Lecture 10: Review Gibbs sampling

Idea: Sequentially sampling from univariate conditional distributions (which are often available in closed form).

1. Select starting values $\boldsymbol{x}_0$ and set $i = 0$.
2. Repeatedly:

$$
\begin{aligned}
&\text{Sample} \quad x^1_{i+1}|\cdot \sim \pi(x^1|x^2_i, \dots, x^p_i)\\
&\text{Sample} \quad x^2_{i+1}|\cdot \sim \pi(x^2|x^1_{i+1}, x^3_i, \dots, x^p_i)\\
&\vdots\\
&\text{Sample} \quad x^{p-1}_{i+1}|\cdot \sim \pi(x^{p-1}|x^1_{i+1}, x^2_{i+1}, \dots, x^{p-2}_{i+1}, x^p_i)\\
&\text{Sample} \quad x^p_{i+1}|\cdot \sim \pi(x^p|x^1_{i+1}, \dots, x^{p-1}_{i+1})
\end{aligned}
$$

   where $|\cdot$ denotes conditioning on the most recent updates of all other elements of $\boldsymbol{x}$.

3. Increment i and go to step 2.

Why is the acceptance rate 1?

For ease of notation let x denote the current state and $x^\star$ the proposed new state where we update the $j-$th component of x, so that:

$$
\begin{aligned}
x &= (x^1, \dots, x^{j-1}, x^j, x^{j+1}, \dots, x^p)^\top\\
x^\star &= (x^1, \dots, x^{j-1}, x^{\star,j}, x^{j+1}, \dots, x^p)^\top
\end{aligned}
$$

where $x^{\star,j}$ denotes the propsed value for the $j-$th component. Then

$$
\begin{aligned}
\frac{\pi(x^\star)}{\pi(x)} \cdot \frac{Q(x \mid x^\star)}{Q(x^\star \mid x)} &= \frac{\pi(x^{\star,j} \mid x^{\star,-j})\pi(x^{\star,-j})}{\pi(x^j \mid x^{-j})\pi(x^{-j})} \cdot \frac{Q(x \mid x^\star)}{Q(x^\star \mid x)}\\
&= \frac{\pi(x^{\star,j} \mid x^{-j})\pi(x^{-j})}{\pi(x^j \mid x^{-j})\pi(x^{-j})} \cdot \frac{Q(x \mid x^\star)}{Q(x^\star \mid x)}\\
&= \frac{\pi(x^{\star,j} \mid x^{-j})\pi(x^{-j})}{\pi(x^j \mid x^{-j})\pi(x^{-j})} \cdot \frac{\pi(x^j \mid x^{\star,-j})}{\pi(x^{\star,j} \mid x^{-j})}\\
&= 1
\end{aligned}
$$

Example: Deriving full-conditionals

Assume $y_i|\mu, \kappa \sim \mathcal{N}(\mu, \kappa^{-1})$, $i = 1, \dots, n$. As prior for μ and κ we choose a normal and gamma distribution, respectively, where:

$$
\begin{aligned}
\mu &\sim \mathcal{N}(\mu_0, \kappa_0^{-1})\\
\kappa &\sim \mathcal{G}(a, b)
\end{aligned}
$$

The full-conditionals are

$$
\begin{aligned}
\mu|\kappa, \boldsymbol{y} &\sim \mathcal{N}\left(\frac{\mu_0\kappa_0 + \bar{y}n\kappa}{\kappa_0 + n\kappa}, (\kappa_0, n\kappa)^{-1}\right)\\
\kappa|\mu, \boldsymbol{y} &\sim \mathcal{G}\left(a + \frac{n}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2\right)
\end{aligned}
$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ denotes the mean over all y.

Implementation and convergence diagnostics

Source: http://i.telegraph.co.uk/multimedia/archive/02365/coding_alamy_2365972b.jpg

Numerical note

How should you compute

$$\alpha = \min\left(1, \frac{\pi(x^\star)}{\pi(x_{i-1})} \times \frac{Q(x_{i-1}|x^\star)}{Q(x^\star|x_{i-1})}\right)$$

See blackboard

Burn-in

In practice, one waits until the Markov chain is converged. Let K denote the burn-in period. Then the realisations $\boldsymbol{x}_{K+1}, \boldsymbol{x}_{K+2}, \ldots, \boldsymbol{x}_{K+N}$ are used to estimate characteristics of the target distribution.

The empirical determination of K is difficult. Often it is determined based on the trace plot of the Markov chain.

Convergence diagnostics

Valid inferences from sequences of MCMC outputs are based on the assumption that the outputs are from the desired target distribution.

