelation plots and histograms of the Metropolis within Gibbs sampler output

4.7 Summary

- i Bayesian inference can be complicated when not using a conjugate prior distribution
- ii One solution is to use Markov chain Monte Carlo (MCMC) methods
- iii These work by producing realisations from the posterior distribution by constructing a Markov chain which has the posterior distribution as its stationary distribution
- iv The MCMC methods we have studied are the Gibbs sampler, Metropolis within Gibbs algorithm and the Metropolis–Hastings algorithm
- v When obtaining output from these algorithms, we need to assess whether there needs to be a burn-in and whether the output needs to be thinned (by looking at traceplots and autocorrelation plots) using `mcmcAnalysis` and `mcmcProcess`

(vi) The (converged and thinned) MCMC output are realisations from the posterior distribution. It can be used to

- obtain the posterior distribution for any (joint) functions of the parameters (such as $\sigma = 1/\sqrt{\tau}$ or $(\theta_1 = \mu - \tau, \theta_2 = e^{\mu + \tau/2})^T$)
- look at bivariate posterior distributions via scatter plots
- look at univariate marginal posterior distributions via histograms or boxplots
- obtain numerical summaries such as the mean, standard deviation and confidence intervals for single variables and correlations between variables

(vii) Equi-tailed posterior confidence intervals can be determined from the MCMC output using `mcmcCi`

4.8 Learning objectives

By the end of this chapter, you should be able to:

1. explain why not using a conjugate prior generally causes problems in determining the posterior distribution
2. describe the Gibbs sampler, explain why it is a Markov chain and give an outline as to why its stationary distribution is the posterior distribution
3. describe the issues of processing MCMC output (burn-in, autocorrelation, thinning etc.) and interpret numerical/graphical output
4. derive the full conditional densities for any posterior distribution and name these distributions if they are “standard” distributions given in the notes or on the exam paper

5. describe a Metropolis-Hastings algorithm in general terms and when using symmetric random walk proposals or independence proposals or general proposals
6. describe the hybrid methods: *componentwise transitions* and *Metropolis within Gibbs*
7. provide a detailed description of **any** of the MCMC algorithms as they apply to generating realisations from **any** posterior distribution