Christian P. Robert and Jean–Michel Marin
Université Paris Dauphine, Institut de Mathématiques et Modélisation de Montpellier, Université Montpellier 2, & CREST, INSEE, Paris

Bayesian Core: The Complete Solution Manual

October 24, 2009

Springer
Berlin Heidelberg New York
Hong Kong London Singapore
Milan Paris Tokyo

Preface

> The warning could not have been meant for the place
> where it could only be found after approach.
> —**Joseph Conrad**, ***Heart of Darkness***

This solution manual was initially intended for instructors using the book, so that they could entirely rely on the exercises to provide graded homeworks. However, due to repeated criticisms and to requests from readers, as well as to the fact that some exercises were stepping stones for following sections and to the lack of evidence that the exercises were indeed used to set homeworks, we came to realise (albeit late in the day!) that some solutions were needed by some (self-study or not) readers. From there, the move to make the *whole* set of solutions available to *all* readers was a rather natural step. Especially when contemplating the incoming revision of *Bayesian Core* towards a *Use R!* oriented version, with an attempt at reducing the math complexity, another reproach found in some of the published criticisms. Therefore, lo and behold!, by popular request, the solution manual is now available for free use and duplication by anyone, not only by instructors, on the book webpage as well as on Springer Verlag's website.

However, there is a caveat to the opening of the manual to all: since this solution manual was first intended (and written) for instructors, some self-study readers may come to the realisation that the solutions provided here are too sketchy for them because the way we wrote those solutions assumes some minimal familiarity with the maths, the probability theory and with the statistics behind the arguments. There is unfortunately a limit to the time and to the efforts we can put in this solution manual and studying *Bayesian Core* requires some prerequisites in maths (such as matrix algebra and Riemann integrals), in probability theory (such as the use of joint and conditional densities) and some bases of statistics (such as the notions of inference, sufficiency and confidence sets) that we cannot cover here. Casella and Berger (2001) is a good reference in case a reader is lost with the "basic" concepts

or sketchy math derivations. Indeed, we also came to realise that describing the book as"self-contained" was a dangerous add as readers were naturally inclined to always relate this term to their current state of knowledge, a bias resulting in inappropriate expectations. (For instance, some students unfortunately came to one of our short courses with no previous exposure to standard distributions like the t or the gamma distributions.)

We obviously welcome comments and questions on possibly erroneous solutions, as well as suggestions for more elegant or more complete solutions: since this manual is distributed both freely and independently from the book, it can be updated and corrected [almost] in real time! Note however that the R codes given in the following pages are not optimised because we prefer to use simple and understandable codes, rather than condensed and efficient codes, both for time constraints (this manual took about a whole week of August 2007 to complete) and for pedagogical purposes: the readers must be able to grasp the meaning of the R code with a minimum of effort since R programming is not supposed to be an obligatory entry to the book. In this respect, using R replaces the pseudo-code found in other books since it can be implemented as such but does not restrict understanding. Therefore, if you find better [meaning, more efficient/faster] codes than those provided along those pages, we would be glad to hear from you, but that does not mean that we will automatically substitute your R code for the current one, because readability is also an important factor.

Sceaux, France,October 24, 2009
Christian P. Robert

Contents

1

User's Manual

Exercise 1.1 Given a function g on $\mathbb{R}$, state the two basic conditions for g to be a probability density function (pdf) with respect to the Lebesgue measure. Recall the definition of the cumulative distribution function (cdf) associated with g and that of the quantile function of g.

If g is integrable with respect to the Lebesgue measure, g is a pdf if and only if

1. g is non-negative, $g(x) \geq 0$
2. g integrates to 1,

$$\int_{\mathbb{R}} g(x)\,\mathrm{d}x = 1\,.$$

Exercise 1.2 If (x_1, x_2) is a normal $\mathcal{N}_2((\mu_1, \mu_2), \Sigma)$ random vector, with

$$\Sigma = \begin{pmatrix} \sigma^2 & \omega\sigma\tau \\ \omega\sigma\tau & \tau^2 \end{pmatrix}\,,$$

recall the conditions on (ω, σ, τ) for Σ to be a (nonsingular) covariance matrix. Under those conditions, derive the conditional distribution of x_2 given x_1.

The matrix Σ is a covariance matrix if

1. Σ is symmetric and this is the case;
2. Σ is semi-definite positive, i.e. , for every $\mathbf{x} \in \mathbb{R}^2$, $\mathbf{x}^\mathsf{T}\Sigma\mathbf{x} \geq 0$, or, for every (x_1, x_2),

$$\sigma^2 x_1^2 + 2\omega\sigma\tau x_1 x_2 + \tau^2 x_2^2 = (\sigma x_1 + \omega\tau x_2)^2 + \tau^2 x_2^2(1-\omega^2) \geq 0\,.$$

A necessary condition for Σ to be positive semi-definite is that $\det(\Sigma) = \sigma^2\tau^2(1-\omega^2) \geq 0$, which is equivalent to $|\omega| \leq 1$.

In that case, $\mathbf{x}^\mathsf{T}\Sigma\mathbf{x} \geq 0$. The matrix Σ is furthermore nonsingular if $\det(\Sigma) > 0$, which is equivalent to $|\omega| < 1$.

Under those conditions, the conditional distribution of x_2 given x_1 is defined by

$$\begin{aligned} f(x_2|x_1) &\propto \exp\left\{ (x_1-\mu_1 \quad x_2-\mu_2)\Sigma^{-1}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}/2 \right\} \\ &\propto \exp\left\{ (x_1-\mu_1 \quad x_2-\mu_2)\begin{pmatrix} \tau^2 & -\omega\sigma\tau \\ -\omega\sigma\tau & \sigma^2 \end{pmatrix}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}/2\det(\Sigma) \right\} \\ &\propto \exp\left\{ \sigma^2(x_2-\mu_2)^2 - 2\omega\sigma\tau(x_1-\mu_1)(x_2-\mu_2)/2\det(\Sigma) \right\} \\ &\propto \exp\left\{ \frac{\sigma^2}{\sigma^2\tau^2(1-\omega^2)}\left(x_2-\mu_2-\omega\tau\frac{x_1-\mu_1}{\sigma}\right)^2/2 \right\} \end{aligned}$$

Therefore,

$$x_2|x_1 \sim \mathcal{N}\left(\mu_2+\omega\tau\frac{x_1-\mu_1}{\sigma}, \tau^2(1-\omega^2)\right) .$$

Exercise 1.3 Test the `help()` command on the functions `seq()`, `sample()`, and `order()`. (*Hint:* start with `help()`.)

Just type

```
> help()
> help(seq)
> help(sample)
> help(order)
```

and try illustrations like

```
> x=seq(1,500,10)
> y=sample(x,10,rep=T)
> z=order(y)
```

Exercise 1.4 Study the properties of the R function `lm()` using simulated data as in

```
> x=rnorm(20)
> y=3*x+5+rnorm(20,sd=0.3)
> reslm=lm(y~x)
> summary(reslm)
```

Generating the normal vectors x and y and calling the linear regression function lm leads to

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7459 -0.2216  0.1535  0.2130  0.8989

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.02530    0.10283   48.87   <2e-16 ***
x            2.98314    0.09628   30.98   <2e-16 ***
---
Sig. code:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4098 on 18 degrees of freedom
Multiple R-Squared: 0.9816,     Adjusted R-squared: 0.9806
F-statistic:   960 on 1 and 18 DF,  p-value: < 2.2e-16
```

Therefore, in this experiment, the regression coefficients (α, β) in $\mathbb{E}[y|x] = \alpha + \beta x$ are estimated by maximum likelihood as $\hat{\alpha} = 2.98$ and $\hat{\beta} = 5.03$, while they are $\alpha = 3$ and $\beta = 5$ in the simulated dataset.

Exercise 1.5 Of the R functions you have met so far, check which ones are written in R by simply typing their name without parentheses, as in `mean` or `var`.

Since

```
> mean
function (x, ...)
UseMethod("mean")
<environment: namespace:base>
```

and

```
> var
function (x, y = NULL, na.rm = FALSE, use)
{
    if (missing(use))
        use <- if (na.rm)
            "complete.obs"
        else "all.obs"
    na.method <- pmatch(use, c("all.obs", "complete.obs",
```

```
"pairwise.complete.obs"))
    if (is.data.frame(x))
        x <- as.matrix(x)
    else stopifnot(is.atomic(x))
    if (is.data.frame(y))
        y <- as.matrix(y)
    else stopifnot(is.atomic(y))
    .Internal(cov(x, y, na.method, FALSE))
}
<environment: namespace:stats>
```

we can deduce that the first function is written in C, while the second function is written in R.

2

Normal Models

Exercise 2.1 Check your current knowledge of the normal $\mathcal{N}(\mu, \sigma^2)$ distribution by writing down its density function and computing its first four moments.

The density of the normal $\mathcal{N}(\mu, \sigma^2)$ distribution is given by

$$\varphi(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-(x-\mu)^2/2\sigma^2\right\}$$

and, if $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$\begin{aligned}
\mathbb{E}[X] &= \mu + \int_{-\infty}^{+\infty} \frac{x-\mu}{\sqrt{2\pi}\sigma} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} \mathrm{d}x \\
&= \mu + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} y \exp -y^2/2 \mathrm{d}y \\
&= \mu + \frac{\sigma}{\sqrt{2\pi}} [-\exp -y^2/2]_{y=-\infty}^{y=+\infty} \\
&= \mu \,,
\end{aligned}$$

then, using one integration by parts,

$$\begin{aligned}
\mathbb{E}[(X-\mu)^2] &= \int_{-\infty}^{+\infty} \frac{y^2}{\sqrt{2\pi}\sigma} \exp -y^2/2\sigma^2 \mathrm{d}y \\
&= \sigma^2 \int_{-\infty}^{+\infty} \frac{z^2}{\sqrt{2\pi}} \exp -z^2/2 \, \mathrm{d}z \\
&= \frac{\sigma^2}{\sqrt{2\pi}} [-z \exp -z^2/2]_{z=-\infty}^{z=+\infty} + \sigma^2 \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp -z^2/2 \, \mathrm{d}z \\
&= \sigma^2 \,,
\end{aligned}$$

exploiting the fact that $(x-\mu)^3 \exp\left\{-(x-\mu)^2/2\sigma^2\right\}$ is asymmetric wrt the vertical axis $x=\mu$,

$$\mathbb{E}[(X-\mu)^3] = \int_{-\infty}^{+\infty} \frac{y^3}{\sqrt{2\pi}\sigma} \exp -y^2/2\sigma^2 \mathrm{d}y = 0$$

and, using once again one integration by parts,

$$\begin{aligned}\mathbb{E}[(X-\mu)^4] &= \frac{\sigma^4}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} z^4 \exp -z^2/2 \,\mathrm{d}z \\ &= \frac{\sigma^4}{\sqrt{2\pi}} [-z^3 \exp -z^2/2]_{z=-\infty}^{z=+\infty} + \sigma^4 \int_{-\infty}^{+\infty} \frac{3z^2}{\sqrt{2\pi}} \exp -z^2/2 \,\mathrm{d}z \\ &= 3\sigma^4 \,.\end{aligned}$$

Thus, the four first (centered) moments of the normal $\mathcal{N}(\mu, \sigma^2)$ distribution are μ, σ^2, 0 and $3\sigma^4$.

Exercise 2.2 Before exiting to the next page, think of datasets that could be, or could not be, classified as normal. In each case, describe your reason for proposing or rejecting a normal modeling.

A good illustration of the opposition between normal and nonnormal modelings can be found in insurrance claims: for minor damages, the histogram of the data (or of its log-transform) is approximately normal, while, for the highest claims, the tail is very heavy and cannot be modeled by a normal distribution (but rather by an extreme value distribution). Take for instance `http://www.statsci.org/data/general/carinsuk.html`

Exercise 2.3 Reproduce the histogram of Figure 2.1 and the subsequent analysis conducted in this chapter for the relative changes in reported larcenies relative to the 1995 figures, using the 90cntycr.wk1 file available on the Webpage of the book.

A new datafile must be created out of the file 90cntycr.wk1. Then, plotting an histogram and doing inference on this dataset follows from the directions provided within the chapter.

Exercise 2.4 By creating random subwindows of the region plotted in Figure 2.2, represent the histograms of these subsamples and examine whether they strongly differ or not. Pay attention to the possible influence of the few "bright spots" on the image.

While a "random subwindow" is not anything clearly defined, we can create a 800×800 matrix by

```
> cmb=matrix(scan("CMBdata"),nrow=800)
```

and define random subregions by

```
  > cmb1=cmb[sample(1:100,1):sample(101:300,1),
sample(50:150,1):sample(401:600,1)]
  > cmb2=cmb[sample(701:750,1):sample(751:800,1),
sample(650:750,1):sample(751:800,1)]
```

Comparing the histograms can then be done as

```
  > hist(cmb1,proba=T,xlim=range(cmb))
  > par(new=T)
  > hist(cmb2,proba=T,xlim=range(cmb))
```

or, more elaborately, through nonparametric density estimates