- There is no overall minimum number of samples to ensure approximation.
- Consequently methods for testing convergence, known as convergence diagnostics, have to be applied.
- However it has to emphasised that these diagnostics do not guarantee convergence.

Trace plots

An initial possibility for deciding if a MCMC output does not converge to the desired posterior distributions is to look at the sample trace for each variable.

- If our chain is taking a long time to move around the parameter space, then it will take longer to converge.
- If the samples form a homogene band (no wave movements or other rare fluctuations), convergence might be indicated.
- Vastly different values at the beginning of the trace indicate burn-in iterations, which should be discarded.

Autocorrelation

To examine dependencies of successive MCMC samples, the autocorrelation function can be used. Let $x_1, \dots, x_N$, where N denotes the number of samples, denote our MCMC chain.

The lag k autocorrelation $\rho(k)$ is the correlation between every draw and its k-th lag. For N reasonably large

$$\rho(k) \approx \frac{\sum_{i=1}^{N-k}(x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2},$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ is the overall mean.

- With increasing lag k we expect lower autocorrelations.
- If autocorrelation is still relatively high for higher values of k, this indicates high degree of correlation between our draws and slow mixing.

Geweke diagnostics

The MCMC chain is divided into two windows

- the first x%, and
- the last y% of the iterates

(coda default: $x = 10$, $y = 50$). For both windows the mean is calculated.

If the chain is stationary both values should be equal and Geweke's test statistic (z-score) follows an asymptotical standard normal distribution.

Further reading

There are several convergence diagnostics:

- some are based on a single Markov chain run
- some are based on several Markov chain runs

For further reading see for example

- Gilks, W. R., Richardson, S. and Spiegelhalter, D.J. (1996) *Markov Chain Monte Carlo in Practice, Chapman & Hall, London*,

Different approaches are implemented in the

- R-package coda. (Plummer et al., 2006)

Effective sample size

A useful measure to compare the performance of different MCMC samplers is the effective sample size (ESS) Kass et al. (1998) American Statistician 52, 93–100..

- The ESS is the estimated number of independent samples needed to obtain a parameter estimate with the same precision as the MCMC estimate based on N dependent samples.

$$\text{ESS} = \frac{N}{\tau}, \quad \tau = 1 + 2 \cdot \sum_{k=1}^{\infty} \rho(k),$$

where τ is the autocorrelation time and $\rho(k)$ the autocorrelation at lag k.

Autocorrelation time

- There are different stopping criteria for the sum. Geyer (1992, Statistical Science, page 477)) proposed the initial monotone sequence estimator, where

$$\tau = 1 + 2 \cdot \sum_{k=1}^{2m+1} \rho(k)$$

where m is chosen to be the largest integer such that

$$\Gamma_i = \rho(2i) + \rho(2i+1), \quad i = 1, \ldots, m$$

is positive and the sequence $\Gamma_1, \ldots, \Gamma_m$ is monotone decreasing.

Beetle mortality data (Bliss (1935), Annals of Applied Biology, 22: 134–167)

Beetles are exposed to gaseous carbon disulphide at various concentrations for five hours.

- y_i number killed out of n_i at i-th dose level, $i = 1, \ldots, 8$.
- x_i log dose.

Dose, x_i	Number of	Number
($\log_{10} CS_2 mgl^{-1}$)	beetles, n_i	killed, y_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Logistic regression model

- Assuming independence of the beetles, $y_i \sim \text{Bin}(n_i, \pi_i)$:

$$p(\boldsymbol{y}|\pi_i) = \prod_{i=1}^{8} \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}$$

where π_i denotes the probability of being killed at the i-th dose level.

(Comment: Independence assumption would not be appropriate if the deaths were caused by a contagious disease)

- Logistic model:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \alpha + \beta(x_i - \bar{x})$$

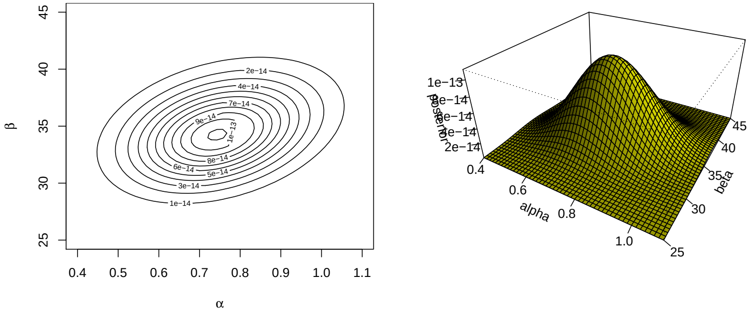
$$\pi_i = \text{expit}(\alpha + \beta(x_i - \bar{x})) = \frac{\exp(\alpha + \beta(x_i - \bar{x}))}{1 + \exp(\alpha + \beta(x_i - \bar{x}))}$$

- Independent normal prior distribution

$$\alpha \sim \mathcal{N}(0, \sigma_\alpha^2) \qquad\qquad \beta \sim \mathcal{N}(0, \sigma_\beta^2)$$

- Choose precisions, $\tau_\alpha = 1/\sigma_\alpha^2$, and $\tau_\beta = 1/\sigma_\beta^2$, to be small; e.g. 10^{-4}.

