The Metropolis-Hastings algorithm

Special cases : – The Metropolis algorithm
– The Gibbs sampler

1. Define a proposal distribution (or jumping distribution) $J_t(\theta^* | \theta^{t-1})$ for proposing a new value θ^* conditional on θ^{t-1}

2. Draw (or set) a starting value θ^0 for which $p(\theta^0 | y) > 0$, might be based on a crude approximate estimate

3. For $t = 1, \dots, T$ (sampling the new values θ^t)

(a) Sample a proposal θ^* from $J_t(\theta^* | \theta^{t-1})$

(b) Calculate

$$r = \frac{p(\theta^* | y) \big/ J_t(\theta^* | \theta^{t-1})}{p(\theta^{t-1} | y) \big/ J_t(\theta^{t-1} | \theta^*)}$$

and $\alpha(\theta^{t-1}, \theta^*) = \min\{r, 1\}$ ← Acceptance probability

(c) Set $\theta^t = \begin{cases} \theta^* & \text{with probability } \alpha(\theta^{t-1}, \theta^*) \\ \theta^{t-1} & \text{otherwise} \end{cases}$ ← Counts as a simulated value

nov 10-12:18

Remarks

– r is always defined, since in order for θ^* to be proposed, we must have $p(\theta^{t-1} | y) > 0$ and $J(\theta^* | \theta^{t-1}) > 0$

– Relatively easy to deduce that the transition distribution of the Markov Chain

$$T_t(\theta^t | \theta^{t-1}) = \alpha(\theta^{t-1}, \theta^t) \cdot J_t(\theta^t | \theta^{t-1}) + \delta(\theta^t = \theta^{t-1}) \cdot \int (1 - \alpha(\theta^{t-1}, \theta')) \cdot J_t(\theta' | \theta^{t-1}) \, d\theta'$$

NB: Sensible to compute

$$r = \exp\{\log p(\theta^* | y) - \log J_t(\theta^* | \theta^{t-1}) - \log p(\theta^{t-1} | y) + \log J_t(\theta^{t-1} | \theta^*)\}$$

Why does it work

From Markov Chain theory, we know that $p(\theta | y)$ is the stationary distr. of the irreducible, aperiodic and recurrent Markov Chain if "the detailed balance" is satisfied:

$$p(\theta^{t-1} | y) \cdot T_t(\theta^t | \theta^{t-1}) = p(\theta^t | y) T_t(\theta^{t-1} | \theta^t) \quad (*)$$

Trivial for $\theta^t = \theta^{t-1}$

For $\theta^t \neq \theta^{t-1}$ we have $T_t(\theta^t / \theta^{t-1}) = \alpha(\theta^{t-1}, \theta^t) J_t(\theta^t | \theta^{t-1})$

If $\alpha(\theta^{t-1}, \theta^t) = \dfrac{p(\theta^t | y) / J_t(\theta^t | \theta^{t-1})}{p(\theta^{t-1} | y) / J_t(\theta^{t-1} / \theta^t)}$ then $r \leq 1$

then $\alpha(\theta^t, \theta^{t-1}) = 1$ and $T_t(\theta^{t-1} | \theta^t) = J_t(\theta^t | \theta^{t-1})$

and $(*)$ is satisfied, and obviously if $\alpha(\theta^t, \theta^{t-1}) = r$, then $(*)$ is also satisfied

nov 10-12:35

Conditions: The constructed Markov Chain must be irreducible, aperiodic, recurrent

Must construct J_t such that there is a positive probability of eventually jumping to all values of θ for which $p(\theta|y)>0$

Holds for all practical choices of J_t

The Metropolis algorithm

$J_t(\cdot|\cdot)$ is required to be symmetric, i.e. $J_t(\theta^a|\theta^b) = J_t(\theta^b|\theta^a)$

Then, r simplifies to

$$r = \frac{p(\theta^*|y)}{p(\theta^{t-1}|y)}$$

Simple, illustrative example

Target density: $p(\theta|y) = N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1&0\\0&1\end{pmatrix}\right)$