```
  > cnp1=density(cmb1,ad=3)  # Smooth out the bandwith
  > cnp2=density(cmb2,ad=3)  # Smooth out the bandwith
  > plot(cnp1,xlim=range(cmb),type="l",lwd=2,col="tomato3",
main="comparison")
  > lines(cnp2,lty=5,col="steelblue4",lwd=2)
```

which leads to Figure 2.1. In that case, both subsamples are roughly normal but with different parameters.

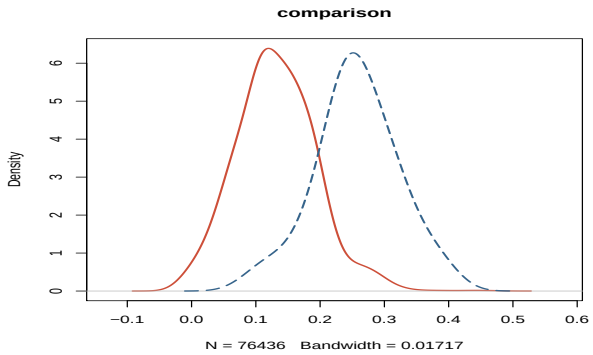

Fig. 2.1. Comparison of two density estimates for two random subregions.

Exercise 2.5 Show that (2.2) can be derived by first setting θ as a random variable with density function π and then $\mathscr{D}$ conditionally on θ as distributed from $\ell(\theta|\mathscr{D})$.

If $\pi(\theta)$ is the density of the *marginal* distribution of θ and then $\ell(\theta|\mathscr{D})$ the density of the conditional distribution of $\mathscr{D}$ given θ, the density of the joint distribution of θ and of $\mathscr{D}$ is given by

$$\ell(\theta|\mathscr{D})\pi(\theta)\,.$$

Therefore, Bayes's theorem simply is the derivation of the density of the conditional distribution of θ given $\mathscr{D}$ from this joint density.

Exercise 2.6 Show that the minimization (in $\hat{\theta}(\mathscr{D})$) of the expectation $\mathbb{E}[\mathsf{L}(\theta,\hat{\theta}))|\mathscr{D}]$—that is, of the expectation of the quadratic loss function under the distribution with density $\pi(\theta|\mathscr{D})$—produces the posterior expectation as the solution in $\hat{\theta}$.

Since

$$\begin{aligned}\mathbb{E}[\mathrm{L}(\theta,\hat{\theta}))|\mathscr{D}] &= \mathbb{E}[||\theta-\hat{\theta}||^2|\mathscr{D}]\\ &= \mathbb{E}[(\theta-\hat{\theta})^{\mathsf{T}}(\theta-\hat{\theta})|\mathscr{D}]\\ &= \mathbb{E}[||\theta||^2-2\theta^{\mathsf{T}}\hat{\theta}+||\hat{\theta}||^2|\mathscr{D}]\\ &= \mathbb{E}[||\theta||^2|\mathscr{D}]-2\hat{\theta}^{\mathsf{T}}\mathbb{E}[\theta|\mathscr{D}]+||\hat{\theta}||^2\\ &= \mathbb{E}[||\theta||^2|\mathscr{D}]-||\mathbb{E}[\theta|\mathscr{D}]||^2+||\mathbb{E}[\theta|\mathscr{D}]-\hat{\theta}||^2\,,\end{aligned}$$

minimising $\mathbb{E}[\mathrm{L}(\theta,\hat{\theta}))|\mathscr{D}]$ is equivalent to minimising $||\mathbb{E}[\theta|\mathscr{D}]-\hat{\theta}||^2$ and hence the solution is

$$\hat{\theta}=\mathbb{E}[\theta|\mathscr{D}]\,.$$

Exercise 2.7 Show that the normal, binomial, geometric, Poisson, and exponential distributions are all exponential families.

For each of those families of distributions, it is enough to achieve the standard form of exponential families

$$f_\theta(y)=h(y)\,\exp\left\{\theta\cdot R(y)-\Psi(\theta)\right\}\,, \tag{2.1}$$

as defined in the book.

In the normal $\mathscr{N}(\theta,1)$ case,

$$f_\theta(y)=\frac{1}{\sqrt{2\pi}}\,\exp\frac{1}{2}\,\left\{-y^2+2y\theta-\theta^2\right\}$$

and so it fits the representation (2.1) with $R(y)=y$, $h(y)=\exp(-y^2/2)/\sqrt{2\pi}$ and $\Psi(\theta)=\theta^2/2$.

In the binomial $\mathscr{B}(n,p)$ case,

$$f_p(y) = \binom{n}{y} \exp\{y\log(p) + (n-y)\log(1-p)\}\ , \quad y \in \{0,1,\ldots,n\}\,,$$

and it also fits the representation (2.1) with $\theta = \log(p/(1-p))$, $R(y) = y$, $h(y) = \binom{n}{y}$ and $\Psi(\theta) = -n\log(1+e^\theta)$.

In the geometric $\mathscr{G}(p)$ case [corresponding to the number of failures before a success],

$$f_p(y) = \exp\{y\log(1-p) + \log(p)\}\ , \quad y = 0,1,\ldots,$$

and it also fits the representation (2.1) with $\theta = \log(1-p)$, $R(y) = y$, $h(y) = 1$ and $\Psi(\theta) = -\log(1-e^\theta)$.

In the Poisson $\mathscr{P}(\lambda)$ case,

$$f_\lambda(y) = \frac{1}{y!}\exp\{y\log(\lambda) - \lambda\}$$

and it also fits the representation (2.1) with $\theta = \log(\lambda)$, $R(y) = y$, $h(y) = 1/y!$ and $\Psi(\theta) = \exp(\theta)$.

In the exponential $\mathscr{E}xp(\lambda)$ case,

$$f_\lambda(y) = \exp\{-\lambda y + \log(\lambda)\}$$

and it also fits the representation (2.1) with $\theta = \lambda$, $R(y) = -y$, $h(y) = 1$ and $\Psi(\theta) = -\log(\theta)$.

Exercise 2.8 Show that, for an exponential family, $\Psi(\theta)$ is defined by the constraint that f_θ is a probability density and that the expectation of this distribution can be written as $\partial\Psi(\theta)/\partial\theta$, the vector of the derivatives of $\Psi(\theta)$ with respect to the components of θ.

Using the representation (2.1),

$$\int f_\theta(y)\,\mathrm{d}y = \int h(y)\,\exp\{\theta\cdot R(y) - \Psi(\theta)\}\,\mathrm{d}y = 1$$

implies that $\Psi(\theta)$ is uniquely defined by

$$\int h(y)\,\exp\{\theta\cdot R(y)\}\,\mathrm{d}y = \int h(y)\,\mathrm{d}y \exp\{\Psi(\theta)\}\ .$$

When considering the expectation of $R(Y)$,

$$\begin{aligned}
\mathbb{E}_\theta[R(Y)] &= \int R(y)h(y)\,\exp\{\theta\cdot R(y)-\Psi(\theta)\}\,\mathrm{d}y \\
&= \int \frac{\partial}{\partial\theta}\{\theta\cdot R(y)\}\,h(y)\,\exp\{\theta\cdot R(y)-\Psi(\theta)\}\,\mathrm{d}y \\
&= \int \frac{\partial}{\partial\theta}\{\theta\cdot R(y)-\Psi(\theta)+\Psi(\theta)\}\,h(y)\,\exp\{\theta\cdot R(y)-\Psi(\theta)\}\,\mathrm{d}y \\
&= \frac{\partial\Psi(\theta)}{\partial\theta}\int h(y)\,\exp\{\theta\cdot R(y)-\Psi(\theta)\}\,\mathrm{d}y \\
&\quad + \int h(y)\,\frac{\partial}{\partial\theta}\left[\exp\{\theta\cdot R(y)-\Psi(\theta)\}\right]\mathrm{d}y \\
&= \frac{\partial\Psi(\theta)}{\partial\theta}\times 1 + \frac{\partial}{\partial\theta}\left\{\int h(y)\,\exp\{\theta\cdot R(y)-\Psi(\theta)\}\,\mathrm{d}y\right\} \\
&= \frac{\partial\Psi(\theta)}{\partial\theta}\,.
\end{aligned}$$

Exercise 2.9 Show that the updated hyperparameters in (2.5) are given by

$$\xi'(y)=\xi+R(y)\,,\quad \lambda'(y)=\lambda+1\,.$$

Find the corresponding expressions for $\pi(\theta|\xi,\lambda,y_1,\ldots,y_n)$.

If we start with

$$f_\theta(y)=h(y)\,\exp\{\theta\cdot R(y)-\Psi(\theta)\}\,,\quad\text{and}\quad \pi(\theta|\xi,\lambda)\propto\exp\{\theta\cdot\xi-\lambda\Psi(\theta)\}\,,$$

Bayes theorem implies that

$$\begin{aligned}
\pi(\theta|\xi,\lambda,y) &\propto f_\theta(y)\pi(\theta|\xi,\lambda) \\
&\propto \exp\{\theta\cdot R(y)-\Psi(\theta)\}\,\exp\{\theta\cdot\xi-\lambda\Psi(\theta)\} \\
&= \exp\{\theta\cdot[R(y)+\xi]-(\lambda+1)\Psi(\theta)\}\,.
\end{aligned}$$

Therefore,

$$\xi'(y)=\xi+R(y)\,,\quad \lambda'(y)=\lambda+1\,.$$

Similarly,

$$\begin{aligned}
\pi(\theta|\xi,\lambda,y_1,\ldots,y_n) &\propto \prod_{i=1}^n f_\theta(y_i)\pi(\theta|\xi,\lambda) \\
&\propto \exp\left\{\sum_{i=1}^n[\theta\cdot R(y_i)-\Psi(\theta)]\right\}\exp\{\theta\cdot\xi-\lambda\Psi(\theta)\} \\
&= \exp\left\{\theta\cdot[\sum_{i=1}^n R(y_i)+\xi]-(\lambda+n)\Psi(\theta)\right\}\,.
\end{aligned}$$

Therefore,

$$\xi'(y_1,\ldots,y_n) = \xi + \sum_{i=1}^{n} R(y_i)\,, \quad \lambda'(y_1,\ldots,y_n) = \lambda + n\,.$$

Exercise 2.10 erive the posterior distribution for an iid sample $\mathscr{D} = (y_1,\ldots,y_n)$ from $\mathscr{N}(\theta,1)$ and show that it only depends on the sufficient statistic $\overline{y} = \sum_{i=1}^{n} y_i/n$.

Since (see Exercice 2.7)

$$f_\theta(y) = \frac{1}{\sqrt{2\pi}}\,\exp\frac{1}{2}\,\left\{-y^2 + 2y\theta - \theta^2\right\}$$

fits the representation (2.1) with $R(y) = y$, $h(y) = \exp(-y^2/2)/\sqrt{2\pi}$ and $\Psi(\theta) = \theta^2/2$, a conjugate prior is

$$\pi(\theta|\xi,\lambda) \propto \exp\frac{1}{2}\,\left\{2\xi\theta - \lambda\theta^2\right\}\,,$$

which is equivalent to a $\mathscr{N}(\xi/\lambda, 1/\lambda)$ prior distribution. Following the updating formula given in Exercice 2.9, the posterior distribution is a

$$\mathscr{N}(\xi'(y_1,\ldots,y_n)/\lambda'(y_1,\ldots,y_n), 1/\lambda'(y_1,\ldots,y_n))$$

distribution, i.e.

$$\mu|y_1,\ldots,y_n \sim \mathscr{N}\left(\frac{\xi + n\overline{y}}{\lambda + n}, \frac{1}{\lambda + n}\right)\,.$$

It obviously only depends on the sufficient statistics $\overline{y}$.

Exercise 2.11 Give the range of values of the posterior mean (2.6) as the pair $(\lambda, \lambda^{-1}\xi)$ varies over $\mathbb{R}^+ \times \mathbb{R}$.

While

$$\frac{\lambda^{-1}}{1 + \lambda^{-1}} \leq 1\,,$$

the fact that ξ can take any value implies that this posterior mean has an unrestricted range, which can be seen as a drawback of this conjugate modeling.

Exercise 2.12 A Weibull distribution $\mathscr{W}(\alpha, \beta, \gamma)$ is defined as the power transform of a gamma $\mathscr{G}(\alpha, \beta)$ distribution: if $X \sim \mathscr{W}(\alpha, \beta, \gamma)$, then $X^\gamma \sim \mathscr{G}(\alpha, \beta)$. Show that, when γ is known, $\mathscr{W}(\alpha, \beta, \gamma)$ is an exponential family but that it is not an exponential family when γ is unknown.

The Weibull random variable X has the density

$$\frac{\gamma\alpha^\beta}{\Gamma(\beta)}\, x^{(\beta+1)\gamma-1}\, e^{-x^\gamma\alpha}\,,$$

since the Jacobian of the change of variables $y = x^\gamma$ is $\gamma x^{\gamma-1}$. So, checking the representation (2.1) leads to

$$f(x|\alpha,\beta,\gamma) = \frac{\gamma\alpha^\beta}{\Gamma(\beta)}\, \exp\left\{[(\beta+1)\gamma - 1]\log(x) - \alpha x^\gamma\right\}\,,$$

with $R(x) = (\gamma\log(x), -x^\gamma)$, $\theta = (\beta, \alpha)$ and $\Psi(\theta) = \log\Gamma(\beta) - \log\gamma\alpha^\beta$.

If γ is unknown, the term $x^\gamma\alpha$ in the exponential part makes it impossible to recover the representation (2.1).

Exercise 2.13 Show that, when the prior on $\theta = (\mu, \sigma^2)$ is $\mathscr{N}(\xi, \sigma^2/\lambda_\mu) \times \mathscr{IG}(\lambda_\sigma, \alpha)$, the marginal prior on μ is a Student's t distribution $\mathscr{T}(2\lambda_\sigma, \xi, \alpha/\lambda_\mu\lambda_\sigma)$ (see Example 2.3 below for the definition of a Student's t density). Give the corresponding marginal prior on σ^2. For an iid sample $\mathscr{D} = (x_1, \ldots, x_n)$ from $\mathscr{N}(\mu, \sigma^2)$, derive the parameters of the posterior distribution of (μ, σ^2).

Since the joint prior distribution of (μ, σ^2) is

$$\pi(\mu,\sigma^2) \propto (\sigma^2)^{-\lambda_\sigma-1-1/2}\, \exp\frac{-1}{2\sigma^2}\left\{\lambda_\mu(\mu-\xi)^2 + 2\alpha\right\}$$

(given that the Jacobian of the change of variable $\omega = \sigma^{-2}$ is ω^{-2}), integrating out σ^2 leads to

$$\begin{aligned}
\pi(\mu) &\propto \int_0^\infty (\sigma^2)^{-\lambda_\sigma-3/2}\, \exp\frac{-1}{2\sigma^2}\left\{\lambda_\mu(\mu-\xi)^2 + 2\alpha\right\}\, \mathrm{d}\sigma^2 \\
&\propto \int_0^\infty \omega^{\lambda_\sigma-1/2}\, \exp\frac{-\omega}{2}\left\{\lambda_\mu(\mu-\xi)^2 + 2\alpha\right\}\, \mathrm{d}\omega \\
&\propto \left\{\lambda_\mu(\mu-\xi)^2 + 2\alpha\right\}^{-\lambda_\sigma-1/2} \\
&\propto \left\{1 + \frac{\lambda_\sigma\lambda_\mu(\mu-\xi)^2}{2\lambda_\sigma\alpha}\right\}^{-\frac{2\lambda_\sigma+1}{2}}\,,
\end{aligned}$$

which is the proper density of a Student's t distribution $\mathcal{T}(2\lambda_\sigma, \xi, \alpha/\lambda_\mu\lambda_\sigma)$.

By definition of the joint prior on (μ, σ^2), the marginal prior on σ^2 is a inverse gamma $\mathcal{IG}(\lambda_\sigma, \alpha)$ distribution.

The joint posterior distribution of (μ, σ^2) is

$$\pi((\mu,\sigma^2)|\mathscr{D}) \propto (\sigma^2)^{-\lambda_\sigma(\mathscr{D})} \exp\left\{-\left(\lambda_\mu(\mathscr{D})(\mu-\xi(\mathscr{D}))^2+\alpha(\mathscr{D})\right)/2\sigma^2\right\},$$

with

$$\begin{aligned}
\lambda_\sigma(\mathscr{D}) &= \lambda_\sigma + 3/2 + n/2\,,\\
\lambda_\mu(\mathscr{D}) &= \lambda_\mu + n\,,\\
\xi(\mathscr{D}) &= (\lambda_\mu\xi + n\overline{x})/\lambda_\mu(\mathscr{D})\,,\\
\alpha(\mathscr{D}) &= 2\alpha + \frac{\lambda_\mu(\mathscr{D})}{n\lambda_\mu}(\overline{x}-\xi)^2 + s^2(\mathscr{D})\,.
\end{aligned}$$

This is the product of a marginal inverse gamma

$$\mathcal{IG}\left(\lambda_\sigma(\mathscr{D}) - 3/2, \alpha(\mathscr{D})/2\right)$$

distribution on σ^2 by a conditional normal

$$\mathcal{N}\left(\xi(\mathscr{D}), \sigma^2/\lambda_\mu(\mathscr{D})\right)$$

on μ. (Hence, we do get a conjugate prior.) Integrating out σ^2 leads to

$$\begin{aligned}
\pi(\mu|\mathscr{D}) &\propto \int_0^\infty (\sigma^2)^{-\lambda_\sigma(\mathscr{D})} \exp\left\{-\left(\lambda_\mu(\mathscr{D})(\mu-\xi(\mathscr{D}))^2+\alpha(\mathscr{D})\right)/2\sigma^2\right\} \mathrm{d}\sigma^2\\
&\propto \int_0^\infty \omega^{\lambda_\sigma(\mathscr{D})-2} \exp\left\{-\left(\lambda_\mu(\mathscr{D})(\mu-\xi(\mathscr{D}))^2+\alpha(\mathscr{D})\right)\omega/2\right\} \mathrm{d}\omega\\
&\propto \left(\lambda_\mu(\mathscr{D})(\mu-\xi(\mathscr{D}))^2+\alpha(\mathscr{D})\right)^{-(\lambda_\sigma(\mathscr{D})-1)},
\end{aligned}$$

which is the generic form of a Student's t distribution.

Exercise 2.14 Show that, for location and scale models, Jeffreys' prior is given by $\pi^J(\theta) = 1$ and $\pi^J(\theta) = 1/\theta$, respectively.

In the case of a location model, $f(y|\theta) = p(y-\theta)$, the Fisher information matrix of a location model is given by

$$\begin{aligned}
I(\theta) &= \mathbb{E}_\theta\left[\frac{\partial \log p(Y-\theta)}{\partial\theta}^{\mathsf{T}} \frac{\partial \log p(Y-\theta)}{\partial\theta}\right]\\
&= \int \left[\frac{\partial p(y-\theta)}{\partial\theta}\right]^{\mathsf{T}} \left[\frac{\partial p(y-\theta)}{\partial\theta}\right] / p(y-\theta)\,\mathrm{d}y\\
&= \int \left[\frac{\partial p(z)}{\partial z}\right]^{\mathsf{T}} \left[\frac{\partial p(z)}{\partial z}\right] / p(z)\,\mathrm{d}z
\end{aligned}$$

it is indeed constant in θ. Therefore the determinant of $I(\theta)$ is also constant and Jeffreys' prior can be chosen as $\pi^J(\theta) = 1$ [or any other constant as long as the parameter space is not compact].

In the case of a scale model, if $y \sim f(y/\theta)/\theta$, a change of variable from y to $z = \log(y)$ [if $y > 0$] implies that $\eta = \log(\theta)$ is a location parameter for z. Therefore, the Jacobian transform of $\pi^J(\eta) = 1$ is $\pi^J(\theta) = 1/\theta$. When y can take both negative and positive values, a transform of y into $z = \log(|y|)$ leads to the same result.

Exercise 2.15 In the case of an exponential family, derive Jeffreys' prior in terms of the Hessian matrix of $\Psi(\theta)$, i.e. the matrix of second derivatives of $\Psi(\theta)$.

Using the representation (2.1)

$$\log f_\theta(y) = \log h(y) + \theta \cdot R(y) - \Psi(\theta)\,,$$

we get

$$\frac{\partial^2}{\partial\theta\partial\theta^\mathsf{T}} \log f_\theta(y) = -\frac{\partial^2\Psi(\theta)}{\partial\theta\partial\theta^\mathsf{T}}$$

and therefore the Fisher information matrix is the Hessian matrix of $\Psi(\theta)$, $H(\theta)$. This implies that $\pi^J(\theta) = \det H(\theta)$.

Exercise 2.16 Show that, when $\pi(\theta)$ is a probability density, (2.8) necessarily holds for all datasets $\mathscr{D}$.

Given that $\pi(\theta)$ is a (true) probability density and that the likelihood $\ell(\theta|\mathscr{D})$ is also a (true) probability density in $\mathscr{D}$ that can be interpreted as a conditional density, the product

$$\pi(\theta)\ell(\theta|\mathscr{D})$$

is a true joint probability density for $(\theta, \mathscr{D})$. The above integral therefore defines the marginal density of $\mathscr{D}$, which is always defined.

Exercise 2.17 Try to devise a parameterized model and an improper prior such that, no matter the sample size, the posterior distribution does not exist. (If you cannot find such a case, wait until Chapter 6.)

It is sufficient to find a function of the parameter θ that goes to infinity faster than the likelihood goes to 0, no matter what the sample size is. For

instance, take $\pi(\theta) \propto \exp \theta^2$ for a Cauchy $\mathscr{C}(\theta, 1)$ model. Then, for a sample of size n, the likelihood goes to 0 as θ^{-2n} and it cannot beat the exponential increase in the prior.

Exercise 2.18 Show that, under the loss L_{a_0,a_1}, the Bayes estimator associated with a prior π is given by

$$\delta^\pi(x) = \begin{cases} 1 & \text{if} \quad P^\pi(\theta \in \Theta_0|x) > a_1/a_0 + a_1, \\ 0 & \text{otherwise.} \end{cases}$$

The posterior expected loss is

$$\mathbb{E}\left[L_{a_0,a_1}(\theta, d)|x\right] = \begin{cases} a_0\, P^\pi(\theta \in \Theta_0|x) & \text{if} \quad d = 0\,, \\ a_1\, P^\pi(\theta \in \Theta_1|x) & \text{if} \quad d = 1\,, \end{cases}$$

thus the decision minimising this posterior loss is $d = 1$ when $a_1\, P^\pi(\theta \in \Theta_1|x) < a_0\, P^\pi(\theta \in \Theta_0|x)$, i.e.

$$a_1(1 - P^\pi(\theta \in \Theta_0|x)) < a_0\, P^\pi(\theta \in \Theta_0|x)\,,$$

and $d = 0$ otherwise.

Exercise 2.19 When $\theta \in \{\theta_0, \theta_1\}$, show that the Bayesian procedure only depends on the ratio $\varrho_0 f_{\theta_0}(x)/(1 - \varrho_0) f_{\theta_1}(x)$, where ϱ_0 is the prior weight on θ_0.

In this special case, π puts a point mass of ϱ_0 on θ_0 and of $(1 - \varrho_0)$ on θ_1. Therefore,

$$\begin{aligned} P^\pi(\theta = \theta_0|x) &= \frac{\varrho_0\, f_{\theta_0}(x)}{\varrho_0\, f_{\theta_0}(x) + (1 - \varrho_0)\, f_{\theta_1}(x)} \\ &= \frac{1}{1 + (1 - \varrho_0)\, f_{\theta_1}(x)/(1 - \varrho_0)\, f_{\theta_1}(x)}\,, \end{aligned}$$

which only depends on the ratio $\varrho_0 f_{\theta_0}(x)/(1 - \varrho_0) f_{\theta_1}(x)$.

Exercise 2.20 Show that the limit of the posterior probability $P^\pi(\mu < 0|x)$ when ξ goes to 0 and τ goes to ∞ is $\Phi(-x/\sigma)$.

Since

$$\begin{aligned} P^\pi(\mu<0|x) &= \Phi\left(-\xi(x)/\omega\right) \\ &= \Phi\left(\frac{\sigma^2\xi+\tau^2 x}{\sigma^2+\tau^2}\sqrt{\frac{\sigma^2+\tau^2}{\sigma^2\tau^2}}\right) \\ &= \Phi\left(\frac{\sigma^2\xi+\tau^2 x}{\sqrt{\sigma^2+\tau^2}\sqrt{\sigma^2\tau^2}}\right), \end{aligned}$$

when ξ goes to 0 and τ goes to ∞, the ratio

$$\frac{\sigma^2\xi+\tau^2 x}{\sqrt{\sigma^2+\tau^2}\sqrt{\sigma^2\tau^2}}$$

goes to

$$\lim_{\tau\to\infty}\frac{\tau^2 x}{\sqrt{\sigma^2+\tau^2}\sqrt{\sigma^2\tau^2}} = \lim_{\tau\to\infty}\frac{\tau^2 x}{\tau^2\sigma} = \frac{x}{\sigma}\,.$$

Exercise 2.21 We recall that the normalizing constant for a Student's $\mathcal{T}(\nu,\mu,\sigma^2)$ distribution is

$$\frac{\Gamma((\nu+1)/2)/\Gamma(\nu/2)}{\sigma\sqrt{\nu\pi}}\,.$$

Give the value of the integral in the denominator of B^π_{10} above.

We have

$$(\mu-\bar{x})^2+(\mu-\bar{y})^2 = 2\left(\mu-\frac{\bar{x}+\bar{y}}{2}\right)^2+\frac{(\bar{x}-\bar{y})^2}{2}$$

and thus

$$\begin{aligned} &\int\left[(\mu-\bar{x})^2+(\mu-\bar{y})^2+S^2\right]^{-n}\mathrm{d}\mu \\ &\quad= 2^{-n}\int\left[\left(\mu-\frac{\bar{x}+\bar{y}}{2}\right)^2+\frac{(\bar{x}-\bar{y})^2}{4}+\frac{S^2}{2}\right]^{-n}\mathrm{d}\mu \\ &\quad= (2\sigma^2)^{-n}\int\left[1+\left(\mu-\frac{\bar{x}+\bar{y}}{2}\right)^2\Big/\sigma^2\nu\right]^{-(\nu+1)/2}\mathrm{d}\mu\,, \end{aligned}$$

where $\nu=2n-1$ and

$$\sigma^2 = \left[\left(\frac{\bar{x}-\bar{y}}{2}\right)^2+\frac{S^2}{2}\right]\Big/(2n-1)\,.$$

Therefore,

$$\begin{aligned}\int &\left[(\mu-\bar{x})^2+(\mu-\bar{y})^2+S^2\right]^{-n}\,\mathrm{d}\mu\\ &= (2\sigma^2)^{-n}\,\frac{\sigma\sqrt{\nu\pi}}{\Gamma((\nu+1)/2)/\Gamma(\nu/2)}\\ &= \frac{\sqrt{\nu\pi}}{2^n\sigma^{2n-1}\Gamma((\nu+1)/2)/\Gamma(\nu/2)}\\ &= \frac{(2n-1)^{2n-1}\sqrt{\nu\pi}}{2^n\left[\left(\frac{\bar{x}-\bar{y}}{2}\right)^2+\frac{S^2}{2}\right]^{2n-1}\Gamma((\nu+1)/2)/\Gamma(\nu/2)}\,.\end{aligned}$$

Note that this expression is used later in the simplified devivation of B_{01}^{π} without the term $(2n-1)^{2n-1}\sqrt{\nu\pi}/2^n\Gamma((\nu+1)/2)/\Gamma(\nu/2)$ because this term appears in both the numerator and the denominator.

Exercise 2.22 Approximate B_{01}^{π} by a Monte Carlo experiment where ξ is simulated from a Student's t distribution with mean $(\bar{x}+\bar{y})/2$ and appropriate variance, and the integrand is proportional to $\exp -\xi^2/2$. Compare the precision of the resulting estimator with the above Monte Carlo approximation based on the normal simulation.

The integral of interest in B_{01}^{π} is

$$\begin{aligned}\int &\left[(2\xi+\bar{x}-\bar{y})^2/2+S^2\right]^{-n+1/2}\,e^{-\xi^2/2}\,\mathrm{d}\xi/\sqrt{2\pi}\\ &= \left(S^2\right)^{-n+1/2}\int\frac{\exp-\xi^2/2}{\sqrt{2\pi}}\left[\frac{4(n-1)(\xi-(\bar{y}-\bar{x})/2)^2}{(2n-2)S^2}+1\right]^{-n+1/2}\mathrm{d}\xi\\ &= \left(S^2\right)^{-n+1/2}\int\frac{\exp-\xi^2/2}{\sqrt{2\pi}}\left[\frac{(\xi-(\bar{y}-\bar{x})/2)^2}{\nu\sigma^2}+1\right]^{-(\nu+1)/2}\mathrm{d}\xi\\ &= C\int\frac{\exp-\xi^2/2}{\sqrt{2\pi}}\,\mathfrak{t}(\xi|\mu,\sigma,\nu)\,\mathrm{d}\xi\,,\end{aligned}$$

where $\mathfrak{t}(\xi|\mu,\sigma,\nu)$ is the density of the Student's $\mathcal{T}(\nu,\mu,\sigma^2)$ distribution with parameters $\nu=2n-2$, $\mu=(\bar{y}-\bar{x})/2$, and $\sigma^2=S^2/4(n-1)$, and C is the constant

$$C=\left(S^2\right)^{-n+1/2}\Big/\frac{\Gamma((\nu+1)/2)/\Gamma(\nu/2)}{\sigma\sqrt{\nu\pi}}\,.$$

Therefore, we can simulate a sample $\xi_1,\ldots,\xi_n$ from the $\mathcal{T}(\nu,\mu,\sigma^2)$ distribution and approximate the above integral by the average

$$\frac{C}{n}\sum_{i=1}^{n}\frac{\exp-\xi_i^2/2}{\sqrt{2\pi}}\,,$$

using an R program like

```
n=100
N=1000000
nu=2*n-2
barx=.088
bary=.1078
mu=.5*(bary-barx)
stwo=.00875
sigma=sqrt(.5*stwo/nu)
C=log(stwo)*(-n+.5)+log(sigma*sqrt(nu*pi))+
lgamma(.5*nu)-lgamma(.5*nu+.5)

# T simulation
xis=rt(n=N,df=nu)*sigma + mu
B01=-log(cumsum(dnorm(xis))/(1:N))
B01=exp( (-n+.5)*log(.5*(barx-bary)^2+stwo)+B01-C )

# Normal simulation
xis=rnorm(N)
C01=cumsum((stwo+.5*(2*xis+barx-bary)^2)^(-n+.5))/(1:N)
C01=((.5*(barx-bary)^2+stwo)^(-n+.5))/C01

# Comparison of the cumulated averages
plot(C01[seq(1,N,l=1000)],type="l",col="tomato2",lwd=2,
    ylim=c(20,30),xlab=expression(N/100),ylab=expression(1/B[10]))
lines(B01[seq(1,N,l=1000)],col="steelblue3",lwd=2,lty=5)
```

As shown on Figure 2.2, the precision of the estimator based on the $\mathcal{T}(\nu, \mu, \sigma^2)$ simulation is immensely superior to the precision of the estimator based on a normal sample: they both converge to the same value 23.42, but with very different variances.

Exercise 2.23 Discuss what happens to the importance sampling approximation when the support of g is larger than the support of γ.

If the support of γ, $\mathfrak{S}_\gamma$, is smaller than the support of g, the representation

$$\mathfrak{I} = \int \frac{h(x)g(x)}{\gamma(x)}\,\gamma(x)\,\mathrm{d}x$$

is not valid and the importance sampling approximation evaluates instead the integral

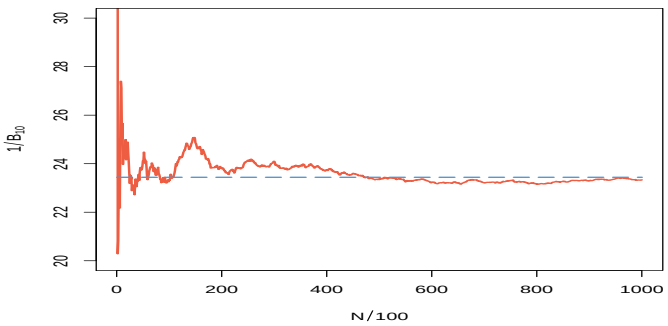

Fig. 2.2. Comparison of two approximations of the Bayes factor B_{01} based on 10^6 simulations.

$$\int_{\mathfrak{S}_\gamma} \frac{h(x)g(x)}{\gamma(x)}\,\gamma(x)\,\mathrm{d}x.$$

Exercise 2.24 Show that the importance weights of Example 2.2 have infinite variance.

The importance weight is

$$\exp\left\{(\theta-\mu)^2/2\right\}\prod_{i=1}^{n}[1+(x_i-\theta)^2]^{-1}$$

with $\theta \sim \mathcal{N}(\mu,\sigma^2)$. While its expectation is finite—it would be equal to 1 were we to use the right normalising constants—, the expectation of its square is not:

$$\int \exp\left\{(\theta-\mu)^2/2\right\}\prod_{i=1}^{n}[1+(x_i-\theta)^2]^{-2}\,\mathrm{d}\theta = \infty\,,$$

due to the dominance of the exponential term over the polynomial term.

Exercise 2.25 Show that, when γ is the normal $\mathcal{N}(0,\nu/(\nu-2))$ density, the ratio

$$\frac{f_\nu^2(x)}{\gamma(x)} \propto \frac{e^{x^2(\nu-2)/2\nu}}{[1+x^2/\nu]^{(\nu+1)}}$$

does not have a finite integral. What does this imply about the variance of the importance weights?

This is more or less a special case of Exercise 2.24, with again the exponential term dominating the polynomial term, no matter what the value of $\nu > 2$ is. The importance weights have no variance. When running an experiment like the following one

```
nu=c(3,5,10,100,1000,10000)
N=length(nu)
T=1000

nors=rnorm(T)
par(mfrow=c(2,N/2),mar=c(4,2,4,1))

for (nnu in nu){

   y=sqrt(nnu/(nnu-2))*nors
   isw=dt(y,df=nnu)/dnorm(y)

   hist(log(isw),prob=T,col="wheat4",nclass=T/20)
}
```

the output in Figure 2.3 shows that the value of ν still matters very much in the distribution of the weights. When ν is small, the probability to get very large weights is much higher than with large ν's, and the dispersion decreases with ν.

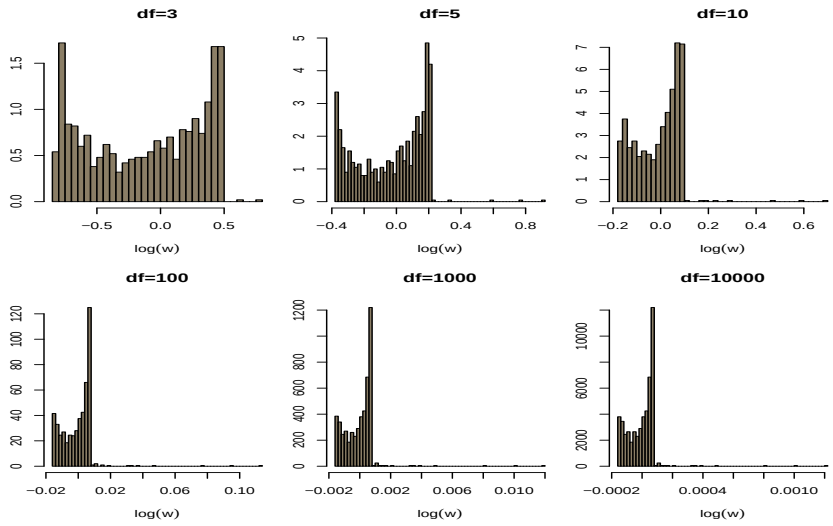

Fig. 2.3. Distributions of the log-importance weights for a normal importance distribution against a Student's t target for several values of ν.

Exercise 2.26 Given two model densities $f_1(\mathscr{D}|\theta)$ and $f_2(\mathscr{D}|\theta)$ with the same parameter θ and corresponding priors densities $\pi_1(\theta)$ and $\pi_2(\theta)$, denote $\tilde{\pi}_1(\theta|\mathscr{D}) = f_1(\mathscr{D}|\theta)\pi_1(\theta)$ and $\tilde{\pi}_2(\theta|\mathscr{D}) = f_2(\mathscr{D}|\theta)\pi_2(\theta)$, and show that the Bayes factor corresponding to the comparison of both models satisfies

$$B_{12}^{\pi} = \frac{\int \tilde{\pi}_1(\theta|\mathscr{D})\alpha(\theta)\pi_2(\theta|\mathscr{D})\mathrm{d}\theta}{\int \tilde{\pi}_2(\theta|\mathscr{D})\alpha(\theta)\pi_1(\theta|\mathscr{D})\mathrm{d}\theta}$$

for every positive function α and deduce that

$$n_1 \sum_{i=1}^{n_2} \tilde{\pi}_1(\theta_{2i}|\mathscr{D})\alpha(\theta_{2i}) \Big/ n_2 \sum_{i=1}^{n_1} \tilde{\pi}_2(\theta_{1i}|\mathscr{D})\alpha(\theta_{1i})$$

is a convergent approximation of the Bayes factor B_{12}^{π} when $\theta_{ji} \sim \pi_j(\theta|\mathscr{D})$ $(i = 1, 2,\ j = 1, \ldots, n_j)$.

The missing normalising constants in $\tilde{\pi}_1(\theta|\mathscr{D})$ and $\tilde{\pi}_2(\theta|\mathscr{D})$ are the marginal densities $m_1(\mathscr{D})$ and $m_2(\mathscr{D})$, in the sense that $(i = 1, 2)$

$$\pi_i(\theta|\mathscr{D}) = \tilde{\pi}_i(\theta|\mathscr{D})/m_i(\mathscr{D})\,.$$

Therefore,

$$\begin{aligned}
&\frac{\int \tilde{\pi}_1(\theta|\mathscr{D})\alpha(\theta)\pi_2(\theta|\mathscr{D})\mathrm{d}\theta}{\int \tilde{\pi}_2(\theta|\mathscr{D})\alpha(\theta)\pi_1(\theta|\mathscr{D})\mathrm{d}\theta} \\
&= \frac{\int m_1(\mathscr{D})\pi_1(\theta|\mathscr{D})\alpha(\theta)\pi_2(\theta|\mathscr{D})\mathrm{d}\theta}{\int m_2(\mathscr{D})pi_1(\theta|\mathscr{D})\alpha(\theta)\pi_2(\theta|\mathscr{D})\mathrm{d}\theta} \\
&= \frac{m_1(\mathscr{D})}{m_2(\mathscr{D})} = B_{12}^{\pi}
\end{aligned}$$

and α is irrelevant for the computation of the ratio of integrals.

A Monte Carlo implementation of this remark is to represent each integral in the ratio as an expectation under $\pi_2(\theta|\mathscr{D})$ and $\pi_1(\theta|\mathscr{D})$ respectively. Simulations θ_{ji}'s from both posteriors then produce convergent estimators of the corresponding integrals. This method is called *bridge sampling* and the choice of α is relevant in the variance of the corresponding estimator.

Exercise 2.27 Show that, when n goes to infinity and when the prior has an unlimited support, the predictive distribution converges to the exact (sampling) distribution of x_{n+1}.

This property follows from the fact that the posterior distribution converges to a Dirac mass at the true value $\theta^\star$ of the parameter when n goes to infinity [under some regularity conditions on both π and $f(x|\theta)$, as well as identifiability constraints]. Therefore,

$$\int f(x_{n+1}|\theta)\pi(\theta|\mathscr{D}_n)\,\mathrm{d}\theta$$

converges to $f(x_{n+1}|\theta^\star)$.

Exercise 2.28 Show that, when X is distributed from an increasing and continuous cdf F, $F(X)$ has a uniform distribution.

If F is increasing and continuous, it is invertible and we have

$$P(F(X) \le u) = P(X \le F^{-1}(u)) = F(F^{-1}(u)) = u\,,$$

when $0 \le u \le 1$. This demonstrates that $F(X)$ is uniformely distributed and this property can be exploited for simulation purposes when F is available in closed form.

3

Regression and Variable Selection

Exercise 3.1 Show that the matrix X is of full rank if and only if the matrix $X^\mathsf{T}X$ is invertible (where X^T denotes the transpose of the matrix X, which can produced in R using the `t(X)` command). Deduce that this cannot happen when $k+1>n$.

The matrix X is a $(n, k+1)$ matrix. It is of full rank if the $k+1$ columns of X induce a subspace of $\mathbb{R}^n$ of dimension $(k+1)$, or, in other words, if those columns are linearly independent: there exists no solution to $X\gamma = \mathbf{0}_n$ other than $\gamma = \mathbf{0}_n$, where $\mathbf{0}_{k+1}$ denotes the $(k+1)$-dimensional vector made of 0's. If $X^\mathsf{T}X$ is invertible, then $X\gamma = \mathbf{0}_n$ implies $X^\mathsf{T}X\gamma = X^\mathsf{T}\mathbf{0}_n = \mathbf{0}_{k+1}$ and thus $\gamma = (X^\mathsf{T}X)^{-1}\mathbf{0}_{k+1} = \mathbf{0}_{k+1}$, therefore X is of full rank. If $X^\mathsf{T}X$ is not invertible, there exist vectors β and $\gamma \neq \beta$ such that $X^\mathsf{T}X\beta = X^\mathsf{T}X\gamma$, i.e. $X^\mathsf{T}X(\beta-\gamma) = \mathbf{0}_{k+1}$. This implies that $||X(\beta-\gamma)||^2 = 0$ and hence $X(\beta-\gamma) = \mathbf{0}_n$ for $\beta-\gamma \neq \mathbf{0}_{k+1}$, thus X is not of full rank.

Obviously, the matrix $(k+1, k+1)$ matrix $X^\mathsf{T}X$ cannot be invertible if $k+1>n$ since the columns of X are then necessarily linearly dependent.

Exercise 3.2 Show that solving the minimization program above requires solving the system of equations $(X^\mathsf{T}X)\beta = X^\mathsf{T}\mathbf{y}$. Check that this can be done via the R command