Posterior distribution



The posterior distribution is

$$p(\alpha, \beta | \boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x}) \propto p(\alpha)\, p(\beta) \prod_{i=1}^{8} p(y_i | \alpha, \beta, n_i, x_i),$$

which is no standard distribution. For estimating α and β we implement an Metropolis-Hastings algorithm with

- two univariate random walk proposals (Metropolis-within-Gibbs).
- one bivariate random walk proposal.

Target densities

Univariate update

- The full-conditional distributions are:

$$p(\alpha|\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x}, \beta) \propto p(\alpha) \prod_{i=1}^{8} p(y_i|\alpha, \beta, n_i, x_i)$$

$$p(\beta|\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x}, \alpha) \propto p(\beta) \prod_{i=1}^{8} p(y_i|\alpha, \beta, n_i, x_i)$$

- For each parameter we choose a normal proposal with mean equal to the current value and variances tuned to get acceptance rates between $20 - 50\%$.

Target densities

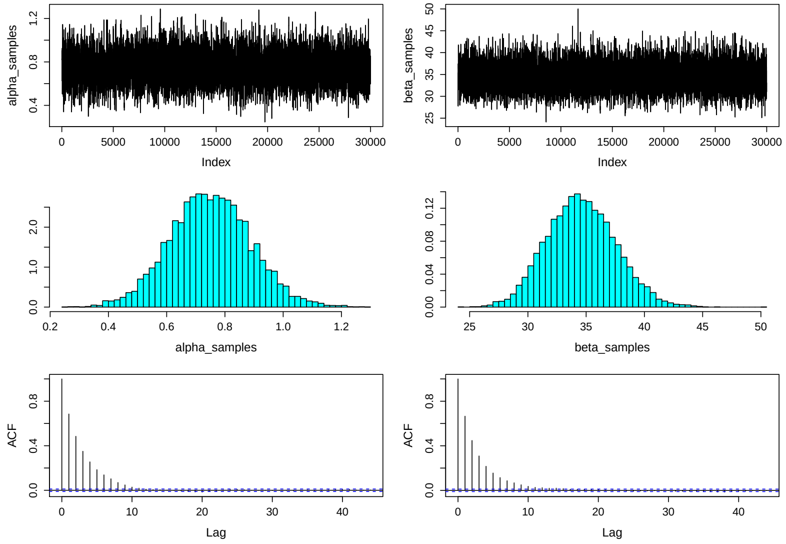
Bivariate update

- Here, the target density is the posterior distribution.
- Choose a normal proposal with mean equal to the current value and covariance matrix

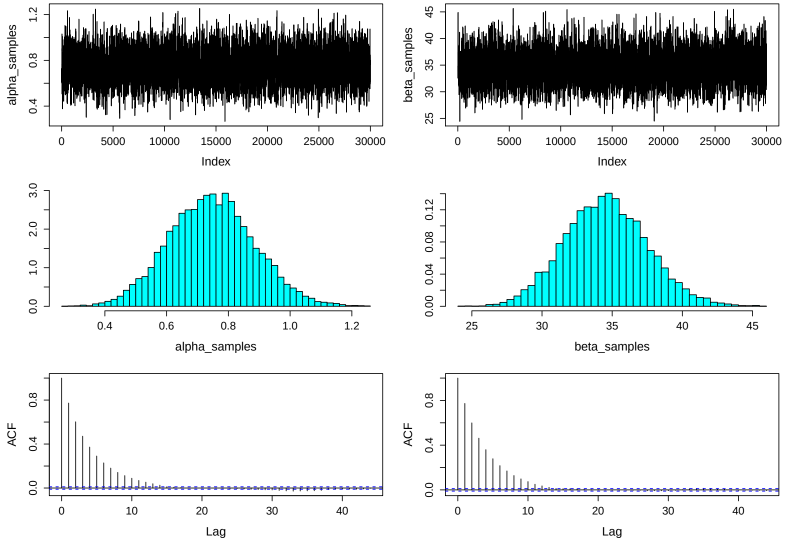
$$\Sigma = c \cdot \boldsymbol{I}_p^{-1},$$

where $\boldsymbol{I}_p^{-1}$ denotes the negative inverse curvature of the log posterior at the posterior mode and c is a factor to tune the acceptance rate.