Pretend we don't know this, only that

$p(\theta|y) \propto \exp\{-\frac{1}{2}\theta^T\theta\}$

Choose $J_t(\theta^*|\theta^{t-1}) = N\left(\theta^* \,\middle|\, \theta^{t-1}, 0.2^2\cdot\begin{pmatrix}1&0\\0&1\end{pmatrix}\right)$

nov 10-12:54

The Gibbs sampler

- Devide $\underset{\text{length } p}{\theta}$ into d subvectors $\theta = (\theta_{(1)}, \dots, \theta_{(d)})$ with $d \leq p$
- Define iteration t consist of d steps
  - Step j of iteration t consists of updating $\theta_{(j)}$ conditional on the current values of the other subvectors of θ, $\theta_{(-j)}$, and on y. The jumping distribution is then

$$J_{j,t}^{\text{Gibbs}}(\theta^*_{(j)}|\theta^{t-1}) = \underbrace{p(\theta^*_{(j)}|\theta^{t-1}_{(-j)}, y)}_{\text{Full cond. distr. , prop. to the joint posterior distr w.r.t } \theta_{(j)}} \text{ if } \theta^*_{(-j)} = \theta^{t-1}_{(-j)} \text{ (otherwise 0)}$$

Then

$$r = \frac{p(\theta^*|y)/J_{j,t}^{\text{Gibbs}}(\theta^*_{(j)}|\theta^{t-1})}{p(\theta^{t-1}|y)/J_{j,t}^{\text{Gibbs}}(\theta^{t-1}_{(j)}|\theta^*)}$$

$$= \frac{p(\theta^*|y)/p(\theta^*_{(j)}|\overbrace{\theta^{t-1}_{(-j)}}^{\theta^*_{(-j)}}, y)}{p(\theta^{t-1}|y)/p(\theta^{t-1}_{(j)}|\underbrace{\theta^*_{(-j)}}_{\theta^{t-1}_{(-j)}}, y)}$$

$$= \frac{p(\theta^{t-1}_{(-j)}|y)}{p(\theta^{t-1}_{(-j)}|y)} = 1$$

$$P(A,B|y) = \underbrace{P(B|y)}\cdot P(A|B,y)$$

nov 10-13:34

Metropolis-Hastings within Gibbs

- If some, or all, of $p(\theta_{(j)} | \theta_{(-j)}, y)$ are not possible to sample from, we can use Metropolis-Hastings (MH) for sampling from $p(\theta_{(j)} | \theta_{(-j)}, y)$
  - Then we construct a jumping distribution that only proposes a change in $\theta_{(j)}$, hence all proposals θ^* from $J_{j,t}(\theta^* | \theta^{t-1})$ are such that $\theta^*_{(-j)} = \theta^{t-1}_{(-j)}$
  - Acceptance/rejection of $\theta^*_{(j)}$ is decided as described as for MH.

Gibbs sampling is most efficient when the subvectors $\theta_{(1)}, \dots, \theta_{(d)}$ are such that there is high dependence within $\theta_{(j)}, j=1,\dots,d$

and low dependence between $\theta_{(1)}, \dots, \theta_{(d)}$

Blocking

At the extreme: Single-site Gibbs: $d=p$
(Slow if there is dependence between the parameters)

nov 10-13:49

Jumping rules and efficiency of simulations

Ideally, $J(\theta^* | \theta) = p(\theta^* | y)$, $\forall \theta$, then $r=1$ and we have a iid samples from $p(\theta | y)$.

Important $J_t(\theta^* | \theta)$:

- Easy to sample from for $\forall \theta$
- Easy to compute r
- Each jump is long enough so that the chain does not move too slowly
- The jumps are not rejected too often

D Must be balanced, the first implies lower r-values, the second high r values

Approximate rules of thumb:

- One-dimensional θ : r lie around 0.44
- High-dimensional θ: r lie around 0.23

nov 10-13:56

- Disregard burn-in (warm-up) iterations to reduce the dependence on the starting values.