```
> solve(t(X)%*%(X),t(X)%*%y)
```

If we decompose $(\mathbf{y}-X\beta)^\mathsf{T}(\mathbf{y}-X\beta)$ as

$$\mathbf{y}^\mathsf{T}\mathbf{y} - 2\mathbf{y}^\mathsf{T}X\beta + \beta^\mathsf{T}X^\mathsf{T}X\beta$$

and differentiate this expression in β, we obtain the equation

$$-2\mathbf{y}^\mathsf{T}X + 2\beta^\mathsf{T}X^\mathsf{T}X = \mathbf{0}_{k+1}\,,$$

i.e.

$$(X^\mathsf{T}X)\beta = X^\mathsf{T}\mathbf{y}$$

by transposing the above.

As can be checked via `help(solve)`, `solve(A,b)` is the R function that solves the linear equation system $Ax = b$. Defining X and y from caterpillar, we get

```
> solve(t(X)%*%X,t(X)%*%y)

                     [,1]
rep(1, 33) 10.998412367
V1         -0.004430805
V2         -0.053830053
V3          0.067939357
V4         -1.293636435
V5          0.231636755
V6         -0.356799738
V7         -0.237469094
V8          0.181060170
V9         -1.285316143
V10        -0.433105521
```

which [obviously] gives the same result as the call to the linear regression function lm():

```
> lm(y~X-1)

Call:
lm(formula = y ~ X - 1)

Coefficients:
Xrep(1, 33)     XV1       XV2       XV3       XV4         XV5
 10.998412  -0.004431  -0.053830   0.067939  -1.29363   0.23163
      XV6       XV7       XV8       XV9       XV10
 -0.356800  -0.237469   0.181060  -1.285316  -0.43310
```

Note the use of the `-1` in the formula `y~X-1` that eliminates the intercept already contained in X.

Exercise 3.3 Show that $\mathbb{V}(\hat{\beta}|\sigma^2, X) = \sigma^2(X^\mathsf{T}X)^{-1}$.

Since $\hat{\beta} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}\mathbf{y}$ is a linear transform of $\mathbf{y} \sim \mathcal{N}(X\beta, \sigma^2 I_n)$, we have

$$\hat{\beta} \sim \mathcal{N}\left((X^\mathsf{T}X)^{-1}X^\mathsf{T}X\beta, \sigma^2(X^\mathsf{T}X)^{-1}X^\mathsf{T}X(X^\mathsf{T}X)^{-1}\right) ,$$

i.e.

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2(X^\mathsf{T}X)^{-1}\right) .$$

Exercise 3.4 Taking advantage of the matrix identities

$$\begin{aligned}\left(M + X^\mathsf{T}X\right)^{-1} &= M^{-1} - M^{-1}\left(M^{-1} + (X^\mathsf{T}X)^{-1}\right)^{-1}M^{-1} \\ &= (X^\mathsf{T}X)^{-1} - (X^\mathsf{T}X)^{-1}\left(M^{-1} + (X^\mathsf{T}X)^{-1}\right)^{-1}(X^\mathsf{T}X)^{-1}\end{aligned}$$

and

$$\begin{aligned}X^\mathsf{T}X(M + X^\mathsf{T}X)^{-1}M &= \left(M^{-1}(M + X^\mathsf{T}X)(X^\mathsf{T}X)^{-1}\right)^{-1} \\ &= \left(M^{-1} + (X^\mathsf{T}X)^{-1}\right)^{-1} ,\end{aligned}$$

establish that (3.3) and (3.4) are the correct posterior distributions.

Starting from the prior distribution

$$\beta|\sigma^2, X \sim \mathcal{N}_{k+1}(\tilde{\beta}, \sigma^2 M^{-1}) , \quad \sigma^2|X \sim \mathcal{IG}(a, b) ,$$

the posterior distribution is

$$\begin{aligned}\pi(\beta, \sigma^2|\hat{\beta}, s^2, X) &\propto \sigma^{-k-1-2a-2-n} \exp\frac{-1}{2\sigma^2}\Big\{(\beta - \tilde{\beta})^\mathsf{T}M(\beta - \tilde{\beta}) \\ &\quad +(\beta - \hat{\beta})^\mathsf{T}(X^\mathsf{T}X)(\beta - \hat{\beta}) + s^2 + 2b\Big\} \\ &= \sigma^{-k-n-2a-3} \exp\frac{-1}{2\sigma^2}\Big\{\beta^\mathsf{T}(M + X^\mathsf{T}X)\beta - 2\beta^\mathsf{T}(M\tilde{\beta} + X^\mathsf{T}X\hat{\beta}) \\ &\quad +\tilde{\beta}^\mathsf{T}M\tilde{\beta} + \hat{\beta}^\mathsf{T}(X^\mathsf{T}X)\hat{\beta} + s^2 + 2b\Big\} \\ &= \sigma^{-k-n-2a-3} \exp\frac{-1}{2\sigma^2}\left\{(\beta - \mathbb{E}[\beta|y, X])^\mathsf{T}(M + X^\mathsf{T}X)(\beta - \mathbb{E}[\beta|y, X])\right. \\ &\quad \left.+\beta^\mathsf{T}M\tilde{\beta} + \hat{\beta}^\mathsf{T}(X^\mathsf{T}X)\hat{\beta} - \mathbb{E}[\beta|y, X]^\mathsf{T}(M + X^\mathsf{T}X)\mathbb{E}[\beta|y, X] + s^2 + 2b\right\}\end{aligned}$$

with

$$\mathbb{E}[\beta|y, X] = (M + X^\mathsf{T}X)^{-1}(M\tilde{\beta} + X^\mathsf{T}X\hat{\beta}) .$$

Therefore, (3.3) is the conditional posterior distribution of β given σ^2. Integrating out β leads to

$$\begin{aligned}\pi(\sigma^2|\hat{\beta}, s^2, X) &\propto \sigma^{-n-2a-2} \exp\frac{-1}{2\sigma^2}\left\{\beta^{\mathsf{T}} M\tilde{\beta} + \hat{\beta}^{\mathsf{T}}(X^{\mathsf{T}}X)\hat{\beta}\right.\\ &\quad -\mathbb{E}[\beta|y,X]^{\mathsf{T}}(M+X^{\mathsf{T}}X)\mathbb{E}[\beta|y,X] + s^2 + 2b\}\\ &= \sigma^{-n-2a-2}\exp\frac{-1}{2\sigma^2}\left\{\beta^{\mathsf{T}}M\tilde{\beta} + \hat{\beta}^{\mathsf{T}}(X^{\mathsf{T}}X)\hat{\beta} + s^2 + 2b\right.\\ &\quad \left.-(M\tilde{\beta}+X^{\mathsf{T}}X\hat{\beta})^{\mathsf{T}}(M+X^{\mathsf{T}}X)^{-1}(M\tilde{\beta}+X^{\mathsf{T}}X\hat{\beta})\right\}\end{aligned}$$

Using the first matrix identity, we get that

$$\begin{aligned}(M\tilde{\beta}+X^{\mathsf{T}}X\hat{\beta})^{\mathsf{T}}&\left(M+X^{\mathsf{T}}X\right)^{-1}(M\tilde{\beta}+X^{\mathsf{T}}X\hat{\beta})\\ &= \tilde{\beta}^{\mathsf{T}}M\tilde{\beta} - \tilde{\beta}^{\mathsf{T}}\left(M^{-1}+(X^{\mathsf{T}}X)^{-1}\right)^{-1}\tilde{\beta}\\ &+ \hat{\beta}^{\mathsf{T}}(X^{\mathsf{T}}X)\hat{\beta} - \hat{\beta}^{\mathsf{T}}\left(M^{-1}+(X^{\mathsf{T}}X)^{-1}\right)^{-1}\hat{\beta}\\ &+ 2\hat{\beta}^{\mathsf{T}}(X^{\mathsf{T}}X)\left(M+X^{\mathsf{T}}X\right)^{-1}M\tilde{\beta}\\ &= \tilde{\beta}^{\mathsf{T}}M\tilde{\beta} + \hat{\beta}^{\mathsf{T}}(X^{\mathsf{T}}X)\hat{\beta}\\ &- (\tilde{\beta}-\hat{\beta})^{\mathsf{T}}\left(M^{-1}+(X^{\mathsf{T}}X)^{-1}\right)^{-1}(\tilde{\beta}-\hat{\beta})\end{aligned}$$

by virtue of the second identity. Therefore,

$$\begin{aligned}\pi(\sigma^2|\hat{\beta}, s^2, X) \propto \sigma^{-n-2a-2}\exp\frac{-1}{2\sigma^2}&\left\{(\tilde{\beta}-\hat{\beta})^{\mathsf{T}}\left(M^{-1}\right.\right.\\ &\left.\left.+(X^{\mathsf{T}}X)^{-1}\right)^{-1}(\tilde{\beta}-\hat{\beta}) + s^2 + 2b\right\}\end{aligned}$$

which is the distribution (3.4).

Exercise 3.5 Give a $(1-\alpha)$ HPD region on β based on (3.6).

As indicated just before this exercise,

$$\beta|\mathbf{y}, X \sim \mathscr{T}_{k+1}\left(n+2a, \hat{\mu}, \hat{\Sigma}\right).$$

This means that

$$\pi(\beta|\mathbf{y}, X) \propto \frac{1}{2}\left\{1 + \frac{(\beta-\hat{\mu})^{\mathsf{T}}\hat{\Sigma}^{-1}(\beta-\hat{\mu})}{n+2a}\right\}^{(n+2a+k+1)}$$

and therefore that an HPD region is of the form

$$\mathfrak{H}_\alpha = \left\{\beta;\, , (\beta-\hat{\mu})^{\mathsf{T}}\hat{\Sigma}^{-1}(\beta-\hat{\mu}) \le k_\alpha\right\},$$

where k_α is determined by the coverage probability α.

Now, $(\beta - \hat{\mu})^{\mathsf{T}}\hat{\Sigma}^{-1}(\beta - \hat{\mu})$ has the same distribution as $||z||^2$ when $z \sim \mathscr{T}_{k+1}(n + 2a, 0, I_{k+1})$. This distribution is Fisher's $\mathcal{F}(k + 1, n + 2a)$ distribution, which means that the bound k_α is determined by the quantiles of this distribution.

Exercise 3.6 The regression model can also be used in a predictive sense: for a given $(m, k + 1)$ explanatory matrix $\tilde{X}$, the corresponding outcome $\tilde{\mathbf{y}}$ can be inferred through the *predictive distribution* $\pi(\tilde{\mathbf{y}}|\sigma^2, \mathbf{y}, X, \tilde{X})$. Show that $\pi(\tilde{\mathbf{y}}|\sigma^2, \mathbf{y}, X, \tilde{X})$ is a Gaussian density with mean

$$\begin{aligned}\mathbb{E}[\tilde{\mathbf{y}}|\sigma^2, \mathbf{y}, X, \tilde{X}] &= \mathbb{E}[\mathbb{E}(\tilde{\mathbf{y}}|\beta, \sigma^2, \mathbf{y}, X, \tilde{X})|\sigma^2, \mathbf{y}, X, \tilde{X}] \\ &= \mathbb{E}[\tilde{X}\beta|\sigma^2, \mathbf{y}, X, \tilde{X}] \\ &= \tilde{X}(M + X^{\mathsf{T}}X)^{-1}(X^{\mathsf{T}}X\hat{\beta} + M\tilde{\beta})\end{aligned}$$

and covariance matrix

$$\begin{aligned}\mathbb{V}(\tilde{\mathbf{y}}|\sigma^2, \mathbf{y}, \tilde{X})x &= \mathbb{E}[\mathbb{V}(\tilde{\mathbf{y}}|\beta, \sigma^2, \mathbf{y}, X, \tilde{X})|\sigma^2, \mathbf{y}, X, \tilde{X}] \\ &\quad + \mathbb{V}(\mathbb{E}[\tilde{\mathbf{y}}|\beta, \sigma^2, \mathbf{y}, X, \tilde{X}]|\sigma^2, \mathbf{y}, X, \tilde{X}) \\ &= \mathbb{E}[\sigma^2 I_m|\sigma^2, \mathbf{y}, X, \tilde{X}] + \mathbb{V}(\tilde{X}\beta|\sigma^2, \mathbf{y}, X, \tilde{X}) \\ &= \sigma^2(I_m + \tilde{X}(M + X^{\mathsf{T}}X)^{-1}\tilde{X}^{\mathsf{T}})\,.\end{aligned}$$

Deduce that

$$\begin{aligned}\tilde{\mathbf{y}}|\mathbf{y}, X, \tilde{X} \sim \mathscr{T}_m \Big(&n + 2a, \tilde{X}(M + X^{\mathsf{T}}X)^{-1}(X^{\mathsf{T}}X\hat{\beta} + M\tilde{\beta}), \\ &\frac{2b + s^2 + (\tilde{\beta} - \hat{\beta})^{\mathsf{T}}\left(M^{-1} + (X^{\mathsf{T}}X)^{-1}\right)^{-1}(\tilde{\beta} - \hat{\beta})}{n + 2a} \\ &\times \left\{I_m + \tilde{X}(M + X^{\mathsf{T}}X)^{-1}\tilde{X}^{\mathsf{T}}\right\}\Big)\,.\end{aligned}$$

Since

$$\tilde{\mathbf{y}}|\tilde{X}, \beta, \sigma \sim \mathscr{N}\left(\tilde{X}\beta, \sigma^2 I_m\right)$$

and since the posterior distribution of β conditional on σ is given by (3.3), we have that

$$\tilde{\mathbf{y}}|\mathbf{y}, X, \tilde{X}, \sigma \sim \mathscr{N}\left(\tilde{X}\mathbb{E}[\beta|\sigma^2, \mathbf{y}, X], \sigma^2 I_m + \tilde{X}\mathrm{var}(\beta|\mathbf{y}, X, \sigma)\tilde{X}^{\mathsf{T}}\right)\,,$$

with mean $\tilde{X}(M + X^{\mathsf{T}}X)^{-1}(X^{\mathsf{T}}X\hat{\beta} + M\tilde{\beta})$, as shown by the derivation of Exercise 3.4. The variance is equal to $\sigma^2\left(I_m + \tilde{X}(M + X^{\mathsf{T}}X)^{-1}\tilde{X}^{\mathsf{T}}\right)$.

Integrating σ^2 against the posterior distribution (3.4) means that

$$\pi(\tilde{\mathbf{y}}|\mathbf{y},X,\tilde{X}) \propto \int_0^\infty \sigma^{-m-n-2a-1} \exp\frac{-1}{2\sigma^2}\left\{2b+s^2+\right.$$
$$(\tilde{\beta}-\hat{\beta})^\mathsf{T}\left(M^{-1}+(X^\mathsf{T}X)^{-1}\right)^{-1}(\tilde{\beta}-\hat{\beta})+(\tilde{\mathbf{y}}-\mathbb{E}[\tilde{\mathbf{y}}|\sigma^2,\mathbf{y},X,\tilde{X}])^\mathsf{T}(I_m$$
$$\left.+\tilde{X}(M+X^\mathsf{T}X)^{-1}\tilde{X}^\mathsf{T})^{-1}(\tilde{\beta}-\hat{\beta})(\tilde{\mathbf{y}}-\mathbb{E}[\tilde{\mathbf{y}}|\sigma^2,\mathbf{y},X,\tilde{X}])^\mathsf{T}\right\}\mathrm{d}\sigma^2$$
$$\propto \left\{2b+s^2+(\tilde{\beta}-\hat{\beta})^\mathsf{T}\left(M^{-1}+(X^\mathsf{T}X)^{-1}\right)^{-1}(\tilde{\beta}-\hat{\beta})+(\tilde{\mathbf{y}}\right.$$
$$-\mathbb{E}[\tilde{\mathbf{y}}|\sigma^2,\mathbf{y},X,\tilde{X}])^\mathsf{T}(I_m+\tilde{X}(M+X^\mathsf{T}X)^{-1}\tilde{X}^\mathsf{T})^{-1}(\tilde{\beta}-\hat{\beta})(\tilde{\mathbf{y}}$$
$$\left.-\mathbb{E}[\tilde{\mathbf{y}}|\sigma^2,\mathbf{y},X,\tilde{X}])^\mathsf{T}\right\}^{-(m+n+2a)/2}$$

which corresponds to a Student's $\mathcal{T}$ distribution with $(n+2a)$ degrees of freedom, a location parameter equal to $\mathbb{E}[\tilde{\mathbf{y}}|\sigma^2,\mathbf{y},X,\tilde{X}]$ [that does not depend on σ] and a scale parameter equal to

$$\left\{2b+s^2+(\tilde{\beta}-\hat{\beta})^\mathsf{T}\left(M^{-1}+(X^\mathsf{T}X)^{-1}\right)^{-1}(\tilde{\beta}-\hat{\beta})\right\}$$
$$\times\left[I_m+\tilde{X}(M+X^\mathsf{T}X)^{-1}\tilde{X}^\mathsf{T}\right]/(n+2a)\,.$$

Exercise 3.7 Show that the marginal distribution of $\mathbf{y}$ associated with (3.3) and (3.4) is given by

$$\mathbf{y}|X \sim \mathscr{T}_n\left(2a, X\tilde{\beta}, \frac{b}{a}(I_n+XM^{-1}X^\mathsf{T})\right).$$

This is a direct consequence of Exercise 3.6 when replacing $(\tilde{\mathbf{y}},\tilde{X})$ with $(\mathbf{y},X)$ and $(\mathbf{y},X)$ with the empty set. This is indeed equivalent to take $m=n$, $n=0$, $X=0$, $s^2=0$ and

$$(\tilde{\beta}-\hat{\beta})^\mathsf{T}\left(M^{-1}+(X^\mathsf{T}X)^{-1}\right)^{-1}(\tilde{\beta}-\hat{\beta})=0$$

in the previous exercice.

Exercise 3.8 Given the null hypothesis $H_0: R\beta=0$, where R is a (q,p) matrix of rank q, show that the restricted model on $\mathbf{y}$ given X can be represented as

$$\mathbf{y}|\beta_0,\sigma_0^2,X_0 \overset{H_0}{\sim} \mathscr{N}_n\left(X_0\beta_0,\sigma_0^2 I_n\right)$$

where X_0 is a $(n,k-q)$ matrix and β_0 is a $(k-q)$ dimensional vector. (*Hint:* Give the form of X_0 and β_0 in terms of X and β.) Under the hypothesis specific prior $\beta_0|H_0,\sigma_0^2 \sim \mathscr{N}_{k-q}\left(\tilde{\beta}_0,\sigma^2(M_0)^{-1}\right)$ and $\sigma_0^2|H_0 \sim \mathscr{IG}(a_0,b_0)$, construct the Bayes factor associated with the test of H_0.

When $R\beta = 0$, β satisfies q independent linear constraints, which means that q coordinates of β can be represented as linear combinations of the $(k-q)$ others, denoted by β_0, e.g.

$$\beta_{i_1} = \mathfrak{s}_{i_1}^\mathsf{T}\beta_0, \ldots, \beta_{i_q} = \mathfrak{s}_{i_q}^\mathsf{T}\beta_0\,,$$

where $i_1 < \cdots < i-q$ are the above coordinates. This implies that

$$X\beta = X\begin{pmatrix}\mathfrak{s}_1^\mathsf{T}\\ \cdots\\ \mathfrak{s}_q^\mathsf{T}\end{pmatrix}\beta_0 = X_0\beta_0\,,$$

where $\mathfrak{s}_i$ is either one of the above linear coefficients or contains 0 except for a single 1. Therefore, when H_0 holds, the expectation of $\mathbf{y}$ conditional on X can be written as $X_0\beta_0$ where X_0 is a $(n, k-q)$ matrix and β_0 is a $(k-q)$ dimensional vector. (Actually, there is an infinite number of ways to write $\mathbb{E}[\mathbf{y}|X]$ in this format.) The change from σ to σ_0 is purely notational to indicate that the variance σ_0^2 is associated with another model.

If we set a conjugate prior on (β_0, σ_0), the result of Exercise 3.7 also applies for this (sub-)model, in the sense that the marginal distribution of $\mathbf{y}$ for this model is

$$\mathbf{y}|X_0 \sim \mathscr{T}_n\left(2a_0, X_0\tilde{\beta}_0, \frac{b_0}{a_0}(I_n + X_0M_0^{-1}X_0^\mathsf{T})\right).$$

Therefore, the Bayes factor for testing H_0 can be written in closed form as

$$B_{01} = \frac{\Gamma((2a_0+n)/2)/\Gamma(2a_0/2)/\sigma\sqrt{2a_0}(b_0/a_0)^{n/2}|I_n + X_0M_0^{-1}X_0^\mathsf{T}|^{1/2}}{\Gamma((2a+n)/2)/\Gamma(2a/2)/\sigma\sqrt{2a}(b/a)^{n/2}|I_n + XM^{-1}X^\mathsf{T}|^{1/2}}$$
$$\times \frac{\left\{1 + (\mathbf{y} - X_0\tilde{\beta}_0)^\mathsf{T}(I_n + X_0M_0^{-1}X_0^\mathsf{T})^{-1}(\mathbf{y} - X_0\tilde{\beta}_0)/2b_0\right\}^{-(2a_0+n)/2}}{\left\{1 + (\mathbf{y} - X\tilde{\beta})^\mathsf{T}(I_n + XM^{-1}X^\mathsf{T})^{-1}(\mathbf{y} - X\tilde{\beta})/2b\right\}^{-(2a+n)/2}}$$

Note that, in this case, the normalising constants matter because they differ under H_0 and under the alternative.

Exercise 3.9 Show that

$$\beta|\mathbf{y}, X \sim \mathscr{T}_{k+1}\left(n, \frac{c}{c+1}\left(\frac{\tilde{\beta}}{c} + \hat{\beta}\right),\right.$$
$$\left.\frac{c(s^2 + (\tilde{\beta}-\hat{\beta})^\mathsf{T}X^\mathsf{T}X(\tilde{\beta}-\hat{\beta}))}{n(c+1)}(X^\mathsf{T}X)^{-1}\right).$$

Since

$$\beta|\sigma^2,\mathbf{y},X \sim \mathcal{N}_{k+1}\left(\frac{c}{c+1}(\tilde{\beta}/c+\hat{\beta}),\frac{\sigma^2 c}{c+1}(X^{\mathsf{T}}X)^{-1}\right),$$
$$\sigma^2|\mathbf{y},X \sim \mathcal{IG}\left(\frac{n}{2},\frac{s^2}{2}+\frac{1}{2(c+1)}(\tilde{\beta}-\hat{\beta})^{\mathsf{T}}X^{\mathsf{T}}X(\tilde{\beta}-\hat{\beta})\right),$$

we have that

$$\sqrt{\frac{c+1}{c}}\,\left[X^{\mathsf{T}}X\right]^{1/2}\left\{\beta-\frac{c}{c+1}(\tilde{\beta}/c+\hat{\beta})\right\} \sim \mathcal{N}_{k+1}\left(0,\sigma^2 I_n\right),$$

with

$$\left[s^2+\frac{1}{(c+1)}(\tilde{\beta}-\hat{\beta})^{\mathsf{T}}X^{\mathsf{T}}X(\tilde{\beta}-\hat{\beta})\right]/\sigma^2 \sim \chi^2_n,$$

which is the definition of the Student's

$$\mathcal{T}_{k+1}\left(n,\frac{c}{c+1}\left(\frac{\tilde{\beta}}{c}+\hat{\beta}\right),\frac{c(s^2+(\tilde{\beta}-\hat{\beta})^{\mathsf{T}}X^{\mathsf{T}}X(\tilde{\beta}-\hat{\beta})/(c+1))}{n(c+1)}(X^{\mathsf{T}}X)^{-1}\right)$$

distribution.

Exercise 3.10 Show that $\pi(\tilde{\mathbf{y}}|\sigma^2,\mathbf{y},X,\tilde{X})$ is a Gaussian density.

Conditional on σ^2, there is no difference with the setting of Exercise 3.6 since the only difference in using Zellner's G-prior compared with the conjugate priors is in the use of the noninformative prior $\pi(\sigma^2|X) \propto \sigma^{-2}$. Therefore, this is a consequence of Exercise 3.6.

Exercise 3.11 The posterior predictive distribution is obtained by integration over the marginal posterior distribution of σ^2. Derive $\pi(\tilde{\mathbf{y}}|\mathbf{y},X,\tilde{X})$.

Once more, integrating the normal distribution over the inverse gamma random variable σ^2 produces a Student's $\mathcal{T}$ distribution. Since

$$\sigma^2|\mathbf{y},X \sim \mathcal{IG}\left(\frac{n}{2},\frac{s^2}{2}+\frac{1}{2(c+1)}(\tilde{\beta}-\hat{\beta})^{\mathsf{T}}X^{\mathsf{T}}X(\tilde{\beta}-\hat{\beta})\right)$$

under Zellner's G-prior, the predictive distribution is a

$$\tilde{\mathbf{y}}|\mathbf{y},X,\tilde{X} \sim \mathcal{T}_{k+1}\left(n,\tilde{X}\frac{\tilde{\beta}+c\hat{\beta}}{c+1},\frac{c(s^2+(\tilde{\beta}-\hat{\beta})^{\mathsf{T}}X^{\mathsf{T}}X(\tilde{\beta}-\hat{\beta})/(c+1))}{n(c+1)}\right.$$
$$\left.\times\left\{I_m+\frac{c}{c+1}\tilde{X}(X^{\mathsf{T}}X)^{-1}\tilde{X}^{\mathsf{T}}\right\}\right)$$

distribution.

Exercise 3.12 Give a joint $(1-\alpha)$ HPD region on β.

Since we have (Exercise 3.9)

$$\beta|\mathbf{y}, X \sim \mathscr{T}_{k+1}\left(n, \frac{c}{c+1}\left(\frac{\tilde{\beta}}{c}+\hat{\beta}\right), \frac{c(s^2+(\tilde{\beta}-\hat{\beta})^\mathsf{T}X^\mathsf{T}X(\tilde{\beta}-\hat{\beta}))}{n(c+1)}(X^\mathsf{T}X)^{-1}\right),$$

with

$$\hat{\Sigma} = \frac{c(s^2+(\tilde{\beta}-\hat{\beta})^\mathsf{T}X^\mathsf{T}X(\tilde{\beta}-\hat{\beta}))}{n(c+1)}(X^\mathsf{T}X)^{-1},$$

an HPD region is of the form

$$\mathfrak{H}_\alpha = \left\{\beta;\, , (\beta-\hat{\mu})^\mathsf{T}\hat{\Sigma}^{-1}(\beta-\hat{\mu}) \leq k_\alpha\right\},$$

where k_α is determined by the coverage probability α and

$$\hat{\mu} = \frac{c}{c+1}\left(\frac{\tilde{\beta}}{c}+\hat{\beta}\right), \quad \hat{\Sigma} = \frac{c(s^2+(\tilde{\beta}-\hat{\beta})^\mathsf{T}X^\mathsf{T}X(\tilde{\beta}-\hat{\beta}))}{n(c+1)}(X^\mathsf{T}X)^{-1}.$$

As in Exercise 3.5, the distribution of $(\beta-\hat{\mu})^\mathsf{T}\hat{\Sigma}^{-1}(\beta-\hat{\mu})$ is a Fisher's $\mathcal{F}(k+1,n)$ distribution.

Exercise 3.13 Show that the matrix $(I_n + cX(X^\mathsf{T}X)^{-1}X^\mathsf{T})$ has 1 and $c+1$ as eigenvalues. (*Hint:* Show that the eigenvectors associated with $c+1$ are of the form $X\beta$ and that the eigenvectors associated with 1 are those orthogonal to X, i.e. z's such that $X^\mathsf{T}z = 0$.) Deduce that the determinant of the matrix $(I_n + cX(X^\mathsf{T}X)^{-1}X^\mathsf{T})$ is indeed $(c+1)^{(k+1)/2}$.

Given the hint, this is somehow obvious:

$$(I_n + cX(X^\mathsf{T}X)^{-1}X^\mathsf{T})X\beta = X\beta + cX(X^\mathsf{T}X)^{-1}X^\mathsf{T}X\beta = (c+1)X\beta$$
$$(I_n + cX(X^\mathsf{T}X)^{-1}X^\mathsf{T})z = z + cX(X^\mathsf{T}X)^{-1}X^\mathsf{T}z = z$$

for all β's in $\mathbb{R}^{k+1}$ and all z's orthogonal to X. Since the addition of those two subspaces generates a vector space of dimension n, this defines the whole

set of eigenvectors for both eigenvalues. And since the vector subspace generated by X is of dimension $(k+1)$, this means that the determinant of $(I_n + cX(X^\mathsf{T}X)^{-1}X^\mathsf{T})$ is $(c+1)^{k+1} \times 1^{n-k-1}$.

Exercise 3.14 Derive the marginal posterior distribution of β for this model.

The joint posterior is given by

$$\beta|\sigma^2, \mathbf{y}, X \sim \mathcal{N}_{k+1}\left(\hat{\beta}, \sigma^2(X^\mathsf{T}X)^{-1}\right),$$
$$\sigma^2|\mathbf{y}, X \sim \mathcal{IG}((n-k-1)/2, s^2/2).$$

Therefore,

$$\beta|\mathbf{y}, X \sim \mathcal{T}_{k+1}\left(n-k-1, \hat{\beta}, \frac{s^2}{n-k-1}(X^\mathsf{T}X)^{-1}\right)$$

by the same argument as in the previous exercises.

Exercise 3.15 Show that the marginal posterior distribution of β_i $(1 \leq i \leq k)$ is a $\mathcal{T}_1(n-k-1, \hat{\beta}_i, \omega_{(i,i)}s^2/(n-k-1))$ distribution. (*Hint:* Recall that $\omega_{(i,i)} = (X^\mathsf{T}X)^{-1}_{(i,i)}$.)

The argument is straightforward: since $\beta|\sigma^2, \mathbf{y}, X \sim \mathcal{N}_{k+1}\left(\hat{\beta}, \sigma^2(X^\mathsf{T}X)^{-1}\right)$, $\beta_i|\sigma^2, \mathbf{y}, X \sim \mathcal{N}\left(\hat{\beta}_i, \sigma^2\omega_{(i,i)}\right)$. Integrating out σ^2 as in the previous exercise leads to

$$\beta_i|\sigma^2, \mathbf{y}, X \sim \mathcal{T}_1(n-k-1, \hat{\beta}_i, \omega_{(i,i)}s^2/(n-k-1)).$$

Exercise 3.16 Give the predictive distribution of $\tilde{\mathbf{y}}$, the m dimensional vector corresponding to the (m,k) matrix of explanatory variables $\tilde{X}$.

This predictive can be derived from Exercise 3.6. Indeed, Jeffreys' prior is nothing but a special case of conjugate prior with $a = b = 0$. Therefore, Exercise 3.6 implies that, in this limiting case,

$$\tilde{\mathbf{y}}|\mathbf{y}, X, \tilde{X} \sim \mathcal{T}_m\left(n, \tilde{X}(M + X^\mathsf{T}X)^{-1}(X^\mathsf{T}X\hat{\beta} + M\tilde{\beta}),\right.$$
$$\frac{s^2 + (\tilde{\beta}-\hat{\beta})^\mathsf{T}\left(M^{-1} + (X^\mathsf{T}X)^{-1}\right)^{-1}(\tilde{\beta}-\hat{\beta})}{n}$$
$$\left.\times\left\{I_m + \tilde{X}(M + X^\mathsf{T}X)^{-1}\tilde{X}^\mathsf{T}\right\}\right).$$

Exercise 3.17 When using the prior distribution $\pi(c) = 1/c^2$, compare the results with Table 3.6.

In the file #3.txt provided on the Webpage, it suffices to replace `cc^(-1)` with `cc^(-2)` : for instance, the point estimate of β is now