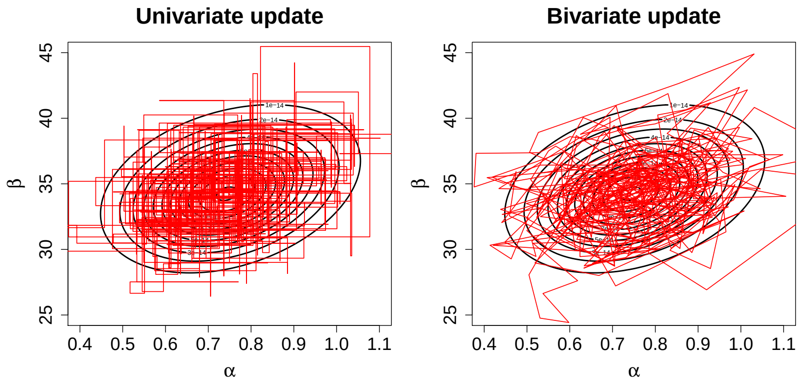
Univariate update: Diagnostic checks



Bivariate update: Diagnostic checks



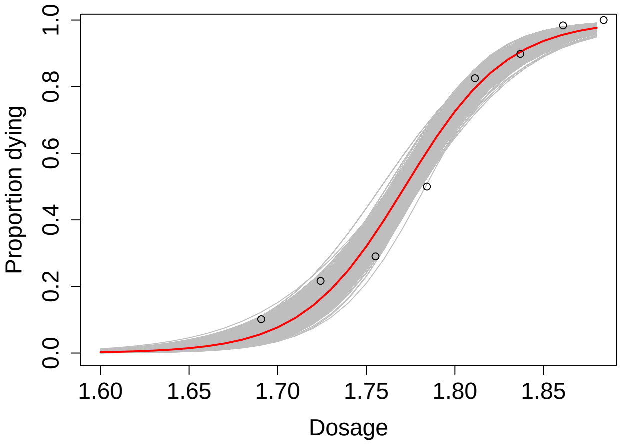
Exploration of posterior



Results

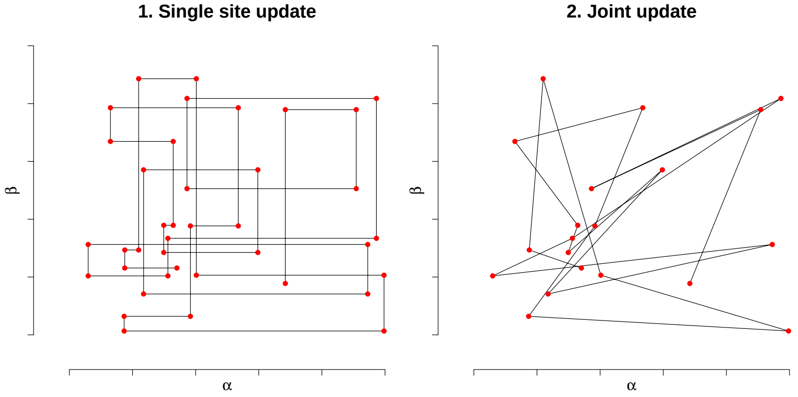
```
> ## Fit a generalized linear model to compare
> m1 <- glm(formula = cbind(y, n - y) ~ x, family = binomial)
> #            Estimate Std. Error z value Pr(>|z|)
> #(Intercept)   0.7438     0.1379   5.396 6.83e-08 ***
> #x            34.2703     2.9121  11.768  < 2e-16 ***
>
> ## Univariate Update
> #> summary(alpha_samples)
> #   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
> # 0.2256  0.6582  0.7505  0.7501  0.8378  1.3340
> #> summary(beta_samples)
> #   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
> #  24.26   32.58   34.47   34.56   36.46   46.76
>
> ## Bivariate Update
> #> summary(alpha_samples)
> #   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
> # 0.2569  0.6566  0.7470  0.7505  0.8400  1.3540
> #> summary(beta_samples)
> #   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
> #  23.77   32.54   34.50   34.57   36.51   47.59
```

Dose-response curve



Updating schemes

1. Update α and β separately $\Rightarrow$ Two acceptance steps.
2. Update α and β jointly $\Rightarrow$ One acceptance step.



Joint updates might be more efficient, however for some parameter combinations the acceptance rates can be very low.