  Conservative: Disregard the first half of iterations

- Assessing convergence:

  - More than one chain with overdispersed starting values, far away from eachother

    ↳ Allows to assess two important criteria for convergence

    1) Stationarity:

    2) Mixing: All chains reflect the whole global distribution, not getting "stuck" locally

       ↳ Must have at least $c=2$ chains to assess this

Formal check of stationarity and mixing:

1. Split (after removing burn-ins) each of chains in two, resulting in $m = 2 \cdot c$ chains, each of length $n = \frac{T}{2 \cdot 2}$

2. Compare between-and within-sequence variances:

nov 10-14:18

For the scalar estimand ψ we have the simulated values (after burn-ins have removed, and after splitting)

$\psi_{ij}, \quad i=1,\dots,n, \quad j=1,\dots,m$

We have the between-chain variance

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\overline{\psi}_{\cdot j} - \overline{\psi}_{\cdot\cdot})^2, \quad \text{where } \overline{\psi}_{\cdot j} = \frac{1}{n} \sum_{i=1}^{n} \psi_{ij}, \quad \overline{\psi}_{\cdot\cdot} = \frac{1}{m} \sum_{j=1}^{m} \overline{\psi}_{\cdot j}$$

and then the within-sequence variance

$$W = \frac{1}{m} \sum_{j=1}^{m} s_j^2, \quad s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\psi_{ij} - \psi_{\cdot j})^2$$

Now, the marginal posterior can be estimated by a weighted average of W and B:

$$\widehat{\text{var}}^{+}(\psi|y) = \frac{n-1}{n} W + \frac{1}{n} B$$

Now, consider $\hat{R} = \sqrt{\dfrac{\widehat{\text{var}}^{+}(\psi|y)}{W}}$

which is an estimate of the factor by which the scale of the current simulated distribution of ψ may be reduced as $n \to \infty$.

$$\hat{R} \xrightarrow[n\to\infty]{} 1$$

Be smart about starting values

- $\hat{R}$ can be used to monitor convergence

  Should approach 1 for all scalar estimands of interest. Rule of thumb $\hat{R} \leq 1.1$ ok!

nov 10-14:33

Intro to Bayesian regression modelling

We have a response variable y

k explanatory variables $x = (x_1, \ldots, x_k)$

Typically we observe y and x for n subjects, so we have

y_i, $x_i = (x_{i1}, \ldots, x_{ik})$ for $i = 1, \ldots, n$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \underset{n\times k}{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

The simplest regression model: The normal ordinary linear model:

$$E[y_i | \beta, x] = \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

↑ Often $x_{i1} = 1$ $\forall i$

β_1 is an intercept

and $\mathrm{Var}[y_i | \sigma^2, X] = \sigma^2$, $\forall i$

$$y | \beta, \sigma, X \sim N(X\beta, \sigma^2 I)$$

Need a prior: $p(\beta, \sigma^2 | X)$

Convenient non-inf. : $p(\beta, \sigma^2 | X) \propto \sigma^{-2}$ ← OK $\begin{cases} n \text{ large} \\ k \text{ small} \end{cases}$

nov 10-14:46

Posterior distribution

$$\beta | \sigma, y \sim N(\hat\beta, V_\beta \sigma^2), \quad \hat\beta = (X^T X)^{-1} X^T y$$

$$V_\beta = (X^T X)^{-1}$$

$$\sigma^2 | y \sim \mathrm{Inv}\text{-}\chi^2(n-k, s^2)$$

where $s^2 = \frac{1}{n-k} (y - X\hat\beta)^T (y - X\hat\beta)$

Generalized linear models

- Linear predictor: $\eta = X\beta$ (for n.l.r : $\eta = E[y_i | \beta, X]$)
- Link function: $g(\mu)$, linking the linear predictor to mean of the response variable:

  $\mu = E[y_i | \beta, X] = g^{-1}(\eta) = g^{-1}(X\beta)$

  For n.l.r. $g(\mu) = \mu$
- A specification of the distr. of y with mean $E[y | \beta, X] = \mu$. This can depend on a dispersion parameter ϕ.

nov 10-14:51