```
> facto=sum(cc/(cc+1)*cc^(-2)*(cc+1)^(-11/2)*
+ (t(y)%*%y-cc/(cc+1)*t(y)%*%P%*%y)^(-33/2))/
+ sum(cc^(-2)*(cc+1)^(-11/2)*(t(y)%*%y-cc/
+ (cc+1)*t(y)%*%P%*%y)^(-33/2))
> facto*betahat
               [,1]
 [1,]  8.506662193
 [2,] -0.003426982
 [3,] -0.041634562
 [4,]  0.052547326
 [5,] -1.000556061
 [6,]  0.179158187
 [7,] -0.275964816
 [8,] -0.183669178
 [9,]  0.140040003
[10,] -0.994120776
[11,] -0.334983109
```

Exercise 3.18 Show that both series (3.10) and (3.11) converge.

Given that

$$f(\mathbf{y}|X,c) \propto (c+1)^{-(k+1)/2} \left[\mathbf{y}^\mathsf{T}\mathbf{y} - \frac{c}{c+1}\mathbf{y}^\mathsf{T}X(X^\mathsf{T}X)^{-1}X^\mathsf{T}\mathbf{y}\right]^{-n/2}$$
$$\approx c^{-(k+1)/2} \left[\mathbf{y}^\mathsf{T}\mathbf{y} - \mathbf{y}^\mathsf{T}X(X^\mathsf{T}X)^{-1}X^\mathsf{T}\mathbf{y}\right]^{-n/2}$$

when c goes to ∞, the main term in the series goes to 0 as a $o(c^{-(k+3)/2})$ and the series converges.

Obviously, if the first series converges, then so does the second series.

Exercise 3.19 Give the predictive distribution of $\tilde{\mathbf{y}}$, the m-dimensional vector corresponding to the (m,k) matrix of explanatory variables $\tilde{X}$.

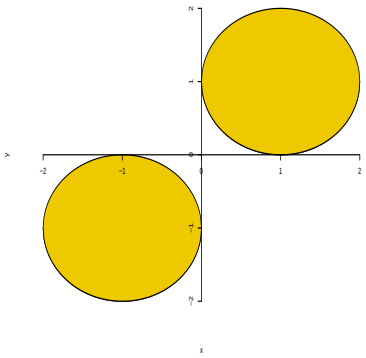

Fig. 3.1. Support of the uniform distribution.

The predictive of $\tilde{\mathbf{y}}$ given $\mathbf{y}, X, \tilde{X}$ is then the weighted average of the predictives given $\mathbf{y}, X, \tilde{X}$ and c:

$$\pi(\tilde{\mathbf{y}}|\mathbf{y}, X, \tilde{X}) \propto \sum_{c=1}^{\infty} \pi(\tilde{\mathbf{y}}|\mathbf{y}, X, \tilde{X}, c) f(\mathbf{y}|X, c)\, c^{-1}$$

where $\pi(\tilde{\mathbf{y}}|\mathbf{y}, X, \tilde{X}, c)$ is the Student's $\mathscr{T}$ distribution obtained in Exercise 3.11.

Exercise 3.20 If (x_1, x_2) is distributed from the uniform distribution on

$$\big\{(x_1, x_2);\ (x_1-1)^2 + (x_2-1)^2 \leq 1\big\} \cup \big\{(x_1, x_2);\ (x_1+1)^2 + (x_2+1)^2 \leq 1\big\}\ ,$$

show that the Gibbs sampler does not produce an irreducible chain. For this distribution, find an alternative Gibbs sampler that works. (*Hint:* Consider a rotation of the coordinate axes.)

The support of this uniform distribution is made of two disks with respective centers $(-1, -1)$ and $(1, 1)$, and with radius 1. This support is not connected (see Figure 3.1) and conditioning on $x_1 < 0$ means that the conditional distribution of x_2 is $\mathscr{U}(-1 - \sqrt{1 - x_1^2}, -1 + \sqrt{1 - x_1^2}$, thus cannot produce a value in $[0, 1]$. Similarly, when simulating the next value of x_1, it necessarily remains negative. The Gibbs sampler thus produces two types of chains, depending on whether or not it is started from the negative disk. If we now consider the Gibbs sampler for the new parameterisation

$$y_1 = x_1 + x_2, \quad y_2 = x_2 - x_1\,,$$

conditioning on y_1 produces a uniform distribution on the union of a negative and of a positive interval. Therefore, one iteration of the Gibbs sampler is sufficient to jump [with positive probability] from one disk to the other one.

Exercise 3.21 If a joint density $g(y_1, y_2)$ corresponds to the conditional distributions $g_1(y_1|y_2)$ and $g_2(y_2|y_1)$, show that it is given by

$$g(y_1, y_2) = \frac{g_2(y_2|y_1)}{\int g_2(v|y_1)/g_1(y_1|v)\,\mathrm{d}v}.$$

If the joint density $g(y_1, y_2)$ exists, then

$$\begin{aligned} g(y_1, y_2) &= g^1(y_1) g_2(y_2|y_1) \\ &= g^2(y_2) g_1(y_1|y_2) \end{aligned}$$

where g^1 and g^2 denote the densities of the marginal distributions of y_1 and y_2, respectively. Thus,

$$\begin{aligned} g^1(y_1) &= \frac{g_1(y_1|y_2)}{g_2(y_2|y_1)} g^2(y_2) \\ &\propto \frac{g_1(y_1|y_2)}{g_2(y_2|y_1)}\,, \end{aligned}$$

as a function of y_1 [$g^2(y_2)$ is irrelevant]. Since g^1 is a density,

$$g^1(y_1) = \frac{g_1(y_1|y_2)}{g_2(y_2|y_1)} \Big/ \int \frac{g_1(u|y_2)}{g_2(y_2|u)} \mathrm{d}u$$

and

$$g(y_1, y_2) = g_1(y_1|y_2) \Big/ \int \frac{g_1(u|y_2)}{g_2(y_2|u)} \mathrm{d}u\,.$$

Since y_1 and y_2 play symmetric roles in this derivation, the symmetric version also holds.

Exercise 3.22 Check that the starting value of μ in the setting of Example 3.2 has no influence on the output of the above Gibbs sampler after $N = 1000$ iterations.

The core of the Gibbs program is

```
> mu = rnorm(1,sum(x*omega)/sum(omega+.05),
+ sqrt(1/(.05+2*sum(omega)))
> omega = rexp(2,1+(x-mu)^2)
```

which needs to be iterated $N = 1000$ times to produce a Gibbs N-sample from $\pi(\mu|\mathcal{D})$. A R program evaluating the [lack of] influence of the starting value $\mu^{(0)}$ will thus need to compare histograms of $\mu^{(1000)}$'s for different starting values. It therefore requires three loops:

```
x=c(3.2,-1.5) # observations
mu0=seq(-10,10,length=20) # starting values
muk=rep(0,250)

par(mfrow=c(5,4),mar=c(4,2,4,1)) # multiple histograms

for (i in 1:20){

  for (t in 1:250){

    mu=mu0[i]
    for (iter in 1:1000){

      omega = rexp(2,1+(x-mu)^2)
      mu = rnorm(1,sum(x*omega)/sum(omega+.05),
           sqrt(1/(.05+2*sum(omega))))
    }
    muk[t]=mu
  }
  hist(muk,proba=T,col="wheat",main=paste(mu0[i]))

}
```

Be warned that the use of this triple loop induces a long wait on most machines!

Exercise 3.23 In the setup of Section 3.5.3, show that

$$\pi(\gamma|\mathbf{y},X) \propto \sum_{c=1}^{\infty} c^{-1}(c+1)^{-(q_\gamma+1)/2} \left[\mathbf{y}^\mathsf{T}\mathbf{y} - \frac{c}{c+1}\mathbf{y}^\mathsf{T}X_\gamma\left(X_\gamma^\mathsf{T}X_\gamma\right)^{-1}X_\gamma^\mathsf{T}\mathbf{y}\right]^{-n/2}$$

and that the series converges. If $\pi(c) \propto c^{-\alpha}$, find which values of α lead to a proper posterior.

We can take advantage of Section 3.5.2: when c is fixed in Zellner's informative G-prior and $\tilde{\beta}_\gamma = \mathbf{0}_{q_\gamma+1}$ for all γ's,

$$\pi(\gamma|\mathbf{y}, X, c) \propto (c+1)^{-(q_\gamma+1)/2} \left[\mathbf{y}^\mathsf{T}\mathbf{y} - \frac{c}{c+1}\mathbf{y}^\mathsf{T}X_\gamma \left(X_\gamma^\mathsf{T}X_\gamma\right)^{-1} X_\gamma^\mathsf{T}\mathbf{y}\right]^{-n/2},$$

thus

$$\pi(\gamma, c|\mathbf{y}, X, c) \propto c^{-1}(c+1)^{-(q_\gamma+1)/2} \left[\mathbf{y}^\mathsf{T}\mathbf{y} - \frac{c}{c+1}\mathbf{y}^\mathsf{T}X_\gamma \left(X_\gamma^\mathsf{T}X_\gamma\right)^{-1} X_\gamma^\mathsf{T}\mathbf{y}\right]^{-n/2}.$$

and

$$\begin{aligned}\pi(\gamma|\mathbf{y}, X) &= \sum_{c=1}^{\infty} \pi(\gamma, c|\mathbf{y}, X, c)\\ &\propto \sum_{c=1}^{\infty} c^{-1}(c+1)^{-(q_\gamma+1)/2} \left[\mathbf{y}^\mathsf{T}\mathbf{y} - \frac{c}{c+1}\mathbf{y}^\mathsf{T}X_\gamma \left(X_\gamma^\mathsf{T}X_\gamma\right)^{-1} X_\gamma^\mathsf{T}\mathbf{y}\right]^{-n/2}.\end{aligned}$$

For $\pi(c) \propto c^{-\alpha}$, the series

$$\sum_{c=1}^{\infty} c^{-\alpha}(c+1)^{-(q_\gamma+1)/2} \left[\mathbf{y}^\mathsf{T}\mathbf{y} - \frac{c}{c+1}\mathbf{y}^\mathsf{T}X_\gamma \left(X_\gamma^\mathsf{T}X_\gamma\right)^{-1} X_\gamma^\mathsf{T}\mathbf{y}\right]^{-n/2}$$

converges if and only if, for all γ's,

$$\alpha + \frac{q_\gamma + 1}{2} > 1\,,$$

which is equivalent to $2\alpha + q_\gamma > 1$. Since $\min(q_\gamma) = 0$, the constraint for propriety of the posterior is

$$\alpha > 1/2\,.$$

4

Generalized Linear Models

Exercise 4.1 For bank, derive the maximum likelihood estimates of β_0 and β_1 found in the previous analysis. Using Jeffreys prior on the parameters $(\beta_0, \beta_1, \sigma^2)$ of the linear regression model, compute the corresponding posterior expectation of (β_0, β_1).

The code is provided in the file `#4.txt` on the Webpage. If the `bank` dataset is not available, it can be downloaded from the Webpage and the following code can be used:

```
bank=matrix(scan("bank"),byrow=T,ncol=5)
y=as.vector(bank[,5])
X=cbind(rep(1,200),as.vector(bank[,1]),as.vector(bank[,2]),
     as.vector(bank[,3]),as.vector(bank[,4]))
summary(lm(y~X[,5]))
```

which produces the output [leading to eqn. (4.1) in the book]:

```
> summary(lm(y~X[,5]))

Call:
lm(formula = y ~ X[, 5])

Residuals:
     Min       1Q   Median       3Q      Max
-0.76320 -0.21860 -0.06228  0.18322  1.04046

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.02282    0.14932  -13.55   <2e-16 ***
```

```
X[, 5]       0.26789    0.01567   17.09   <2e-16 ***
---
Sig. codes: 0 '***' .001 '**' .01 '*' .05 '.' .1 ' ' 1

Residual standard error: 0.3194 on 198 degrees of freedom
Multiple R-Squared: 0.596,      Adjusted R-squared: 0.594
F-statistic: 292.2 on 1 and 198 DF,  p-value: < 2.2e-16
```

As shown in Exercise 3.14, the [marginal] posterior in β associated with the Jeffreys prior is

$$\beta|\mathbf{y}, X \sim \mathscr{T}_{k+1}\left(n-k-1, \hat{\beta}, \frac{s^2}{n-k-1}(X^{\mathsf{T}}X)^{-1}\right)$$

so the posterior expectation of (β_0, β_1) is again $\hat{\beta}$.

Exercise 4.2 Show that, in the setting of Example 4.1, the statistic $\sum_{i=1}^n y_i\,\mathbf{x}^i$ is sufficient when conditioning on the $\mathbf{x}^i$'s $(1 \le i \le n)$ and give the corresponding family of conjugate priors.

Since the likelihood is

$$\begin{aligned}&\exp\left\{\sum_{i=1}^n y_i\,\mathbf{x}^{i\mathsf{T}}\beta\right\}\Big/\prod_{i=1}^n\left[1+\exp(\mathbf{x}^{i\mathsf{T}}\beta)\right]\\ &\quad=\exp\left\{\sum_{i=1}^n\left[y_i\,\mathbf{x}^i\right]^{\mathsf{T}}\beta\right\}\Big/\prod_{i=1}^n\left[1+\exp(\mathbf{x}^{i\mathsf{T}}\beta)\right]\,,\end{aligned}$$

it depends on the observations $(y_1, \ldots, y_n)$ only through the sum $\sum_{i=1}^n y_i\,\mathbf{x}^i$ which is thus a sufficient statistic in this conditional sense.

The family of priors $(\xi \in \mathbb{R}^k\,, \lambda > 0)$

$$\pi(\beta|\xi,\lambda) \propto \exp\left\{\xi^{\mathsf{T}}\beta\right\}\Big/\prod_{i=1}^n\left[1+\exp(\mathbf{x}^{i\mathsf{T}}\beta)\right]^{\lambda}\,,$$

is obviously conjugate. The corresponding posterior is

$$\pi\left(\beta\,\middle|\,\xi+\sum_{i=1}^n y_i\,\mathbf{x}^i, \lambda+1\right)\,,$$

whose drawback is to have λ updated in $\lambda+1$ rather than $\lambda+n$ as in other conjugate settings. This is due to the fact that the prior is itself conditional on X and therefore on n.

Exercise 4.3 Show that the logarithmic link is the canonical link function in the case of the Poisson regression model.

The likelihood of the Poisson regression model is

$$\begin{aligned}\ell(\beta|\mathbf{y}, X) &= \prod_{i=1}^{n} \left(\frac{1}{y_i!}\right) \exp\left\{y_i\, \mathbf{x}^{i\mathsf{T}}\beta - \exp(\mathbf{x}^{i\mathsf{T}}\beta)\right\} \\ &= \prod_{i=1}^{n} \frac{1}{y_i!} \exp\left\{y_i\, \log(\mu_i) - \mu_i\right\}\,,\end{aligned}$$

so $\log(\mu_i) = \mathbf{x}^{i\mathsf{T}}\beta$ and the logarithmic link is indeed the canonical link function.

Exercise 4.4 Suppose $y_1, \ldots, y_k$ are independent Poisson $\mathcal{P}(\mu_i)$ random variables. Show that, conditional on $n = \sum_{i=1}^{k} y_i$,

$$\mathbf{y} = (y_1, \ldots, y_k) \sim \mathcal{M}_k(n; \alpha_1, \ldots, \alpha_k)$$

and determine the α_i's.

The joint distribution of $\mathbf{y}$ is

$$f(\mathbf{y}|\mu_1, \ldots, \mu_k) = \prod_{i=1}^{k} \left(\frac{\mu_i^{y_i}}{y_i!}\right) \exp\left\{-\sum_{i=1}^{k} \mu_i\right\}\,,$$

while $n = \sum_{i=1}^{k} y_i \sim \mathcal{P}(\sum_{i=1}^{k} \mu_i)$ [which can be established using the moment generating function of the $\mathcal{P}(\mu)$ distribution]. Therefore, the conditional distribution of $\mathbf{y}$ given n is

$$\begin{aligned}f(\mathbf{y}|\mu_1, \ldots, \mu_k, n) &= \frac{\prod_{i=1}^{k} \left(\frac{\mu_i^{y_i}}{y_i!}\right) \exp\left\{-\sum_{i=1}^{k} \mu_i\right\}}{\frac{[\sum_{i=1}^{k} \mu_i]^n}{n!} \exp\left\{-\sum_{i=1}^{k} \mu_i\right\}} \mathbb{I}_n\left(\sum_{i=1}^{k} y_i\right) \\ &= \frac{n!}{\prod_{i=1}^{k} y_i!} \prod_{i=1}^{k} \left(\frac{\mu_i}{\sum_{i=1}^{k} \mu_i}\right)^{y_i} \mathbb{I}_n\left(\sum_{i=1}^{k} y_i\right)\,,\end{aligned}$$

which is the pdf of the $\mathcal{M}_k(n; \alpha_1, \ldots, \alpha_k)$ distribution, with

$$\alpha_i = \frac{\mu_i}{\sum_{j=1}^{k} \mu_j}\,, \qquad i = 1, \ldots, k\,.$$

This conditional representation is a standard property used in the statistical analysis of contingency tables (Section 4.5): when the margins are random, the cells are Poisson while, when the margins are fixed, the cells are multinomial.

Exercise 4.5 Show that the detailed balance equation also holds for the *Boltzmann* acceptance probability

$$\rho(x,y) = \frac{\pi(y)q(y,x)}{\pi(y)q(y,x)+\pi(x)q(x,y)}\,.$$

The detailed balance equation is

$$\pi(x)q(x,y)\rho(x,y) = \pi(y)q(y,x)\rho(y,x)\,.$$

Therefore, in the case of the Boltzmann acceptance probability

$$\begin{aligned}\pi(x)q(x,y)\rho(x,y) &= \pi(x)q(x,y)\frac{\pi(y)q(y,x)}{\pi(y)q(y,x)+\pi(x)q(x,y)}\\ &= \frac{\pi(x)q(x,y)\pi(y)q(y,x)}{\pi(y)q(y,x)+\pi(x)q(x,y)}\\ &= \pi(y)q(y,x)\frac{\pi(x)q(x,y)}{\pi(y)q(y,x)+\pi(x)q(x,y)}\\ &= \pi(y)q(y,x)\rho(y,x)\,.\end{aligned}$$

Note that this property also holds for the generalized Boltzmann acceptance probability

$$\rho(x,y) = \frac{\pi(y)q(y,x)\alpha(x,y)}{\pi(y)q(y,x)\alpha(x,y)+\pi(x)q(x,y)\alpha(y,x)}\,,$$

where $\alpha(x,y)$ is an arbitrary positive function.

Exercise 4.6 For π the density of an inverse normal distribution with parameters $\theta_1 = 3/2$ and $\theta_2 = 2$,

$$\pi(x) \propto x^{-3/2}\exp(-3/2x - 2/x)\mathbb{I}_{x>0},$$

write down and implement an independence MH sampler with a Gamma proposal with parameters $(\alpha,\beta) = (4/3,1)$ and $(\alpha,\beta) = (0.5\sqrt{4/3},0.5)$.

A R possible code for running an independence Metropolis–Hastings sampler in this setting is as follows:

```
# target density
target=function(x,the1=1.5,the2=2){
  x^(-the1)*exp(-the1*x-the2/x)
  }

al=4/3
bet=1

# initial value
mcmc=rep(1,1000)

for (t in 2:1000){

  y = rgamma(1,shape=al,rate=bet)
  if (runif(1)<target(y)*dgamma(mcmc[t-1],shape=al,rate=bet)/
(target(mcmc[t-1])*dgamma(y,shape=al,rate=bet)))
    mcmc[t]=y
    else
      mcmc[t]=mcmc[t-1]
  }

# plots
par(mfrow=c(2,1),mar=c(4,2,2,1))
res=hist(mcmc,freq=F,nclass=55,prob=T,col="grey56",
  ylab="",main="")
lines(seq(0.01,4,length=500),valpi*max(res$int)/max(valpi),
  lwd=2,col="sienna2")
plot(mcmc,type="l",col="steelblue2",lwd=2)
```

The output of this code is illustrated on Figure 4.1 and shows a reasonable fit of the target by the histogram and a proper mixing behaviour. Out of the 1000 iterations in this example, 600 corresponded to an acceptance of the Gamma random variable. (Note that to plot the density on the same scale as the histogram, we resorted to a trick on the maxima of the histogram and of the density.)

Exercise 4.7 Estimate the mean of a $\mathcal{G}a(4.3, 6.2)$ random variable using

1. direct sampling from the distribution via R command
   `> x=rgamma(n,4.3,rate=6.2)`
2. Metropolis–Hastings with a $\mathcal{G}a(4, 7)$ proposal distribution;
3. Metropolis–Hastings with a $\mathcal{G}a(5, 6)$ proposal distribution.

In each case, monitor the convergence of the cumulated average.

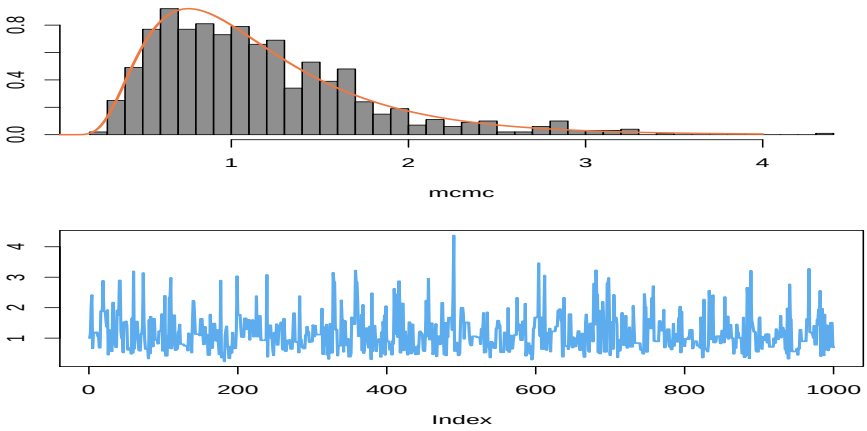

Fig. 4.1. Output of an MCMC simulation of the inverse normal distribution.

Both independence Metropolis–Hastings samplers can be implemented via an R code like

```
al=4.3
bet=6.2

mcmc=rep(1,1000)
for (t in 2:1000){

  mcmc[,t]=mcmc[,t-1]
  y = rgamma(500,4,rate=7)
  if (runif(1)< dgamma(y,al,rate=bet)*dgamma(mcmc[t-1],4,rate=7)/
(dgamma(mcmc[t-1],al,rate=bet)*dgamma(y,4,rate=7))){
    mcmc[t]=y
    }
}
aver=cumsum(mcmc)/1:1000
```

When comparing those samplers, their variability can only be evaluated through repeated calls to the above code, in order to produce a range of outputs for the three methods. For instance, one can define a matrix of cumulated averages `aver=matrix(0,250,1000)` and take the range of the cumulated averages over the 250 repetitions as in `ranj=apply(aver,1,range)`, leading to something similar to Figure 4.2. The complete code for one of the ranges is

```
al=4.3
bet=6.2
```

```
mcmc=matrix(1,ncol=1000,nrow=500)
for (t in 2:1000){
  mcmc[,t]=mcmc[,t-1]
  y = rgamma(500,4,rate=7)
  valid=(runif(500)<dgamma(y,al,rate=bet)*
    dgamma(mcmc[i,t-1],4,rate=7)/(dgamma(mcmc[,t-1],al,rate=bet)*
    dgamma(y,4,rate=7)))
  mcmc[valid,t]=y[valid]
  }
aver2=apply(mcmc,1,cumsum)
aver2=t(aver2/(1:1000))
ranj2=apply(aver2,2,range)
plot(ranj2[1,],type="l",ylim=range(ranj2),ylab="")
polygon(c(1:1000,1000:1),c(ranj2[2,],rev(ranj2[1,])))
```

which removes the Monte Carlo loop over the 500 replications by running the simulations in parallel. We can notice on Figure 4.2 that, while the output from the third sampler is quite similar with the output from the iid sampler [since we use the same scale on the y axis], the Metropolis–Hastings algorithm based on the $\mathcal{G}a(4, 7)$ proposal is rather biased, which may indicate a difficulty in converging to the stationary distribution. This is somehow an expected problem, in the sense that the ratio target-over-proposal is proportional to $x^{0.3} \exp(0.8x)$, which is explosive at both $x = 0$ and $x = \infty$.

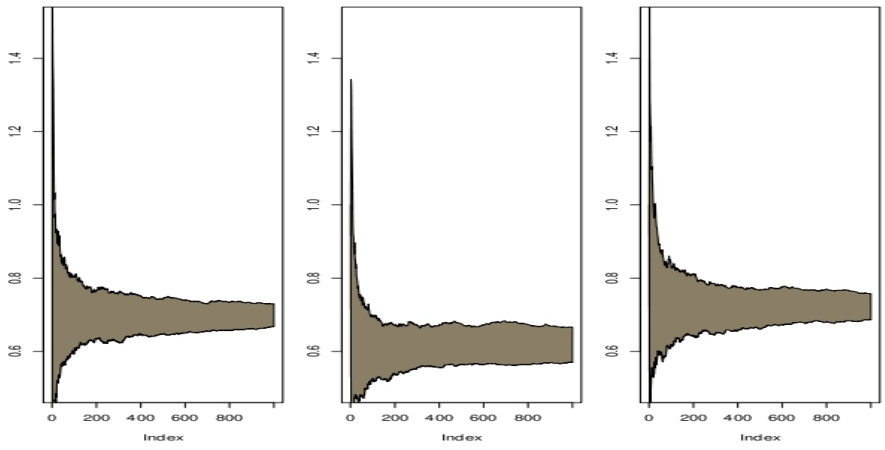

Fig. 4.2. Range of three samplers for the approximation of the $\mathcal{G}a(4.3, 6.2)$ mean: *(left)* iid; *(center)* $\mathcal{G}a(4, 7)$ proposal; *(right)* $\mathcal{G}a(5, 6)$ proposal.

Exercise 4.8 Consider x_1, x_2 and x_3 iid $\mathscr{C}(\theta, 1)$, and $\pi(\theta) \propto \exp(-\theta^2/100)$. Show that the posterior distribution of θ, $\pi(\theta|x_1, x_2, x_3)$, is proportional to

$$\exp(-\theta^2/100)[(1+(\theta-x_1)^2)(1+(\theta-x_2)^2)(1+(\theta-x_3)^2)]^{-1} \tag{4.1}$$

and that it is trimodal when $x_1 = 0$, $x_2 = 5$ and $x_3 = 9$. Using a random walk based on the Cauchy distribution $\mathscr{C}(0, \sigma^2)$, estimate the posterior mean of θ using different values of σ^2. In each case, monitor the convergence.

The function (4.1) appears as the product of the prior by the three densities $f(x_i|\theta)$. The trimodality of the posterior can be checked on a graph when plotting the function (4.1).

A random walk Metropolis–Hastings algorithm can be coded as follows

```
x=c(0,5,9)
# target
targ=function(y){
  dnorm(y,sd=sqrt(50))*dt(y-x[1],df=1)*
  dt(y-x[2],df=1)*dt(y-x[3],df=1)
}

# Checking trimodality
plot(seq(-2,15,length=250),
  targ(seq(-2,15,length=250)),type="l")

sigma=c(.001,.05,1)*9 # different scales
N=100000 # number of mcmc iterations

mcmc=matrix(mean(x),ncol=3,nrow=N)
for (t in 2:N){

   mcmc[t,]=mcmc[t-1,]
   y=mcmc[t,]+sigma*rt(3,1) # rnorm(3)
   valid=(runif(3)<targ(y)/targ(mcmc[t-1,]))
   mcmc[t,valid]=y[valid]
   }
```

The comparison of the three cumulated averages is given in Figure 4.3 and shows that, for the Cauchy noise, both large scales are acceptable while the smallest scale slows down the convergence properties of the chain. For the normal noise, these features are exacerbated in the sense that the smallest scale does not produce convergence for the number of iterations under study [the blue curve leaves the window of observation], the medium scale induces

some variability and it is only the largest scale that gives an acceptable approximation to the mean of the distribution (4.1).

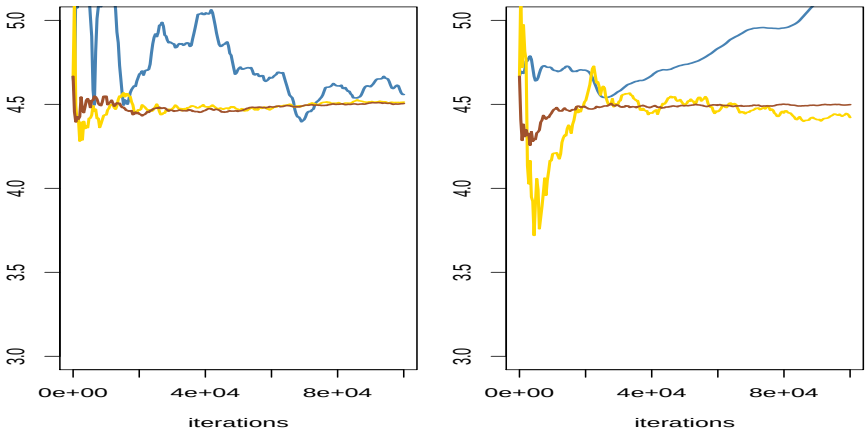

Fig. 4.3. Comparison of the three scale factors $\sigma = .009$ (blue), $\sigma = .45$ (gold) and $\sigma = 9$ (brown), when using a Cauchy noise *(left)* and a normal noise *(right)*.

Exercise 4.9 Rerun the experiment of Example 4.4 using instead a mixture of five random walks with variances $\sigma = 0.01, 0.1, 1, 10, 100$, and equal weights, and compare its output with the output of Figure 4.3.

The original code is provided in the files `#4.R` and `#4.txt` on the webpage. The modification of the `hm1` function is as follows:

```
hmi=function(n,x0,sigma2)
{
x=rep(x0,n)

for (i in 2:n){

  x[i]=x[i-1]
  y=rnorm(1,x[i-1],sqrt(sample(sigma2,1)))
  if (runif(1)<dnorm(y)/dnorm(x[i-1]))
    x[i]=y
  }
x
}
```

Note that picking the variance at random does not modify the random walk structure of the proposal, which is then a mixture of normal distributions all centered in $x^{(t-1)}$. The output compares with Figure 4.3 [from the book] but the histogram is not as smooth and the autocorrelations are higher, which can be easily explained by the fact that using a whole range of scales induces inefficiencies in that the poor scales are chosen for "nothing" from time to time.

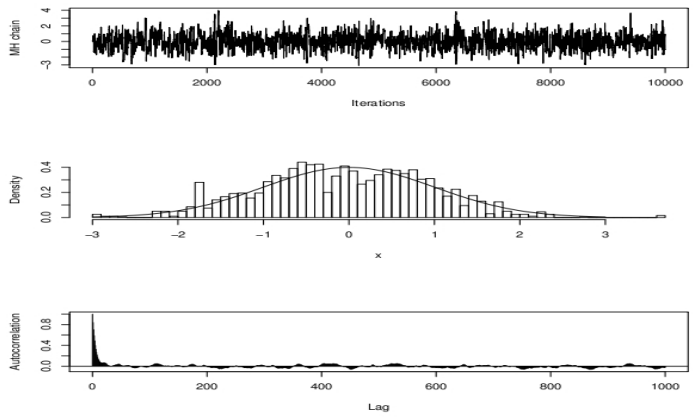

Fig. 4.4. Simulation of a $\mathcal{N}(0,1)$ target with a normal mixture: *top*: sequence of $10,000$ iterations subsampled at every 10-th iteration; *middle*: histogram of the $2,000$ last iterations compared with the target density; *bottom*: empirical autocorrelations using R function plot.acf.

Exercise 4.10 Find conditions on the observed pairs $(\mathbf{x}^i, y_i)$ for the posterior distribution above to be proper.

This distribution is proper (i.e. well-defined) if the integral

$$\mathfrak{I} = \int \prod_{i=1}^{n} \Phi(\mathbf{x}^{i\mathsf{T}}\beta)^{y_i} \left[1 - \Phi(\mathbf{x}^{i\mathsf{T}}\beta)\right]^{1-y_i} \mathrm{d}\beta$$

is finite. If we introduce the latent variable behind $\Phi(\mathbf{x}^{i\mathsf{T}}\beta)$, we get by Fubini that

$$\mathfrak{I} = \int \prod_{i=1}^{n} \varphi(z_i) \int_{\{\beta\,;\mathbf{x}^{i\mathsf{T}}\beta) \gtrless z_i\,,\ i=1,\ldots,n\}} \mathrm{d}\beta\, \mathrm{d}z_1 \cdots \mathrm{d}z_n\,,$$

where $\mathbf{x}^{i\mathsf{T}}\beta \gtrless z_i$ means that the inequality is $\mathbf{x}^{i\mathsf{T}}\beta < z_i$ if $y_i = 1$ and $\mathbf{x}^{i\mathsf{T}}\beta < z_i$ otherwise. Therefore, the inner integral is finite if and only if the set

$$\mathfrak{P} = \left\{\beta\,; \mathbf{x}^{i\mathsf{T}}\beta \gtrless z_i\,,\ i = 1, \ldots, n\right\}$$

is compact. The fact that the whole integral $\mathfrak{I}$ is finite follows from the fact that the volume of the polyhedron defined by $\mathfrak{P}$ grows like $|z_i|^k$ when z_i goes to infinity. This is however a rather less than explicit constraint on the $(\mathbf{x}^i, y_i)$'s!

Exercise 4.11 Include an intercept in the probit analysis of bank and run the corresponding version of Algorithm 4.2 to discuss whether or not the posterior variance of the intercept is high

We simply need to add a column of 1's to the matrix X, as for instance in

```
> X=as.matrix(cbind(rep(1,dim(X)[1]),X))
```

and then use the code provided in the file `#4.txt`, i.e.

```
flatprobit=hmflatprobit(10000,y,X,1)
par(mfrow=c(5,3),mar=1+c(1.5,1.5,1.5,1.5))
for (i in 1:5){
 plot(flatprobit[,i],type="l",xlab="Iterations",
   ylab=expression(beta[i]))
 hist(flatprobit[1001:10000,i],nclass=50,prob=T,main="",
   xlab=expression(beta[i]))
 acf(flatprobit[1001:10000,i],lag=1000,main="",
   ylab="Autocorrelation",ci=F)
}
```

which produces the analysis of bank with an intercept factor. Figure 4.5 gives the equivalent to Figure 4.4 [in the book]. The intercept β_0 has a posterior variance equal to 7558.3, but this must be put in perspective in that the covariates of bank are taking their values in the magnitude of 100 for the three first covariates and of 10 for the last covariate. The covariance of $x_{i1}\beta_1$ is therefore of order 7000 as well. A noticeable difference with Figure 4.4 [in the book] is that, with the inclusion of the intercept, the range of β_1's supported by the posterior is now negative.

Exercise 4.12 Using the latent variable representation of Example 4.2, introduce $z_i|\beta \sim \mathcal{N}\left(\mathbf{x}^{i\mathsf{T}}\beta, 1\right)$ $(1 \leq i \leq n)$ such that $y_i = \mathbb{B}_{z_i \leq 0}$. Deduce that

$$z_i|y_i, \beta \sim \begin{cases} \mathcal{N}_+\left(\mathbf{x}^{i\mathsf{T}}\beta, 1, 0\right) & \text{if} \quad y_i = 1 \\ \mathcal{N}_-\left(\mathbf{x}^{i\mathsf{T}}\beta, 1, 0\right) & \text{if} \quad y_i = 0 \end{cases} \tag{4.2}$$

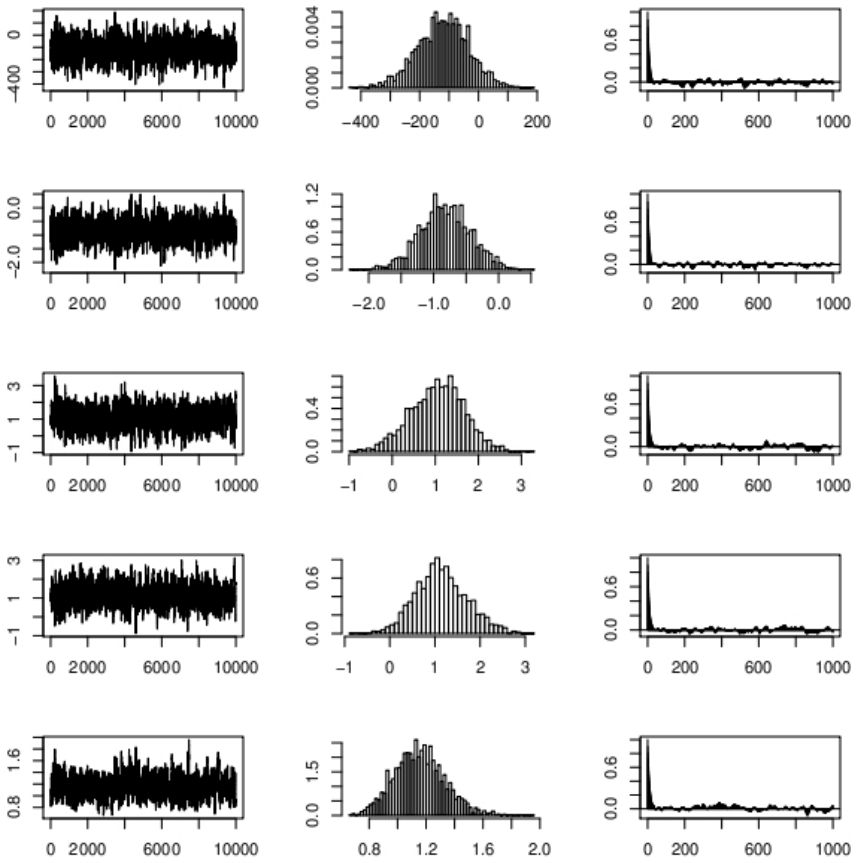

Fig. 4.5. bank: estimation of the probit coefficients [including one intercept β_0] via Algorithm 4.2 and a flat prior. *Left:* β_i's ($i = 0, \ldots, 4$); *center:* histogram over the last $9,000$ iterations; *right:* auto-correlation over the last $9,000$ iterations.

where $\mathcal{N}_+(\mu, 1, 0)$ and $\mathcal{N}_-(\mu, 1, 0)$ are the normal distributions with mean μ and variance 1 that are left-truncated and right-truncated at 0, respectively. Check that those distributions can be simulated using the R commands

```
> xp=qnorm(runif(1)*pnorm(mu)+pnorm(-mu))+mu
> xm=qnorm(runif(1)*pnorm(-mu))+mu
```

Under the flat prior $\pi(\beta) \propto 1$, show that

$$\beta|\mathbf{y}, \mathbf{z} \sim \mathcal{N}_k\left((X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{z}, (X^{\mathsf{T}}X)^{-1}\right) ,$$

where $\mathbf{z} = (z_1, \ldots, z_n)$ and derive the corresponding Gibbs sampler, sometimes called the *Albert–Chib* sampler. (*Hint:* A good starting point is the maximum likelihood estimate of β.) Compare the application to bank with the output in Figure 4.4. (*Note:* Account for computing time differences.)

If $z_i|\beta \sim \mathcal{N}\left(\mathbf{x}^{i\mathsf{T}}\beta, 1\right)$ is a latent [unobserved] variable, it can be related to y_i via the function

$$y_i = \mathbb{I}_{z_i \leq 0}\,,$$

since $P(y_i = 1) = P(z_i \geq 0) = 1 - \Phi\left(-\mathbf{x}^{i\mathsf{T}}\beta\right) = \Phi\left(\mathbf{x}^{i\mathsf{T}}\beta\right)$. The conditional distribution of z_i given y_i is then a constrained normal distribution: if $y_i = 1$, $z_i \leq 0$ and therefore

$$z_i|y_i = 1, \beta \sim \mathcal{N}_+\left(\mathbf{x}^{i\mathsf{T}}\beta, 1, 0\right)\,.$$

(The symmetric case is obvious.)

The command `qnorm(runif(1)*pnorm(mu)+pnorm(-mu))+mu` is a simple application of the inverse cdf transform principle given in Exercise 2.28: the cdf of the $\mathcal{N}_+(\mu, 1, 0)$ distribution is

$$F(x) = \frac{\Phi(x-\mu) - \Phi(-\mu)}{\Phi(\mu)}\,.$$

If we condition on both $\mathbf{z}$ and $\mathbf{y}$ [the conjunction of which is defined as the "completed model"], the y_i's get irrelevant and we are back to a linear regression model, for which the posterior distribution under a flat prior is given in Section 3.3.1 and is indeed $\mathcal{N}_k\left((X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{z}, (X^{\mathsf{T}}X)^{-1}\right)$.

This closed-form representation justifies the introduction of the latent variable $\mathbf{z}$ in the simulation process and leads to the Gibbs sampler that simulates β given $\mathbf{z}$ and $\mathbf{z}$ given β and $\mathbf{y}$ as in (4.2). The R code of this sampler is available in the file `#4.R` as the function `gibbsprobit`. The output of this function is represented on Figure 4.6. Note that the output is somehow smoother than on Figure 4.5. (This does not mean that the Gibbs sampler is converging faster but rather than its component-wise modification of the Markov chain induces slow moves and smooth transitions.)

When comparing the computing times, the increase due to the simulation of the z_i's is not noticeable: for the `bank` dataset, using the codes provided in `#4.txt` require $27s$ and $26s$ over $10,000$ iterations for `hmflatprobit` and `gibbsprobit`. respectively.

Exercise 4.13 Find conditions on $\sum_i y_i$ and on $\sum_i (1 - y_i)$ for the posterior distribution defined by (4.3) to be proper.

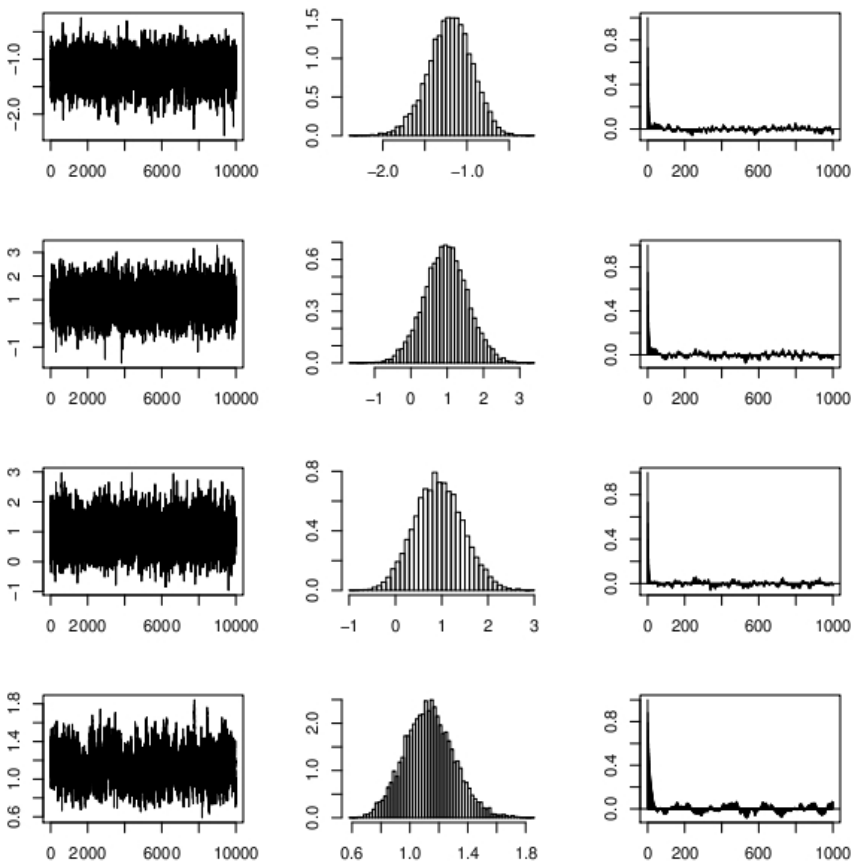

Fig. 4.6. bank: estimation of the probit coefficients [including one intercept β_0] by a Gibbs sampler 4.2 under a flat prior. *Left:* β_i's ($i = 0, \ldots, 4$); *center:* histogram over the last 9,000 iterations; *right:* auto-correlation over the last 9,000 iterations.

There is little difference with Exercise 4.10 because the additional term $\left(\beta^{\mathsf{T}}(X^{\mathsf{T}}X)\beta\right)^{-(2k-1)/4}$ is creating a problem only when β goes to 0. This difficulty is however superficial since the power in $||X\beta||^{(2k-1)/2}$ is small enough to be controlled by the power in $||X\beta||^{k-1}$ in an appropriate polar change of variables. Nonetheless, this is the main reason why we need a $\pi(\sigma^2) \propto \sigma^{-3/2}$ prior rather than the traditional $\pi(\sigma^2) \propto \sigma^{-2}$ which is not controlled in $\beta = 0$. (This is the limiting case, in the sense that the posterior is well-defined for $\pi(\sigma^2) \propto \sigma^{-2+\epsilon}$ for all $\epsilon > 0$.)

Exercise 4.14 For bank, compute the Bayes factor associated with the null hypothesis $H_0 : \beta_2 = \beta_3 = 0$.

The Bayes factor is given by

$$B_{01}^{\pi} = \frac{\pi^{-k/2}\Gamma((2k-1)/4)}{\pi^{-(k-2)/2}\Gamma\{(2k-5)/4\}} \times \frac{\int \left(\beta^{\mathsf{T}}(X^{\mathsf{T}}X)\beta\right)^{-(2k-1)/4} \prod_{i=1}^{n} \Phi(\mathbf{x}^{i\mathsf{T}}\beta)^{y_i} \left[1-\Phi(\mathbf{x}^{i\mathsf{T}}\beta)\right]^{1-y_i} \mathrm{d}\beta}{\int \left\{(\beta^0)^{\mathsf{T}}(X_0^{\mathsf{T}}X_0)\beta^0\right\}^{-(2k-5)/4} \prod_{i=1}^{n} \Phi(x_0^{i\mathsf{T}}\beta^0)^{y_i} \left[1-\Phi(x_0^{i\mathsf{T}}\beta^0)\right]^{1-y_i} \mathrm{d}\beta^0} .$$

For its approximation, we can use simulation from a multivariate normal as suggested in the book or even better from a multivariate $\mathscr{T}$: a direct adaptation from the code in `#4.txt` is

```
noinfprobit=hmnoinfprobit(10000,y,X,1)

library(mnormt)

mkprob=apply(noinfprobit,2,mean)
vkprob=var(noinfprobit)
simk=rmvnorm(100000,mkprob,2*vkprob)
usk=probitnoinflpost(simk,y,X)-
  dmnorm(simk,mkprob,2*vkprob,log=TRUE)

noinfprobit0=hmnoinfprobit(10000,y,X[,c(1,4)],1)
mk0=apply(noinfprobit0,2,mean)
vk0=var(noinfprobit0)
simk0=rmvnorm(100000,mk0,2*vk0)
usk0=probitnoinflpost(simk0,y,X[,c(1,4)])-
  dmnorm(simk0,mk0,2*vk0,log=TRUE)
bf0probit=mean(exp(usk))/mean(exp(usk0))
```

(If a multivariate $\mathscr{T}$ is used, the `dmnorm` function must be replaced with the density of the multivariate $\mathscr{T}$.) The value contained in `bf0probit` is 67.74, which is thus an approximation to B_{10}^{π} [since we divide the approximate marginal under the full model with the approximate marginal under the restricted model]. Therefore, H_0 is quite unlikely to hold, even though, independently, the Bayes factors associated with the componentwise hypotheses $H_0^2 : \beta_2 = 0$ and $H_0^3 : \beta_3 = 0$ support those hypotheses.

Exercise 4.15 Compute the Jacobian $|\partial p_1 \cdots \partial p_k / \partial\beta_1 \cdots \partial\beta_k|$ and deduce that the transform of the prior density $\pi(p_1, \ldots, p_k)$ in the prior density above is correct.

Since $p_i = \Phi(\tilde{\mathbf{x}}^{i\mathsf{T}}\beta)$, we have

$$\frac{\partial}{\partial\beta_j} p_i = \frac{\partial}{\partial\beta_j}\Phi(\tilde{\mathbf{x}}^{i\mathsf{T}}\beta) = x_{ij}\varphi(\tilde{\mathbf{x}}^{i\mathsf{T}}\beta)\,,$$

which means that the Jacobian $\mathfrak{J}$ is the determinant of the matrix made of $\tilde{X}$ multiplied by $\varphi(\tilde{\mathbf{x}}^{i\mathsf{T}}\beta)$ on each row. Therefore,

$$\mathfrak{J} = \prod_{i=1}^{k} \varphi(\tilde{\mathbf{x}}^{i\mathsf{T}}\beta)|X|\,.$$

(Note that this is a very special case where $\tilde{X}$ is a square matrix, hence $|\tilde{X}|$ is well-defined.) Since $|\tilde{X}|$ does not depend on β, it does need to appear in $\pi(\beta)$, i.e.

$$\pi(\beta) \propto \prod_{i=1}^{k} \Phi(\tilde{\mathbf{x}}^{i\mathsf{T}}\beta)^{K_i g_i - 1} \left[1 - \Phi(\tilde{\mathbf{x}}^{i\mathsf{T}}\beta)\right]^{K_i(1-g_i)-1} \varphi(\tilde{\mathbf{x}}^{i\mathsf{T}}\beta)\,.$$

Exercise 4.16 In the case of the logit model, i.e. when $p_i = \exp\tilde{\mathbf{x}}^{i\mathsf{T}}\beta/\{1+\exp\tilde{\mathbf{x}}^{i\mathsf{T}}\beta\}$ $(1 \le i \le k)$, derive the prior distribution on β associated with the prior (4.5) on $(p_1, \ldots, p_k)$.

The only difference with Exercise 4.15 is in the use of a logistic density, hence both the Jacobian and the probabilities are modified:

$$\begin{aligned}\pi(\beta) &\propto \prod_{i=1}^{k} \frac{\exp(\{K_i g_i - 1\}\tilde{\mathbf{x}}^{i\mathsf{T}}\beta)}{\{1+\exp(\tilde{\mathbf{x}}^{i\mathsf{T}}\beta)\}^{K_i-2}} \frac{\exp(\tilde{\mathbf{x}}^{i\mathsf{T}}\beta)}{\{1+\exp(\tilde{\mathbf{x}}^{i\mathsf{T}}\beta)\}^{2}} \\ &= \frac{\exp\left(\sum_{i=1}^{n} K_i g_i \tilde{\mathbf{x}}^{i\mathsf{T}}\beta\right)}{\prod_{i=1}^{k}\left\{1+\exp(\tilde{\mathbf{x}}^{i\mathsf{T}}\beta)\right\}^{K_i}}\,.\end{aligned}$$

Exercise 4.17 Examine whether or not the sufficient conditions for propriety of the posterior distribution found in Exercise 4.13 for the probit model are the same for the logit model.

There is little difference with Exercises 4.10 and 4.13 because the only change is [again] in the use of a logistic density, which has asymptotics similar to the normal density. The problem at $\beta = 0$ is solved in the same manner.

Exercise 4.18 For bank and the logit model, compute the Bayes factor associated with the null hypothesis $H_0 : \beta_2 = \beta_3 = 0$ and compare its value with the value obtained for the probit model in Exercise 4.14.

This is very similar to Exercise 4.14, except that the parameters are now estimated for the logit model. The code is provided in file `#4.txt` as

```
# noninformative prior and random walk HM sample
noinflogit=hmnoinflogit(10000,y,X,1)

# log-marginal under full model
mklog=apply(noinflogit,2,mean)
vklog=var(noinflogit)
simk=rmnorm(100000,mklog,2*vklog)
usk=logitnoinflpost(simk,y,X)-
dmnorm(simk,mklog,2*vklog,log=TRUE)

# noninformative prior and random walk HM sample
# for restricted model
noinflogit0=hmnoinflogit(10000,y,X[,c(1,4)],1)

# log-marginal under restricted model
mk0=apply(noinflogit0,2,mean)
vk0=var(noinflogit0)
simk0=rmnorm(100000,mk0,2*vk0)
usk0=logitnoinflpost(simk0,y,X[,c(1,4)])-
dmnorm(simk0,mk0,2*vk0,log=TRUE)

bf0logit=mean(exp(usk))/mean(exp(usk0))
```

The value of `bf0logit` is 127.2, which, as an approximation to B^{π}_{10}, argues rather strongly against the null hypothesis H_0. It thus leads to the same conclusion as in the probit model of Exercise 4.14, except that the numerical value is almost twice as large. Note that, once again, the Bayes factors associated with the componentwise hypotheses $H_0^2 : \beta_2 = 0$ and $H_0^3 : \beta_3 = 0$ support those hypotheses.

Exercise 4.19 In the case of a 2×2 contingency table with fixed total count $n = n_{11} + n_{12} + n_{21} + n_{22}$, we denote by $\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}$ the corresponding probabilities. If the prior on those probabilities is a Dirichlet $\mathcal{D}_4(1/2, \ldots, 1/2)$, give the corresponding marginal distributions of $\alpha = \theta_{11} + \theta_{12}$ and of $\beta = \theta_{11} + \theta_{21}$. Deduce the associated Bayes factor if H_0 is the hypothesis of independence

between the factors and if the priors on the margin probabilities α and β are those derived above.

A very handy representation of the Dirichlet $\mathcal{D}_k(\delta_1,\ldots,\delta_k)$ distribution is

$$\frac{(\xi_1,\ldots,\xi_k)}{\xi_1+\ldots+\xi_k)} \sim \mathcal{D}_k(\delta_1,\ldots,\delta_k) \quad \text{when} \quad \xi_i \sim \mathcal{G}a(\delta_i,1)\,,\ i=1,\ldots,k\,.$$

Therefore, if

$$(\theta_{11},\theta_{12},\theta_{21},\theta_{22}) = \frac{(\xi_{11},\xi_{12},\xi_{21},\xi_{22})}{\xi_{11}+\xi_{12}+\xi_{21}+\xi_{22}}\,, \xi_{ij} \overset{\text{iid}}{\sim} \mathcal{G}a(1/2,1)\,,$$

then

$$(\theta_{11}+\theta_{12},\theta_{21}+\theta_{22}) = \frac{(\xi_{11}+\xi_{12},\xi_{21}+\xi_{22})}{\xi_{11}+\xi_{12}+\xi_{21}+\xi_{22}}\,,$$

and

$$(\xi_{11}+\xi_{12}),(\xi_{21}+\xi_{22}) \overset{\text{iid}}{\sim} \mathcal{G}a(1,1)$$

implies that α is a $\mathcal{B}e(1,1)$ random variable, that is, a uniform $\mathcal{U}(01,)$ variable. The same applies to β. (Note that α and β are dependent in this representation.)

Since the likelihood under the full model is multinomial,

$$\ell(\theta|\mathcal{T}) = \binom{n}{n_{11}\,n_{12}\,n_{21}} \theta_{11}^{n_{11}}\,\theta_{12}^{n_{12}}\,\theta_{21}^{n_{21}}\,\theta_{22}^{n_{22}}\,,$$

where $\mathcal{T}$ denotes the contingency table [or the dataset $\{n_{11},n_{12},n_{21},n_{22}\}$], the [full model] marginal is

$$\begin{aligned} m(\mathcal{T}) &= \frac{\binom{n}{n_{11}\,n_{12}\,n_{21}}}{\pi^2} \int \theta_{11}^{n_{11}-1/2}\,\theta_{12}^{n_{12}-1/2}\,\theta_{21}^{n_{21}-1/2}\,\theta_{22}^{n_{22}-1/2}\,\mathrm{d}\theta \\ &= \frac{\binom{n}{n_{11}\,n_{12}\,n_{21}}}{\pi^2} \frac{\prod_{i,j}\Gamma(n_{ij}+1/2)}{\Gamma(n+2)} \\ &= \frac{\binom{n}{n_{11}\,n_{12}\,n_{21}}}{\pi^2} \frac{\prod_{i,j}\Gamma(n_{ij}+1/2)}{(n+1)!} \\ &= \frac{1}{(n+1)\pi^2} \prod_{i,j} \frac{\Gamma(n_{ij}+1/2)}{\Gamma(n_{ij}+1)}\,, \end{aligned}$$

where the π^2 term comes from $\Gamma(1/2)=\sqrt{\pi}$.

In the restricted model, θ_{11} is replaced with $\alpha\beta$, θ_{12} by $\alpha(1-\beta)$, and so on. Therefore, the likelihood under the restricted model is the product

$$\binom{n}{n_{1\cdot}}\alpha^{n_{1\cdot}}(1-\alpha)^{n-n_{1\cdot}} \times \binom{n}{n_{\cdot 1}}\beta^{n_{\cdot 1}}(1-\beta)^{n-n_{\cdot 1}}\,,$$

where $n_{1\cdot} = n_{11} + n_{12}$ and $n_{\cdot 1} = n_{11} + n_{21}$, and the restricted marginal under uniform priors on both α and β is

$$\begin{aligned} m_0(\mathcal{T}) &= \binom{n}{n_{1\cdot}}\binom{n}{n_{\cdot 1}} \int_0^1 \alpha^{n_{1\cdot}}(1-\alpha)^{n-n_{1\cdot}}\,\mathrm{d}\alpha \int_0^1 \beta^{n_{\cdot 1}}(1-\beta)^{n-n_{\cdot 1}}\,\mathrm{d}\beta \\ &= \binom{n}{n_{1\cdot}}\binom{n}{n_{\cdot 1}} \frac{(n_{1\cdot}+1)!(n-n_{1\cdot}+1)!}{(n+2)!} \frac{(n_{\cdot 1}+1)!(n-n_{\cdot 1}+1)!}{(n+2)!} \\ &= \frac{(n_{1\cdot}+1)(n-n_{1\cdot}+1)}{(n+2)(n+1)} \frac{(n_{\cdot 1}+1)(n-n_{\cdot 1}+1)}{(n+2)(n+1)}\,. \end{aligned}$$

The Bayes factor B_{01}^{π} is then the ratio $m_0(\mathcal{T})/m(\mathcal{T})$.

Exercise 4.20 Given a contingency table with four categorical variables, determine the number of submodels to consider.

Note that the numbers of classes for the different variables do not matter since, when building a non-saturated submodel, a variable is in or out. There are

1. 2^4 single-factor models [including the zero-factor model];
2. $(2^6 - 1)$ two-factor models [since there are $\binom{4}{2} = 6$ ways of picking a pair of variables out of 4 and since the complete single-factor model is already treated];
3. $(2^4 - 1)$ three-factor models.

Thus, if we exclude the saturated model, there are $2^6 + 2^5 - 2 = 94$ different submodels.

Exercise 4.21 Find sufficient conditions on $(\mathbf{y}, X)$ for this posterior distribution proportional to be proper.

First, as in Exercise 4.13, the term $\left(\beta^{\mathsf{T}}(X^{\mathsf{T}}X)\beta\right)^{-(2k-1)/4}$ is not a major problem when β goes to 0, since it is controlled by the power in the Jacobian $||X\beta||^{k-1}$ in an adapted polar change of variables. Moreover, if the matrix X of regressors is of full rank k, then, when β_j $(j = 1, \ldots, k)$ goes to $\pm\infty$, there exists at least one $1 \le i \le n$ such that $\mathbf{x}^{i\mathsf{T}}\beta$ goes to either $+\infty$ or $-\infty$. In the former case, the whole exponential term goes to 0, while, in the later case, it depends on y_i. For instance, if all y_i's are equal to 1, the above quantity is integrable.

5

Capture–Recapture Experiments

Exercise 5.1 Show that the posterior distribution $\pi(N|n^+)$ given by (5.1) while associated with an improper prior, is defined for all values of n^+. Show that the normalization factor of (5.1) is $n^+ \vee 1$ and deduce that the posterior median is equal to $2(n^+ \vee 1)$. Discuss the relevance of this estimator and show that it corresponds to a Bayes estimate of p equal to $1/2$.

Since the main term of the series is equivalent to N^{-2}, the series converges. The posterior distribution can thus be normalised. Moreover,

$$\begin{aligned}\sum_{i=n_0}^{\infty} \frac{1}{i(i+1)} &= \sum_{i=n_0}^{\infty} \left(\frac{1}{i} - \frac{1}{i+1}\right) \\ &= \frac{1}{n_0} - \frac{1}{n_0+1} + \frac{1}{n_0+1} - \frac{1}{n_0+2} + \dots \\ &= \frac{1}{n_0}\,.\end{aligned}$$

Therefore, the normalisation factor is available in closed form and is equal to $n^+ \vee 1$. The posterior median is the value $N^\star$ such that $\pi(N \geq N^\star|n^+) = 1/2$, i.e.

$$\begin{aligned}\sum_{i=N^\star}^{\infty} \frac{1}{i(i+1)} &= \frac{1}{2}\,\frac{1}{n^+ \vee 1} \\ &= \frac{1}{N^\star}\,,\end{aligned}$$

which implies that $N^\star = 2(n^+ \vee 1)$. This estimator is rather intuitive in that $\mathbb{E}[n^+|N,p] = pN$: since the expectation of p is $1/2$, $\mathbb{E}[n^+|N] = N/2$ and $N^\star = 2n^+$ is a moment estimator of N.

Exercise 5.2 Under the prior $\pi(N,p) \propto N^{-1}$, derive the marginal posterior density of N in the case where $n_1^+ \sim \mathscr{B}(N,p)$ and where $k-1$ iid observations

$$n_2^+, \ldots, n_k^+ \overset{\text{iid}}{\sim} \mathscr{B}(n_1^+, p)$$

are observed (the later are in fact recaptures). Apply to the sample

$$(n_1^+, n_2^+, \ldots, n_{11}^+) = (32, 20, 8, 5, 1, 2, 0, 2, 1, 1, 0)\,,$$

which describes a series of tag recoveries over 11 years.

In that case, if we denote $n_\cdot^+ = n_1^+ + \cdots + n_k^+$ the total number of captures, the marginal posterior density of N is

$$\begin{aligned}\pi(N|n_1^+,\ldots,n_k^+) &\propto \frac{N!}{(N-n_1^+)!}\, N^{-1}\mathbb{I}_{N\geq n_1^+} \\ &\qquad \int_0^1 p^{n_1^+ + \cdots + n_k^+}(1-p)^{N-n_1^+ + (n_1 + -n_2^+ + \cdots + n_1^+ - n_k^+}\,\mathrm{d}p \\ &\propto \frac{(N-1)!}{(N-n_1^+)!}\,\mathbb{I}_{N\geq n_1^+}\int_0^1 p^{n_\cdot^+}(1-p)^{N+kn_1^+-n_\cdot^+}\,\mathrm{d}p \\ &\propto \frac{(N-1)!}{(N-n_1^+)!}\,\frac{(N+kn_1^+-n_\cdot^+)!}{(N+kn_1^++1)!}\,\mathbb{I}_{N\geq n_1^+\vee 1}\,,\end{aligned}$$

which does not simplify any further. Note that the binomial coefficients

$$\binom{n_1^+}{n_j^+} \qquad (j \geq 2)$$

are irrelevant for the posterior of N since they only depend on the data.

The R code corresponding to this model is as follows:

```
n1=32
ndo=sum(32,20,8,5,1,2,0,2,1,1,0)

# unnormalised posterior
post=function(N){
   exp(lfactorial(N-1)+lfactorial(N+11*n1-ndo)-
    lfactorial(N-n1)-lfactorial(N+11*n1+1))
   }

# normalising constant and
# posterior mean
```

```
posv=post((n1:10000))

cons=sum(posv)
pmean=sum((n1:10000)*posv)/cons
pmedi=sum(cumsum(posv)<.5*cons)
```

The posterior mean is therefore equal to 282.4, while the posterior median is 243. Note that a crude analysis estimating p by $\hat{p} = (n_2^+ + \ldots + n_{11})/(10n_1^+) = 0.125$ and N by $n_1^+/\hat{p}$ would produce the value $\hat{N} = 256$.

Exercise 5.3 For the two-stage capture-recapture model, show that the distribution of m_2 conditional on both samples sizes n_1 and n_2 is given by (5.2) and does not depend on p. Deduce the expectation $\mathbb{E}[m_2|n_1, n_2, N]$.

Since

$$n_1 \sim \mathscr{B}(N, p)\,, \quad m_2|n_1 \sim \mathscr{B}(n_1, p)$$

and

$$n_2 - m_2|n_1, m_2 \sim \mathscr{B}(N - n_1, p)\,,$$

the conditional distribution of m_2 is given by

$$\begin{aligned} f(m_2|n_1, n_2) &\propto \binom{n_1}{m_2} p^{m_2}(1-p)^{n_1-m_2} \binom{N-n_1}{n_2-m_2} p^{n_2-m_2}(1-p)^{N-n_1-n_2+m_2} \\ &\propto \binom{n_1}{m_2}\binom{N-n_1}{n_2-m_2} p^{m_2+n_2-m_2}(1-p)^{n_1-m_2+N-n_1-n_2+m_2} \\ &\propto \binom{n_1}{m_2}\binom{N-n_1}{n_2-m_2} \\ &\propto \frac{\binom{n_1}{m_2}\binom{N-n_1}{n_2-m_2}}{\binom{N}{n_2}}\,, \end{aligned}$$

which is the hypergeometric $\mathscr{H}(N, n_2, n_1/N)$ distribution. Obviously, this distribution does not depend on p and its expectation is

$$\mathbb{E}[m_2|n_1, n_2] = \frac{n_1 n_2}{N}\,.$$

Exercise 5.4 In order to determine the number N of buses in a town, a capture–recapture strategy goes as follows. We observe $n_1 = 20$ buses during the first day and keep track of their identifying numbers. Then we repeat the experiment the following day by recording the number of buses that have already been spotted on the previous day, say $m_2 = 5$, out of the $n_2 = 30$ buses observed the second day. For the Darroch model, give the posterior expectation of N under the prior $\pi(N) = 1/N$.

Using the derivations of the book, we have that

$$\pi(N|n_1,n_2,m_2) \propto \frac{1}{N}\binom{N}{n^+} B(n^c+1,2N-n^c+1)\mathbb{I}_{N\geq n^+}$$
$$\propto \frac{(N-1)!}{(N-n^+)!}\,\frac{(2N-n^c)!}{(2N+1)!}\,\mathbb{I}_{N\geq n^+}$$

with $n^+ = 45$ and $n^c = 50$. For $n^+ = 45$ and $n^c = 50$, the posterior mean [obtained by an R code very similar to the one of Exercise 5.2] is equal to 130.91.

Exercise 5.5 Show that the maximum likelihood estimator of N for the Darroch model is $\hat{N} = n_1/(m_2/n_2)$ and deduce that it is not defined when $m_2 = 0$.

The likelihood for the Darroch model is proportional to

$$\ell(N) = \frac{(N-n_1)!}{(N-n_2)!}\,\frac{(N-n^+)!}{N!}\,\mathbb{I}_{N\geq n^+}\,.$$

Since

$$\frac{\ell(N+1)}{\ell(N)} = \frac{(N+1-n_1)(N+1-n_2)}{(N+1-n^+)(N+1)} \geq 1$$

for

$$(N+1)^2-(N+1)(n_1+n_2)+n_1n_2 \geq (N+1)^2-(N+1)n^+$$
$$(N+1)(n_1+n_2-n^+) \geq n_1n_2$$
$$(N+1) \leq \frac{n_1n_2}{m_2}\,,$$

the likelihood is increasing for $N \leq n_1n_2/m2$ and decreasing for $N \geq n_1n_2/m2$. Thus $\hat{N} = n_1n_2/m2$ is the maximum likelihood estimator [assuming this quantity is an integer]. If $m_2 = 0$, the likelihood is increasing with N and therefore there is no maximum likelihood estimator.

Exercise 5.6 Give the likelihood of the extension of Darroch's model when the capture–recapture experiments are repeated K times with capture sizes and recapture observations n_k $(1 \leq k \leq K)$ and m_k $(2 \leq k \leq K)$, respectively. (*Hint*: Exhibit first the two-dimensional sufficient statistic associated with this model.)

When extending the two-stage capture-recapture model to a K-stage model, we observe K capture episodes, with $n_i \sim \mathscr{B}(N,p)$ $(1 \leq i \leq K)$, and $K-1$ recaptures,

$$m_i|n_1, n_2, m_2, \ldots, n_{i-1}, m_{i-1}, n_i \sim \mathscr{H}(N, n_i, n_1 + n_2 - m_2 + \cdots - m_{i-1})\,.$$

The likelihood is therefore

$$\prod_{i=1}^{K} \binom{N}{n_i} p^{n_i}(1-p)^{N-n_i} \prod_{i=2}^{K} \frac{\binom{n_1-m_2+\cdots-m_{i-1}}{m_i}\binom{N-n_1+\cdots+m_{i-1}}{n_i-m_i}}{\binom{N}{n_i}}$$
$$\propto \frac{N!}{(N!-n^+)!}\, p^{n^c}(1-p)^{KN-n^c}\,,$$

where $n^+ = n_1 - m_2 + \cdots - m_K$ is the number of captured individuals and where $n^c = n_1 + \cdots + n_K$ is the number of captures. These two statistics are thus sufficient for the K-stage capture-recapture model.

Exercise 5.7 Give both conditional posterior distributions in the case $n^+ = 0$.

When $n^+ = 0$, there is no capture at all during both capture episodes. The likelihood is thus $(1-p)^{2N}$ and, under the prior $\pi(N,p) = 1/N$, the conditional posterior distributions of p and N are

$$p|N, n^+ = 0 \sim \mathscr{B}e(1, 2N+1)\,,$$
$$N|p, n^+ = 0 \sim \frac{(1-p)^{2N}}{N}\,.$$

That the joint distribution $\pi(N,p|n^+ = 0)$ exists is ensured by the fact that $\pi(N|n^+ = 0) \propto 1/N(2N+1)$, associated with a converging series.

Exercise 5.8 Show that, when the prior on N is a $\mathscr{P}(\lambda)$ distribution, the conditional posterior on $N - n_+$ is $\mathscr{P}(\lambda(1-p)^2)$.

The posterior distribution of (N,p) associated with the informative prior $\pi(N,p) = \lambda^N e^{-\lambda}/N!$ is proportional to

$$\frac{N!}{(N-n^+)!N!}\,\lambda^N\, p^{n^c}(1-p)^{2N-n^c}\,\mathbb{I}_{N\geq n^+}\,.$$

The corresponding conditional on N is thus proportional to

$$\frac{\lambda^N}{(N-n^+)!}\, p^{n^c}(1-p)^{2N-n^c}\,\mathbb{I}_{N\geq n^+} \propto \frac{\lambda^{N-n^+}}{(N-n^+)!}\, p^{n^c}(1-p)^{2N-n^c}\,\mathbb{I}_{N\geq n^+}$$

which corresponds to a Poisson $\mathscr{P}(\lambda(1-p)^2)$ distribution on $N - n_+$.

Exercise 5.9 An extension of the T-stage capture-recapture model is to consider that the capture of an individual modifies its probability of being captured from p to q for future captures. Give the likelihood $\ell(N, p, q|n_1, n_2, m_2 \ldots, n_T, m_T)$.

When extending the T-stage capture-recapture model with different probabilities of being captured and recaptured, after the first capture episode, where $n_1 \sim \mathscr{B}(N, p)$, we observe $T-1$ new captures ($i = 2, \ldots, T$)

$$n_i - m_i | n_1, n_2, m_2, \ldots, n_{i-1}, m_{i-1} \sim \mathscr{B}(N - n_1 - n_2 + m_2 + \ldots + m_{i-1}, p)\,,$$

and $T-1$ recaptures ($i = 2, \ldots, T$),

$$m_i | n_1, n_2, m_2, \ldots, n_{i-1}, m_{i-1} \sim \mathscr{B}(n_1 + n_2 - m_2 + \ldots - m_{i-1}, q)\,.$$

The likelihood is therefore

$$\begin{aligned}
\binom{N}{n_1} p^{n_1}(1-p)^{N-n_1} \prod_{i=2}^{T} \binom{N - n_1 + \ldots - m_{i-1}}{n_i - m_i} p^{n_i - m_i}(1-p)^{N - n_1 + \ldots + m_i} \\
\times \prod_{i=2}^{T} \binom{n_1 + n_2 - \ldots - m_{i-1}}{m_i} q^{m_i}(1-q)^{n_1 + \ldots - m_i} \\
\propto \frac{N!}{(N-n^+)!}\, p^{n^+}(1-p)^{TN - n^*}\, q^{m^+}(1-q)^{n^* - n_1},
\end{aligned}$$

where $n^+ = n_1 - m_2 + \cdots - m_T$ is the number of captured individuals,

$$n^* = T n_1 + \sum_{j=2}^{T} (T - j + 1)(n_j - m_j)$$

and where $m^+ = m_1 + \cdots + m_T$ is the number of recaptures. The four statistics (n_1, n^+, n^*, m^+) are thus sufficient for this version of the T-stage capture-recapture model.

Exercise 5.10 Another extension of the 2-stage capture-recapture model is to allow for mark losses. If we introduce q as the probability of losing the mark, r as the probability of recovering a lost mark and k as the number of recovered lost marks, give the associated likelihood $\ell(N, p, q, r|n_1, n_2, m_2, k)$.

There is an extra-difficulty in this extension in that it contains a latent variable: let us denote by z the number of tagged individuals that have lost

their mark. Then $z \sim \mathscr{B}(n_1, q)$ is not observed, while $k \sim \mathscr{B}(z, r)$ is observed. Were we to observe (n_1, n_2, m_2, k, z), the [completed] likelihood would be

$$\ell^\star(N,p,q,r|n_1,n_2,m_2,k,z) = \binom{N}{n_1} p^{n_1}(1-p)^{N-n_1} \binom{n_1}{z} q^z (1-q)^{n_1-z} \times \binom{z}{k} r^k (1-r)^{z-k} \binom{n_1-z}{m_2} p^{m_2}(1-p)^{n_1-z-m_2} \times \binom{N-n_1+z}{n_2-m_2} p^{n_2-m_2}(1-p)^{N-n_1+z-n_2+m_2},$$

since, for the second round, the population gets partitioned into individuals that keep their tag and are/are not recaptured, those that loose their tag and are/are not recaptured, and those that are captured for the first time. Obviously, it is not possible to distinguish between the two last categories. Since z is not known, the [observed] likelihood is obtained by summation over z:

$$\ell(N,p,q,r|n_1,n_2,m_2,k) \propto \frac{N!}{(N-n_1)!} p^{n_1+n_2}(1-p)^{2N-n_1-n_2} \sum_{z=k\vee N-n_1-n_2+m_2}^{n_1-m_2} \binom{n_1}{z} \binom{n_1-z}{m_2} \times \binom{N-n_1+z}{n_2-m_2} q^z (1-q)^{n_1-z} r^k (1-r)^{z-k}.$$

Note that, while a proportionality sign is acceptable for the computation of the likelihood, the terms depending on z must be kept within the sum to obtain the correct expression for the distribution of the observations. A simplified version is thus

$$\ell(N,p,q,r|n_1,n_2,m_2,k) \propto \frac{N!}{(N-n_1)!} p^{n_1+n_2}(1-p)^{2N-n_1-n_2} q^{n_1}(r/(1-r))^k \sum_{z=k\vee N-n_1-n_2+m_2}^{n_1-m_2} \frac{(N-n_1+z)![q(1-r)/(1-q)]^z}{z!(n_1-z-m_2)!(N-n_1-n_2+m_2+z)!},$$

but there is no close-form solution for the summation over z.

Exercise 5.11 Reproduce the analysis of eurodip when switching the prior from $\pi(N,p) \propto \lambda^N/N!$ to $\pi(N,p) \propto N^{-1}$.

The main purpose of this exercise is to modify the code of the function `gibbs1` in the file `#5.R` on the webpage, since the marginal posterior distribution of N is given in the book as

$$\pi(N|n^+,n^c) \propto \frac{(N-1)!}{(N-n^+)!}\,\frac{(TN-n^c)!}{(TN+1)!}\,\mathbb{I}_{N\geq n^+\vee 1}\,.$$

(The conditional posterior distribution of p does not change.) This distribution being non-standard, it makes direct simulation awkward and we prefer to use a Metropolis-Hastings step, using a modified version of the previous Poisson conditional as proposal $q(N'|N,p)$. We thus simulate

$$N^\star - n^+ \sim \mathscr{P}\left(N^{(t-1)}(1-p^{(t-1)})^T\right)$$

and accept this value with probability

$$\frac{\pi(N^\star|n^+,n^c)}{\pi(N^{(t-1)}|n^+,n^c)}\,\frac{q(N^{(t-1)}|N^\star,p^{(t-1)})}{q(N^\star|N^{(t-1)},p^{(t-1)})}\wedge 1\,.$$

The corresponding modified R function is

```
gibbs11=function(nsimu,T,nplus,nc)
{
# conditional posterior
rati=function(N){
  lfactorial(N-1)+lfactorial(T*N-nc)-
    lfactorial(N-nplus)-lfactorial(T*N+1)
  }

N=rep(0,nsimu)
p=rep(0,nsimu)

N[1]=2*nplus
p[1]=rbeta(1,nc+1,T*N[1]-nc+1)
for (i in 2:nsimu){

  # MH step on N
  N[i]=N[i-1]
  prop=nplus+rpois(1,N[i-1]*(1-p[i-1])^T)
  if (log(runif(1))<rati(prop)-rati(N[i])+
        dpois(N[i-1]-nplus,prop*(1-p[i-1])^T,log=T)-
        dpois(prop-nplus,N[i-1]*(1-p[i-1])^T,log=T))
     N[i]=prop
  p[i]=rbeta(1,nc+1,T*N[i]-nc+1)
  }
list(N=N,p=p)
}
```

The output of this program is given in Figure 5.1.

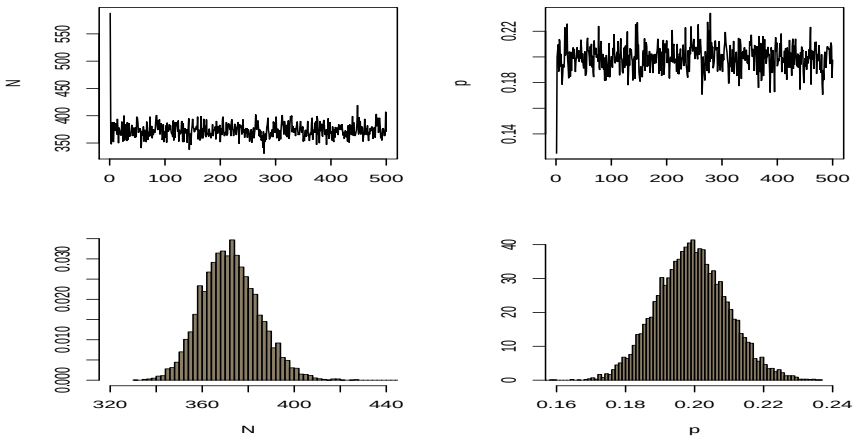

Fig. 5.1. eurodip: MCMC simulation under the prior $\pi(N,p) \propto N^{-1}$.

Exercise 5.12 Show that the conditional distribution of r_1 is indeed proportional to the product (5.4).

The joint distribution of $\mathcal{D}^* = (n_1, c_2, c_3, r_1, r_2)$ is given in the book as

$$\binom{N}{n_1} p^{n_1}(1-p)^{N-n_1} \binom{n_1}{r_1} q^{r_1}(1-q)^{n_1-r_1} \binom{n_1-r_1}{c_2} p^{c_2}(1-p)^{n_1-r_1-c_2} \times \binom{n_1-r_1}{r_2} q^{r_2}(1-q)^{n_1-r_1-r_2} \binom{n_1-r_1-r_2}{c_3} p^{c_3}(1-p)^{n_1-r_1-r_2-c_3}\,.$$

Therefore, if we only keep the terms depending on r_1, we indeed recover

$$\begin{aligned}
&\frac{1}{r_1!(n_1-r_1)!}\, q^{r_1}(1-q)^{n_1-r_1}\, \frac{(n_1-r_1)!}{(n_1-r_1-c_2)!}\,(1-p)^{n_1-r_1-c_2} \\
&\quad\times \frac{(n_1-r_1)!}{(n_1-r_1-r_2)!}\,(1-q)^{n_1-r_1-r_2}\, \frac{(n_1-r_1-r_2)!}{(n_1-r_1-r_2-c_3)!}\,(1-p)^{n_1-r_1-r_2-c_3} \\
&\propto \frac{(n_1-r_1)!}{r_1!(n_1-r_1-c_2)!(n_1-r_1-r_2-c_3)!} \left\{\frac{q}{(1-q)^2(1-p)^2}\right\}^{r_1} \\
&\propto \binom{n_1-c_2}{r_1}\binom{n_1-r_1}{r_2+c_3} \left\{\frac{q}{(1-q)^2(1-p)^2}\right\}^{r_1},
\end{aligned}$$

under the constraint that $r_1 \le \min(n_1, n_1-r_2, n_1-r_2-c_3, n_1-c_2) = \min(n_1-r_2-c_3, n_1-c_2)$.

Exercise 5.13 Show that r_2 can be integrated out in the above joint distribution and leads to the following distribution on r_1:

$$\pi(r_1|p,q,n_1,c_2,c_3) \propto \frac{(n_1-r_1)!(n_1-r_1-c_3)!}{r_1!(n_1-r_1-c_2)!} \times \left(\frac{q}{(1-p)(1-q)[q+(1-p)(1-q)]}\right)^{r_1}. \tag{5.1}$$

Compare the computational cost of a Gibbs sampler based on this approach with a Gibbs sampler using the full conditionals.

Following the decomposition of the likelihood in the previous exercise, the terms depending on r_2 are

$$\frac{1}{r_2!(n_1-r_1-r_2)!}\left\{\frac{q}{(1-p)(1-q)}\right\}^{r_2}\frac{(n_1-r_1-r_2)!}{(n_1-r_1-r_2-c_3)!} = \frac{1}{r_2!(n_1-r_1-r_2-c_3)!}\left\{\frac{q}{(1-p)(1-q)}\right\}^{r_2}.$$

If we sum over $0 \le r_2 \le n_1-r_1-c_3$, we get

$$\frac{1}{(n_1-r_1-c_3)!}\sum_{k=0}^{n_1-r_1-c_3}\binom{n_1-r_1-c_3}{k}\left\{\frac{q}{(1-p)(1-q)}\right\}^{k} = \left\{1+\frac{q}{(1-p)(1-q)}\right\}^{n_1-r_1-c_3}$$

that we can agregate with the remaining terms in r_1

$$\frac{(n-r_1)!}{r_1!(n_1-r_1-c_2)!}\left\{\frac{q}{(1-q)^2(1-p)^2}\right\}^{r_1}$$

to recover (5.1).

Given that a Gibbs sampler using the full conditionals is simulating from standard distributions while a Gibbs sampler based on this approach requires the simulation of this non-standard distribution on r_1, it appears that one iteration of the latter is more time-consuming than for the former.

Exercise 5.14 Show that the likelihood associated with an open population can be written as

$$\ell(N,p|\mathcal{D}^*) = \sum_{(\epsilon_{it},\delta_{it})_{it}}\prod_{t=1}^{T}\prod_{i=1}^{N} q_{\epsilon_{i(t-1)}}^{\epsilon_{it}}(1-q_{\epsilon_{i(t-1)}})^{1-\epsilon_{it}} \times p^{(1-\epsilon_{it})\delta_{it}}(1-p)^{(1-\epsilon_{it})(1-\delta_{it})},$$

where $q_0 = q$, $q_1 = 1$, and δ_{it} and ϵ_{it} are the capture and exit indicators, respectively. Derive the order of complexity of this likelihood, that is, the number of elementary operations necessary to compute it.

This is an alternative representation of the model where each individual capture and life history is considered explicitely. This is also the approach adopted for the Arnason-Schwarz model of Section 5.5. We can thus define the history of individual $1 \leq i \leq N$ as a pair of sequences (ϵ_{it}) and (δ_{it}), where $\epsilon_{it} = 1$ at the exit time t and forever after. For the model given at the beginning of Section 5.3, there are n_1 δ_{i1}'s equal to 1, r_1 ϵ_{i1}'s equal to 1, c_2 δ_{i2}'s equal to 1 among the i's for which $\delta_{i1} = 1$ and so on. If we do not account for these constraints, the likelihood is of order $\mathrm{O}(3^{NT})$ [there are three possible cases for the pair $(\epsilon_{it}, \delta_{it})$ since $\delta_{it} = 0$ if $\epsilon_{it} = 1$]. Accounting for the constraints on the total number of δ_{it}'s equal to 1 increases the complexity of the computation.

Exercise 5.15 Show that, for $M > 0$, if g is replaced with Mg in $\mathscr{S}$ and if (X, U) is uniformly distributed on $\mathscr{S}$, the marginal distribution of X is still g. Deduce that the density g only needs to be known up to a normalizing constant.

The set

$$\mathscr{S} = \{(x, u) : 0 < u < Mg(x)\}$$

has a surface equal to M. Therefore, the uniform distribution on $\mathscr{S}$ has density $1/M$ and the marginal of X is given by

$$\int \mathbb{I}_{(0,Mg(x))} \, \frac{1}{M} \, \mathrm{d}u = \frac{Mg(x)}{M} = g(x) \, .$$

This implies that uniform simulation in $\mathscr{S}$ provides an output from g no matter what the constant M is. In other words, g does not need to be normalised.

Exercise 5.16 For the function $g(x) = (1+\sin^2(x))(2+\cos^4(4x)) \exp[-x^4\{1+\sin^6(x)\}]$ on $[0, 2\pi]$, examine the feasibility of running a uniform sampler on the associated set $\mathscr{S}$.

The function g is non-standard but it is bounded [from above] by the function $\overline{g}(x) = 6 \exp[-x^4]$ since both cos and sin are bounded by 1 or even $\overline{g}(x) = 6$. Simulating uniformly over the set $\mathscr{S}$ associated with g can thus be achieved by simulating uniformly over the set $\mathscr{S}$ associated with $\overline{g}$ until the

output falls within the set $\mathscr{S}$ associated with g. This is the basis of accept-reject algorithms.

Exercise 5.17 Show that the probability of acceptance in Step 2 of Algorithm 5.2 is $1/M$, and that the number of trials until a variable is accepted has a geometric distribution with parameter $1/M$. Conclude that the expected number of trials per simulation is M.

The probability that $U \leq g(X)/(Mf(X))$ is the probability that a uniform draw in the set

$$\mathscr{S} = \{(x, u) : 0 < u < Mg(x)\}$$

falls into the subset

$$\mathscr{S}_0 = \{(x, u) : 0 < u < f(x)\}.$$

The surfaces of $\mathscr{S}$ and $\mathscr{S}_0$ being M and 1, respectively, the probability to fall into $\mathscr{S}_0$ is $1/M$.

Since steps 1. and 2. of Algorithm 5.2 are repeated independently, each round has a probability $1/M$ of success and the rounds are repeated till the first success. The number of rounds is therefore a geometric random variable with parameter $1/M$ and expectation M.

Exercise 5.18 For the conditional distribution of α_t derived from (5.3), construct an Accept-Reject algorithm based on a normal bounding density f and study its performances for $N = 53$, $n_t = 38$, $\mu_t = -0.5$, and $\sigma^2 = 3$.

That the target is only known up to a constant is not a problem, as demonstrated in Exercise 5.15. To find a bound on $\pi(\alpha_t|N, n_t)$ [up to a constant], we just have to notice that

$$(1 + e^{\alpha_t})^{-N} < e^{-N\alpha_t}$$

and therefore

$$
\begin{aligned}
(1+e^{\alpha_t})^{-N} \exp\left\{\alpha_t n_t - \frac{1}{2\sigma^2}(\alpha_t-\mu_t)^2\right\} & \\
&\leq \exp\left\{\alpha_t(n_t-N) - \frac{1}{2\sigma^2}(\alpha_t-\mu_t)^2\right\} \\
&= \exp\left\{-\frac{\alpha_t^2}{2\sigma^2} + 2\frac{\alpha_t}{2\sigma^2}(\mu_t-\sigma^2(N-n_t)) - \frac{\mu_t^2}{2\sigma^2}\right\} \\
&= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha_t-\mu_t+\sigma^2(N-n_t))^2\right\} \\
&\quad \times \sqrt{2\pi}\sigma\exp\left\{-\frac{1}{2\sigma^2}(\mu_t^2-[\mu_t-\sigma^2(N-n_t)]^2)\right\}.
\end{aligned}
$$

The upper bound thus involves a normal $\mathcal{N}(\mu_t-\sigma^2(N-n_t),\sigma^2)$ distribution and the corresponding constant. The R code associated with this decomposition is

```
# constants
N=53
nt=38
mut=-.5
sig2=3
sig=sqrt(sig2)

# log target
ta=function(x){
  -N*log(1+exp(x))+x*nt-(x-mut)^2/(2*sig2)
  }

#bounding constant
bmean=mut-sig2*(N-nt)
uc=0.5*log(2*pi*sig2)+(bmean^2-mut^2)/(2*sig2)

prop=rnorm(1,sd=sig)+bmean
ratio=ta(prop)-uc-dnorm(prop,mean=bmean,sd=sig,log=T)

while (log(runif(1))>ratio){

  prop=rnorm(1,sd=sig)+bmean
  ratio=ta(prop)-uc-dnorm(prop,mean=bmean,sd=sig,log=T)
  }
```

The performances of this algorithm degenerate very rapidly when $N-n_t$ is [even moderately] large.

Exercise 5.19 When uniform simulation on $\mathscr{S}$ is impossible, construct a Gibbs sampler based on the conditional distributions of u and x. (*Hint*: Show that both conditionals are uniform distributions.) This special case of the Gibbs sampler is called the *slice sampler* (see Robert and Casella, 2004, Chapter 8). Apply to the distribution of Exercise 5.16.

Since the joint distribution of (X, U) has the constant density

$$t(x, u) = \mathbb{I}_{0 \le u \le g(x)}\,,$$

the conditional distribution of U given $X = x$ is $\mathscr{U}(0, g(x))$ and the conditional distribution of X given $U = u$ is $\mathscr{U}(\{x; g(x) \ge u\})$, which is uniform over the set of highest values of g. Both conditionals are therefore uniform and this special Gibbs sampler is called the *slice sampler*. In some settings, inverting the condition $g(x) \ge u$ may prove formidable!

If we take the case of Exercise 5.16 and of $\overline{g}(x) = \exp(-x^4)$, the set $\{x; \overline{g}(x) \ge u\}$ is equal to

$$\{x; \overline{g}(x) \ge u\} = \left\{x; x \le (-\log(x))^{1/4}\right\},$$

which thus produces a closed-form solution.

Exercise 5.20 Reproduce the above analysis for the marginal distribution of r_1 computed in Exercise 5.13.

The only change in the codes provided in `#5.R` deals with `seuil`, called by `ardipper`, and with `gibbs2` where the simulation of r_2 is no longer required.

Exercise 5.21 Show that, given a mean and a 95% confidence interval in $[0, 1]$, there exists at most one beta distribution $\mathscr{B}(a, b)$ with such a mean and confidence interval.

If $0 < m < 1$ is the mean $m = a/(a + b)$ of a beta $\mathscr{B}e(a, b)$ distribution, then this distribution is necessarily a beta $\mathscr{B}e(\alpha m, \alpha(1 - m))$ distribution, with $\alpha > 0$. For a given confidence interval $[\ell, u]$, with $0 < \ell < m < u < 1$, we have that

$$\lim_{\alpha \to 0} \int_\ell^u \frac{\Gamma(\alpha)}{\Gamma(\alpha m)\Gamma(\alpha(1-m))}\, x^{\alpha m - 1}(1 - x)^{\alpha(1-m)-1}\, \mathrm{d}x = 0$$

[since, when α goes to zero, the mass of the beta $\mathscr{B}e(\alpha m, \alpha(1-m))$ distribution gets more and more concentrated around 0 and 1, with masses $(1 - m)$ and m, respectively] and

$$\lim_{\alpha\to\infty} \int_\ell^u \frac{\Gamma(\alpha)}{\Gamma(\alpha m)\Gamma(\alpha(1-m))}\, x^{\alpha m-1}(1-x)^{\alpha(1-m)-1}\,\mathrm{d}x = 1$$

[this is easily established using the gamma representation introduced in Exercise 4.19 and the law of large numbers]. Therefore, due to the continuity [in α] of the coverage probability, there must exist one value of α such that

$$B(\ell, u|\alpha, m) = \int_\ell^u \frac{\Gamma(\alpha)}{\Gamma(\alpha m)\Gamma(\alpha(1-m)}\, x^{\alpha m-1}(1-x)^{\alpha(1-m)-1}\,\mathrm{d}x = 0.9\,.$$

Figure 5.2 illustrates this property by plotting $B(\ell, u|\alpha, m)$ for $\ell = 0.1$, $u = 0.6$, $m = 0.4$ and α varying from 0.1 to 50.

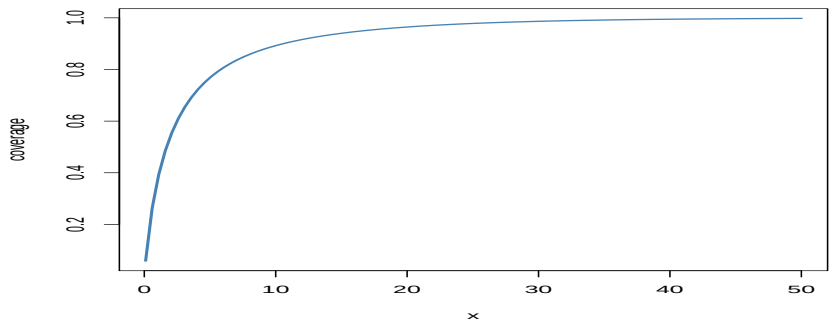

Fig. 5.2. Coverage of the interval $(\ell, u) = (0.1, 0.6)$ by a $\mathscr{B}e(0.4\alpha, 0.6\alpha)$ distribution when α varies.

Exercise 5.22 Show that groups of consecutive unknown locations are independent of one another, conditional on the observations. Devise a way to simulate these groups by blocks rather than one at a time, that is, using the joint posterior distributions of the groups rather than the full conditional distributions of the states.

As will become clearer in Chapter 7, the Arnason-Schwarz model is a very special case of [partly] hidden Markov chain: the locations $z_{(i,t)}$ of an individual i along time constitute a Markov chain that is only observed at times t when the individual is captured. Whether or not $z_{(i,t)}$ is observed has no relevance on the fact that, given $z_{(i,t)}$, $(z_{(i,t-1)}, z_{(i,t-2)}, \ldots)$ is independent from $(z_{(i,t+1)}, z_{(i,t+2)}, \ldots)$. Therefore, conditioning on any time t and on the

corresponding value of $z_{(i,t)}$ makes the past and the future locations independent. In particular, conditioning on the observed locations makes the blocks of unobserved locations in-between independent.

Those blocks could therefore be generated independently and parallely, an alternative which would then speed up the Gibbs sampler compared with the implementation in Algorithm 5.3. In addition, this would bring additional freedom in the choice of the proposals for the simulation of the different blocks and thus could further increase efficiency.

6

Mixture Models

Exercise 6.1 Show that a mixture of Bernoulli distributions is again a Bernoulli distribution. Extend this to the case of multinomial distributions.

By definition, if

$$x \sim \sum_{i=1}^{k} p_i \mathscr{B}(q_i)\,,$$

then x only takes the values 0 and 1 with probabilities

$$\sum_{i=1}^{k} p_i(1-q_i) = 1 - \sum_{i=1}^{k} p_i q_i \quad \text{and} \quad \sum_{i=1}^{k} p_i q_i\,,$$

respectively. This mixture is thus a Bernoulli distribution

$$\mathscr{B}\left(\sum_{i=1}^{k} p_i q_i\right)\,.$$

When considering a mixture of multinomial distributions,

$$x \sim \sum_{i=1}^{k} p_i \mathscr{M}_k(\mathbf{q}_i)\,,$$

with $\mathbf{q}_i = (q_{i1}, \ldots, q_{ik})$, x takes the values $1 \leq j \leq k$ with probabilities

$$\sum_{i=1}^{k} p_i q_{ij}$$

and therefore this defines a multinomial distribution. This means that a mixture of multinomial distributions cannot be identifiable unless some restrictions are set upon its parameters.

Exercise 6.2 Show that the number of nonnegative integer solutions of the decomposition of n into k parts such that $n_1 + \ldots + n_k = n$ is equal to

$$\mathfrak{r} = \binom{n+k-1}{n}.$$

Deduce that the number of partition sets is of order $\mathrm{O}(n^{k-1})$.

This is a usual combinatoric result, detailed for instance in Feller (1970). A way to show that $\mathfrak{r}$ is the solution is to use the "bottomless box" trick: consider a box with k cases and n identical balls to put into those cases. If we remove the bottom of the box, one allocation of the n balls is represented by a sequence of balls (O) and of case separations (|) or, equivalently, of 0's and 1's, of which there are n and $k-1$ respectively [since the box itself does not count, we have to remove the extreme separations]. Picking n positions out of $n+(k-1)$ is exactly $\mathfrak{r}$.

This value is thus the number of "partitions" of an n sample into k groups [we write "partitions" and not partitions because, strictly speaking, all sets of a partition are non-empty]. Since

$$\binom{n+k-1}{n} = \frac{(n+k-1)!}{n!(k-1)!} \approx \frac{n^{k-1}}{(k-1)!}\,,$$

when $n \gg k$, there is indeed an order $\mathrm{O}(n^{k-1})$ of partitions.

Exercise 6.3 For a mixture of two normal distributions with all parameters unknown,

$$p\mathcal{N}(\mu_1, \sigma_1^2) + (1-p)\mathcal{N}(\mu_2, \sigma_2^2)\,,$$

and for the prior distribution $(j = 1, 2)$

$$\mu_j|\sigma_j \sim \mathcal{N}(\xi_j, \sigma_i^2/n_j)\,, \quad \sigma_j^2 \sim \mathcal{IG}(\nu_j/2, s_j^2/2)\,, \quad p \sim \mathcal{B}e(\alpha, \beta)\,,$$

show that

$$p|\mathbf{x}, \mathbf{z} \sim \mathcal{B}e(\alpha + \ell_1, \beta + \ell_2),$$

$$\mu_j|\sigma_j, \mathbf{x}, \mathbf{z} \sim \mathcal{N}\left(\xi_1(\mathbf{z}), \frac{\sigma_j^2}{n_j + \ell_j}\right)\,, \quad \sigma_j^2|\mathbf{x}, \mathbf{z} \sim \mathcal{IG}((\nu_j + \ell_j)/2, s_j(\mathbf{z})/2)$$

where ℓ_j is the number of z_i equal to j, $\bar{x}_j(\mathbf{z})$ and $\hat{s}_j(\mathbf{z})$ are the empirical mean and variance for the subsample with z_i equal to j, and

$$\xi_j(\mathbf{z}) = \frac{n_j\xi_j + \ell_j\bar{x}_j(\mathbf{z})}{n_j + \ell_j}\,, \quad s_j(\mathbf{z}) = s_j^2 + \hat{s}_j^2(\mathbf{z}) + \frac{n_j\ell_j}{n_j + \ell_j}(\xi_j - \bar{x}_j(\mathbf{z}))^2\,.$$

Compute the corresponding weight $\omega(\mathbf{z})$.

If the latent (or missing) variable $\mathbf{z}$ is introduced, the joint distribution of $(\mathbf{x}, \mathbf{z})$ [equal to the completed likelihood] decomposes into

$$\prod_{i=1}^{n} p_{z_i} f(x_i|\theta_{z_i}) = \prod_{j=1}^{2} \prod_{i;z_i=j} p_j f(x_i|\theta_j)$$
$$\propto \prod_{j=1}^{k} p_j^{\ell_j} \prod_{i;z_i=j} \frac{e^{-(x_i-\mu_j)^2/2\sigma_j^2}}{\sigma_j}\,, \tag{6.1}$$

where $p_1 = p$ and $p_2 = (1-p)$. Therefore, using the conjugate priors proposed in the question, we have a decomposition of the posterior distribution of the parameters given $(\mathbf{x}, \mathbf{z})$ in

$$p^{\ell_1+\alpha-1}(1-p)^{\ell 2+\beta-1} \prod_{j=1}^{2} \frac{e^{-(x_i-\mu_j)^2/2\sigma_j^2}}{\sigma_j} \pi(\mu_j, \sigma_j^2)\,.$$

This implies that $p|\mathbf{x}, \mathbf{z} \sim \mathscr{B}e(\alpha + \ell_1, \beta + \ell_2)$ and that the posterior distributions of the pairs (μ_j, σ_j^2) are the posterior distributions associated with the normal observations allocated (via the z_i's) to the corresponding component. The values of the hyperparameters are therefore those already found in Chapter 2 (see, e.g., eqn. (4.6) and Exercise 2.13).

The weight $\omega(\mathbf{z})$ is the marginal [posterior] distribution of $\mathbf{z}$, since

$$\pi(\boldsymbol{\theta}, p|\mathbf{x}) = \sum_{\mathbf{z}} \omega(\mathbf{z}) \pi(\boldsymbol{\theta}, p|\mathbf{x}, \mathbf{z})\,.$$

Therefore, if $p_1 = p$ and $p_2 = 1 - p$,

$$\omega(\mathbf{z}) \propto \int \prod_{j=1}^{2} p_j^{\ell_j} \prod_{i;z_i=j} \frac{e^{-(x_i-\mu_j)^2/2\sigma_j^2}}{\sigma_j} \pi(\boldsymbol{\theta}, p)\, \mathrm{d}\boldsymbol{\theta}\mathrm{d}p$$
$$\propto \frac{\Gamma(\alpha+\ell_1)\Gamma(\beta+\ell_2)}{\Gamma(\alpha+\beta+n)}$$
$$\int \prod_{j=1}^{2} \exp\left[\frac{-1}{2\sigma_j^2}\left\{(n_j+\ell_j)(\mu_j - \xi_j(\mathbf{z}))^2 + s_j(\mathbf{z})\right\}\right] \sigma_j^{-\ell_j-\nu_j-3}\, \mathrm{d}\theta$$
$$\propto \frac{\Gamma(\alpha+\ell_1)\Gamma(\beta+\ell_2)}{\Gamma(\alpha+\beta+n)} \prod_{j=1}^{2} \frac{\Gamma((\ell_j+\nu_j)/2)(s_j(\mathbf{z})/2)^{(\nu_j+\ell_j)/2}}{\sqrt{n_j+\ell_j}}$$

and the proportionality factor can be derived by summing up the rhs over all $\mathbf{z}$'s. (There are 2^n terms in this sum.)

Exercise 6.4 For the normal mixture model of Exercise 6.3, compute the function $Q(\theta_0, \theta)$ and derive both steps of the EM algorithm. Apply this algorithm to a simulated dataset and test the influence of the starting point θ_0.

Starting from the representation (6.1) above,

$$\log \ell(\boldsymbol{\theta}, p|\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} \{\mathbb{I}_1(z_i) \log(p\, f(x_i|\theta_1) + \mathbb{I}_2(z_i) \log((1-p)\, f(x_i|\theta_2)\} \ ,$$

which implies that

$$\begin{aligned}
Q\{(\boldsymbol{\theta}^{(t)}, p^{(t)}), (\boldsymbol{\theta}, p)\} &= \mathbb{E}_{(\theta^{(t)}, p^{(t)})} \left[\log \ell(\boldsymbol{\theta}, p|\mathbf{x}, \mathbf{z})|\mathbf{x}\right] \\
&= \sum_{i=1}^{n} \left\{ \mathrm{P}_{(\theta^{(t)}, p^{(t)})} \left(z_i = 1|\mathbf{x}\right) \log(p\, f(x_i|\boldsymbol{\theta}_1) \right. \\
&\qquad \left. + \mathrm{P}_{(\boldsymbol{\theta}^{(t)}, \mathbf{p}^{(t)})} \left(z_i = 2|\mathbf{x}\right) \log((1-p)\, f(x_i|\boldsymbol{\theta}_2) \right\} \\
&= \log(p/\sigma_1) \sum_{i=1}^{n} \mathrm{P}_{(\boldsymbol{\theta}^{(t)}, p^{(t)})} \left(z_i = 1|\mathbf{x}\right) \\
&\quad + \log((1-p)/\sigma_2) \sum_{i=1}^{n} \mathrm{P}_{(\boldsymbol{\theta}^{(t)}, p^{(t)})} \left(z_i = 2|\mathbf{x}\right) \\
&\quad - \sum_{i=1}^{n} \mathrm{P}_{(\boldsymbol{\theta}^{(t)}, p^{(t)})} \left(z_i = 1|\mathbf{x}\right) \frac{(x_i - \mu_1)^2}{2\sigma_1^2} \\
&\quad - \sum_{i=1}^{n} \mathrm{P}_{(\boldsymbol{\theta}^{(t)}, p^{(t)})} \left(z_i = 2|\mathbf{x}\right) \frac{(x_i - \mu_2)^2}{2\sigma_2^2} \, .
\end{aligned}$$

If we maximise this function in p, we get that

$$\begin{aligned}
p^{(t+1)} &= \frac{1}{n} \sum_{i=1}^{n} \mathrm{P}_{(\boldsymbol{\theta}^{(t)}, p^{(t)})} \left(z_i = 1|\mathbf{x}\right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{p^{(t)} f(x_i|\boldsymbol{\theta}_1^{(t)})}{p^{(t)} f(x_i|\boldsymbol{\theta}_1^{(t)}) + (1-p^{(t)}) f(x_i|\boldsymbol{\theta}_2^{(t)})}
\end{aligned}$$

while maximising in (μ_j, σ_j) $(j = 1, 2)$ leads to

$$\begin{aligned}
\mu_j^{(t+1)} &= \sum_{i=1}^{n} \mathrm{P}_{(\boldsymbol{\theta}^{(t)}, p^{(t)})} \left(z_i = j|\mathbf{x}\right) x_i \Big/ \sum_{i=1}^{n} \mathrm{P}_{(\boldsymbol{\theta}^{(t)}, p^{(t)})} \left(z_i = j|\mathbf{x}\right) \\
&= \frac{1}{n p_j^{(t+1)}} \sum_{i=1}^{n} \frac{x_i p_j^{(t)} f(x_i|\boldsymbol{\theta}_j^{(t)})}{p^{(t)} f(x_i|\boldsymbol{\theta}_1^{(t)}) + (1-p^{(t)}) f(x_i|\boldsymbol{\theta}_2^{(t)})} \, , \\
\sigma_j^{2(t+1)} &= \sum_{i=1}^{n} \mathrm{P}_{(\boldsymbol{\theta}^{(t)}, p^{(t)})} \left(z_i = j|\mathbf{x}\right) (x_i - \mu_j^{(t+1)})^2 \Big/ \sum_{i=1}^{n} \mathrm{P}_{(\boldsymbol{\theta}^{(t)}, p^{(t)})} \left(z_i = j|\mathbf{x}\right) \\
&= \frac{1}{n p_j^{(t+1)}} \sum_{i=1}^{n} \frac{\left[x_i - \mu_j^{(t+1)}\right]^2 p_j^{(t)} f(x_i|\boldsymbol{\theta}_j^{(t)})}{p^{(t)} f(x_i|\boldsymbol{\theta}_1^{(t)}) + (1-p^{(t)}) f(x_i|\boldsymbol{\theta}_2^{(t)})} \, ,
\end{aligned}$$

where $p_1^{(t)} = p^{(t)}$ and $p_2^{(t)} = (1 - p^{(t)})$.

A possible implementation of this algorithm in R is given below:

```
# simulation of the dataset
n=324
tz=sample(1:2,n,prob=c(.4,.6),rep=T)
tt=c(0,3.5)
ts=sqrt(c(1.1,0.8))
x=rnorm(n,mean=tt[tz],sd=ts[tz])

para=matrix(0,ncol=50,nrow=5)
likem=rep(0,50)

# initial values chosen at random
para[,1]=c(runif(1),mean(x)+2*rnorm(2)*sd(x),rexp(2)*var(x))
likem[1]=sum(log( para[1,1]*dnorm(x,mean=para[2,1],
  sd=sqrt(para[4,1]))+(1-para[1,1])*dnorm(x,mean=para[3,1],
  sd=sqrt(para[5,1])) ))

# 50 EM steps
for (em in 2:50){

   # E step
   postprob=1/( 1+(1-para[1,em-1])*dnorm(x,mean=para[3,em-1],
     sd=sqrt(para[5,em-1]))/( para[1,em-1]*dnorm(x,
     mean=para[2,em-1],sd=sqrt(para[4,em-1]))) )

   # M step
   para[1,em]=mean(postprob)
   para[2,em]=mean(x*postprob)/para[1,em]
   para[3,em]=mean(x*(1-postprob))/(1-para[1,em])
   para[4,em]=mean((x-para[2,em])^2*postprob)/para[1,em]
   para[5,em]=mean((x-para[3,em])^2*(1-postprob))/(1-para[1,em])

   # value of the likelihood
   likem[em]=sum(log(para[1,em]*dnorm(x,mean=para[2,em],
     sd=sqrt(para[4,em]))+(1-para[1,em])*dnorm(x,mean=para[3,em],
     sd=sqrt(para[5,em])) ))
}
```

Figure 6.1 represents the increase in the log-likelihoods along EM iterations for 20 different starting points [and the same dataset x]. While most starting points lead to the same value of the log-likelihood after 50 iterations, one starting point induces a different convergence behaviour.

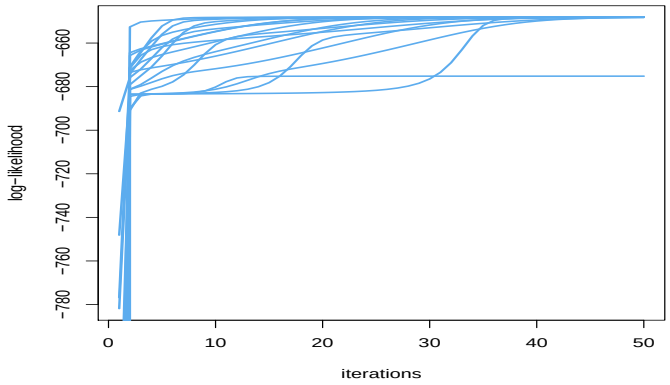

Fig. 6.1. Increase of the log-likelihood along EM iterations for 20 different starting points.

Exercise 6.5 Show that the θ_j's in model (6.2) are dependent on each other given (only) $\mathbf{x}$.

The likelihood associated with model (6.2) being

$$\ell(\boldsymbol{\theta}, p|\mathbf{x}) = \prod_{i=1}^{n} \left[\sum_{j=1}^{k} p_j \, f(x_i|\boldsymbol{\theta}_j) \right] ,$$

it is clear that the posterior distribution will not factorise as a product of functions of the different parameters. It is only given $(\mathbf{x}, \mathbf{z})$ that the $\boldsymbol{\theta}_j$'s are independent.

Exercise 6.6 Construct and test the Gibbs sampler associated with the (ξ, μ_0) parameterization of (6.3) when $\mu_1 = \mu_0 - \xi$ and $\mu_2 = \mu_0 + \xi$.

The simulation of the z_i's is unchanged [since it does not depend on the parameterisation of the components. The conditional distribution of (ξ, μ_0) given $(\mathbf{x}, \mathbf{z})$ is

$$\pi(\xi, \mu_0|\mathbf{x}, \mathbf{z}) \propto \exp \frac{-1}{2} \left\{ \sum_{z_i=1} (x_i - \mu_0 + \xi)^2 + \sum_{z_i=2} (x_i - \mu_0 - \xi)^2 \right\} .$$

Therefore, ξ and μ_0 are not independent given $(\mathbf{x}, \mathbf{z})$, with

$$\mu_0|\xi, \mathbf{x}, \mathbf{z} \sim \mathcal{N}\left(\frac{n\overline{x} + (\ell_1 - \ell_2)\xi}{n}, \frac{1}{n}\right),$$
$$\xi|\mu_0, \mathbf{x}, \mathbf{z} \sim \mathcal{N}\left(\frac{\sum_{z_i=2}(x_i - \mu_0) - \sum_{z_i=1}(x_i - \mu_0)}{n}, \frac{1}{n}\right)$$

The implementation of this Gibbs sampler is therefore a simple modification of the code given in `#6.R` on the webpage: the MCMC loop is now

```
for (t in 2:Nsim){

  # allocation
  fact=.3*sqrt(exp(gu1^2-gu2^2))/.7
  probs=1/(1+fact*exp(sampl*(gu2-gu1)))
  zeds=(runif(N)<probs)

  # Gibbs sampling
  mu0=rnorm(1)/sqrt(N)+(sum(sampl)+xi*(sum(zeds==1)
    -sum(zeds==0)))/N
  xi=rnorm(1)/sqrt(N)+(sum(sampl[zeds==0]-mu0)
    -sum(sampl[zeds==1]-mu0))/N

  # reparameterisation
  gu1=mu0-xi
  gu2=mu0+xi
  muz[t,]=(c(gu1,gu2))

}
```

If we run repeatedly this algorithm, the Markov chain produced is highly dependent on the starting value and remains captive of local modes, as illustrated on Figure 6.2. This reparameterisation thus seems less robust than the original parameterisation.

Exercise 6.7 Give the ratio corresponding to (6.7) when the parameter of interest is in $[0, 1]$ and the random walk proposal is on the logit transform $\log \theta/(1-\theta)$.

Since

$$\frac{\partial}{\partial\theta} \log\left[\theta/(1-\theta)\right] = \frac{1}{\theta} + \frac{1}{1-\theta)} = \frac{1}{\theta(1-\theta)},$$

the Metropolis–Hastings acceptance ratio for the logit transformed random walk is

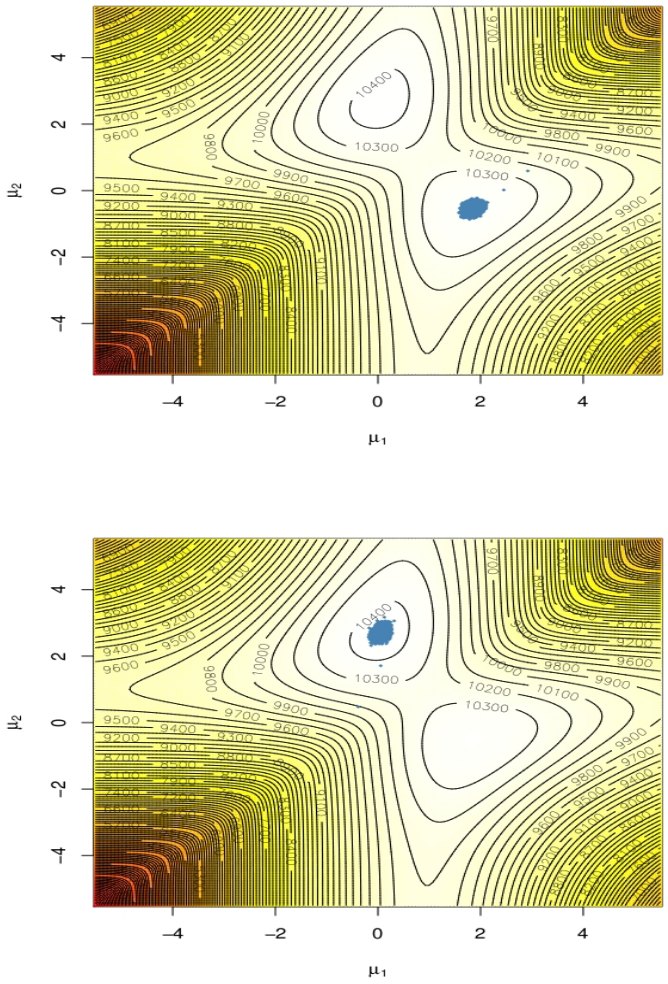

Fig. 6.2. Influence of the starting value on the convergence of the Gibbs sampler associated with the location parameterisation of the mean mixture (10,000 iterations).

$$\frac{\pi(\widetilde{\theta}_j)}{\pi(\theta_j^{(t-1)})}\,\frac{\widetilde{\theta}_j(1-\widetilde{\theta}_j)}{\theta_j^{(t-1)}(1-\theta_j^{(t-1)})}\wedge 1\,.$$

Exercise 6.8 Show that, if an exchangeable prior π is used on the vector of weights $(p_1,\ldots,p_k)$, then, necessarily, $\mathbb{E}^\pi[p_j]=1/k$ and, if the prior on the other parameters $(\theta_1,\ldots,\theta_k)$ is also exchangeable, then $\mathbb{E}^\pi[p_j|x_1,\ldots,x_n]=1/k$ for all j's.

If

$$\pi(p_1,\ldots,p_k)=\pi(p_{\sigma(1)},\ldots,p_{\sigma(k)})$$

for any permutation $\sigma\in\mathfrak{S}_k$, then

$$\mathbb{E}^\pi[p_j]=\int p_j\pi(p_1,\ldots,p_j,\ldots,p_k)\,\mathrm{d}\mathbf{p}=\int p_j\pi(p_j,\ldots,p_1,\ldots,p_k)\,\mathrm{d}\mathbf{p}=\mathbb{E}^\pi[p_1]\,.$$

Given that $\sum_{j=1}^k p_j=1$, this implies $\mathbb{E}^\pi[p_j]=1/k$.

When both the likelihood and the prior are exchangeable in (p_j,θ_j), the same result applies to the posterior distribution.

Exercise 6.9 Show that running an MCMC algorithm with target $\pi(\theta|\mathbf{x})^\gamma$ will increase the proximity to the MAP estimate when $\gamma>1$ is large. Discuss the modifications required in Algorithm 6.2.to achieve simulation from $\pi(\theta|\mathbf{x})^\gamma$ when $\gamma\in\mathbb{N}^*$.

The power distribution $\pi_\gamma(\theta)\propto\pi(\theta)^\gamma$ shares the same modes as π, but the global mode gets more and more mass as γ increases. If $\theta^\star$ is the global mode of π [and of π_γ], then $\{\pi(\theta)/\pi(\theta^\star)\}^\gamma$ goes to 0 as γ goes to ∞ for all θ's different from $\theta^\star$. Moreover, for any $0<\alpha<1$, if we define the α neighbourhood $\mathfrak{N}_\alpha$ of $\theta^\star$ as the set of θ's such that $\pi(\theta)\geq\alpha\pi(\theta^\star)$, then $\pi_\gamma(\mathfrak{N}_\alpha)$ converges to 1 as γ goes to ∞.

The idea behind *simulated annealing* is that, first, the distribution $\pi_\gamma(\theta)\propto\pi(\theta)^\gamma$ is more concentrated around its main mode than $\pi(\theta)$ if γ is large and, second, that it is not necessary to simulate a whole sample from $\pi(\theta)$, then a whole sample from $\pi(\theta)^2$ and so on to achieve a convergent approximation of the MAP estimate. Increasing γ slowly enough along iterations leads to the same result with a much smaller computing requirement.

When considering the application of this idea to a mean mixture as (6.3) [in the book], the modification of Algorithm 6.2 is rather immediate: since we need to simulate from $\pi(\boldsymbol{\theta},p|\mathbf{x})^\gamma$ [up to a normalising constant], this is equivalent to simulate from $\ell(\boldsymbol{\theta},p|\mathbf{x})^\gamma\times\pi(\boldsymbol{\theta},p)^\gamma$. This means that, since the

prior is [normal] conjugate, the prior hyperparameter λ is modified into $\gamma\lambda$ and that the likelihood is to be completed γ times rather than once, i.e.

$$\ell(\boldsymbol{\theta}, p|\mathbf{x})^\gamma = \left(\int f(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}, p)\, \mathrm{d}\mathbf{z}\right)^\gamma = \prod_{j=1}^{\gamma} \int f(\mathbf{x}, \mathbf{z}_j|\boldsymbol{\theta}, p)\, \mathrm{d}\mathbf{z}_j\,.$$

Using this duplication trick, the annealed version of Algorithm 6.2 writes as

Algorithm **Annealed Mean Mixture Gibbs Sampler**

Initialization. Choose $\mu_1^{(0)}$ and $\mu_2^{(0)}$,

Iteration t ($t \geq 1$).

1. For $i = 1, \ldots, n$, $j = 1, \ldots, \gamma$, generate $z_{ij}^{(t)}$ from

$$\mathbb{P}\left(z_{ij} = 1\right) \propto p \exp\left\{-\frac{1}{2}\left(x_i - \mu_1^{(t-1)}\right)^2\right\}$$

$$\mathbb{P}\left(z_{ij} = 2\right) \propto (1-p) \exp\left\{-\frac{1}{2}\left(x_i - \mu_2^{(t-1)}\right)^2\right\}$$

2. Compute

$$\ell = \sum_{j=1}^{\gamma}\sum_{i=1}^{n} \mathbb{I}_{z_{ij}^{(t)}=1} \quad \text{and} \quad \bar{x}_u\left(\mathbf{z}\right) = \sum_{j=1}^{\gamma}\sum_{i=1}^{n} \mathbb{I}_{z_{ij}^{(t)}=u} x_i$$

3. Generate $\mu_1^{(t)}$ from $\mathscr{N}\left(\dfrac{\gamma\lambda\delta + bar x_1\left(\mathbf{z}\right)}{\gamma\lambda + \ell}, \dfrac{1}{\gamma\lambda + \ell}\right)$

4. Generate $\mu_2^{(t)}$ from $\mathscr{N}\left(\dfrac{\gamma\lambda\delta + \bar{x}_2\left(\mathbf{z}\right)}{\gamma\lambda + \gamma n - \ell}, \dfrac{1}{\gamma\lambda + \gamma n - \ell}\right)$.

This additional level of completion means that the Markov chain will have difficulties to move around, compared with the original Gibbs sampling algorithm. While closer visits to the global mode are guaranteed in theory, they may require many more simulations in practice.

Exercise 6.10 In the setting of the mean mixture (6.3), run an MCMC simulation experiment to compare the influence of a $\mathscr{N}(0, 100)$ and of a $\mathscr{N}(0, 10000)$ prior on (μ_1, μ_2) on a sample of 500 observations.

This is straightforward in that the code in `#6.R` simply needs to be modified from

```
# Gibbs samplin
  gu1=rnorm(1)/sqrt(.1+length(zeds[zeds==1]))+
   (sum(sampl[zeds==1]))/(.1+length(zeds[zeds==1]))
  gu2=rnorm(1)/sqrt(.1+length(zeds[zeds==0]))+
   (sum(sampl[zeds==0]))/(.1+length(zeds[zeds==0]))
```

to

```
# Gibbs samplin
  gu1=rnorm(1)/sqrt(.01+length(zeds[zeds==1]))+
    (sum(sampl[zeds==1]))/(.01+length(zeds[zeds==1]))
  gu2=rnorm(1)/sqrt(.01+length(zeds[zeds==0]))+
    (sum(sampl[zeds==0]))/(.01+length(zeds[zeds==0]))
```

for the $\mathcal{N}(0, 100)$ prior and to

```
# Gibbs samplin
  gu1=rnorm(1)/sqrt(.0001+length(zeds[zeds==1]))+
    (sum(sampl[zeds==1]))/(.0001+length(zeds[zeds==1]))
  gu2=rnorm(1)/sqrt(.0001+length(zeds[zeds==0]))+
    (sum(sampl[zeds==0]))/(.0001+length(zeds[zeds==0]))
```

for the $\mathcal{N}(0, 10^4)$ prior. While we do not reproduce the results here, it appears that the sampler associated with the $\mathcal{N}(0, 10^4)$ prior has a higher probability to escape the dubious mode.

Exercise 6.11 Show that, for a normal mixture $0.5\,\mathcal{N}(0, 1) + 0.5\,\mathcal{N}(\mu, \sigma^2)$, the likelihood is unbounded. Exhibit this feature by plotting the likelihood of a simulated sample, using the R image procedure.

This follows from the decomposition of the likelihood

$$\ell(\boldsymbol{\theta}|\mathbf{x}) = \prod_{i=1}^{n} \left[\sum_{j=1}^{2} 0.5\, f(x_i|\boldsymbol{\theta}_j) \right] ,$$

into a sum [over all partitions] of the terms

$$\prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}_{z_i}) = \prod_{i;z_i=1} \varphi(x_i) \prod_{i;z_i=2} \frac{\varphi\{(x_i-\mu)/\sigma\}}{\sigma} .$$

In exactly n of those 2^n partitions, a single observation is allocated to the second component, i.e. there is a single i such that $z_i = 2$. For those particular partitions, if we choose $\mu = x_i$, the second product reduces to $1/\sigma$ which is not bounded when σ goes to 0. Since the observed likelihood is the sume of all those terms, it is bounded from below by terms that are unbounded and therefore it is unbounded.

An R code illustrating this behaviour is

```
# Sample construction
N=100
sampl=rnorm(N)+(runif(N)<.3)*2.7

# Grid
mu=seq(-2.5,5.5,length=250)
sig=rev(1/seq(.001,.01,length=250))  # inverse variance
mo1=mu%*%t(rep(1,length=length(sig)))
mo2=(rep(1,length=length(mu)))%*%t(sig)
ca1=-0.5*mo1^2*mo2
ca2=mo1*mo2
ca3=sqrt(mo2)
ca4=0.5*(1-mo2)

# Likelihood surface
like=0*mo1
for (i in 1:N)
  like=like+log(1+exp(ca1+sampl[i]*ca2+sampl[i]^2*ca4)*ca3)
like=like-min(like)

sig=rev(1/sig)
image(mu,sig,like,xlab=expression(mu),
  ylab=expression(sigma^2),col=heat.colors(250))
contour(mu,sig,like,add=T,nlevels=50)
```

and Figure 6.3 exhibits the characteristic stripes of an explosive likelihood as σ approaches 0 for values of μ close to the values of the sample.

Exercise 6.12 Show that the ratio (6.8) goes to 1 when α goes to 0 when the proposal q is a random walk. Describe the average behavior of this ratio in the case of an independent proposal.

This is obvious since, when the proposal is a random walk [without reparameterisation], the ratio $q(\boldsymbol{\theta},\mathbf{p}|\boldsymbol{\theta}',\mathbf{p}')/q(\boldsymbol{\theta}',\mathbf{p}'|\boldsymbol{\theta},\mathbf{p})$ is equal to 1. The Metropolis–Hastings ratio thus reduces to the ratio of the targets to the power α, which [a.e.] converges to 1 as α goes to 0.

In the case of an independent proposal,

$$\left(\frac{\pi(\boldsymbol{\theta}',\mathbf{p}'|\mathbf{x})}{\pi(\boldsymbol{\theta},\mathbf{p}|\mathbf{x})}\right)^{\alpha} \frac{q(\boldsymbol{\theta},\mathbf{p})}{q(\boldsymbol{\theta}',\mathbf{p}')} \wedge 1$$

is equivalent to $q(\boldsymbol{\theta},\mathbf{p})/q(\boldsymbol{\theta}',\mathbf{p}')$ and therefore does not converge to 1. This situation can however be avoided by picking q^{α} rather than q, in which case the ratio once more converges to 1 as α goes to 0.

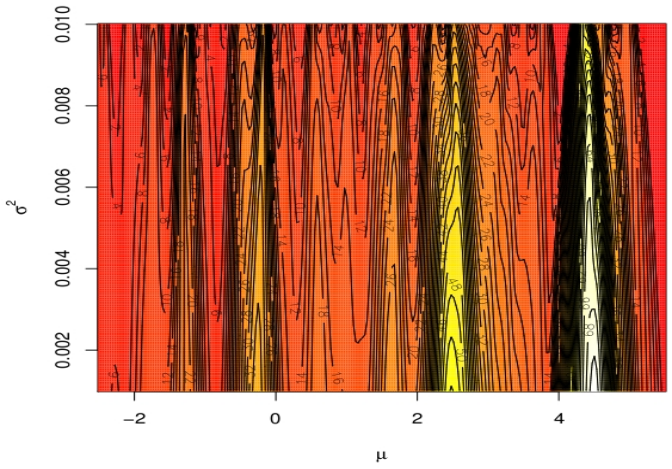

Fig. 6.3. Illustration of an unbounded mixture likelihood.

Exercise 6.13 If one needs to use importance sampling weights, show that the simultaneous choice of several powers α requires the computation of the normalizing constant of π_α.

If samples $(\theta_{i\alpha})_i$ from several tempered versions π_α of π are to be used simultaneously, the importance weights associated with those samples $\pi(\theta_{i\alpha})/\pi_\alpha(\theta_{i\alpha})$ require the computation of the normalizing constants, which is most often impossible. This difficulty explains the appeal of the "pumping mechanism" of Algorithm 6.5, which cancels the need for normalizing constants by using the same π_α twice, once in the numerator and once in the denominator (see Exercice 6.14).

Exercise 6.14 Check that Algorithm 6.5 does not require the normalizing constants of the π_{α_i}'s and show that π is the corresponding stationary distribution.

Since the acceptance probability

$$\min\left\{1, \frac{\pi_{\alpha_1}(x^{(t)})}{\pi(x^{(t)})} \cdots \frac{\pi_{\alpha_p}(x_{p-1}^{(t)})}{\pi_{\alpha_{p-1}}(x_{p-1}^{(t)})} \frac{\pi_{\alpha_{p-1}}(x_p^{(t)})}{\pi_{\alpha_p}(x_p^{(t)})} \cdots \frac{\pi(x_{2p-1}^{(t)})}{\pi_{\alpha_1}(x_{2p-1}^{(t)})}\right\}$$

uses twice each power α_j of π, the unknown normalizing constants of the π_{α_i}'s vanish, which is one of the main reasons for using this algorithm.

The fact that π is stationary can be derived from the "detailed balance" condition: let us assume that each of the MCMC kernels satisfy the corresponding detailed balance equation

$$\pi_{\alpha_j}(x_0)\mathrm{MCMC}(x_1|x_0,\pi_{\alpha_j}) = \pi_{\alpha_j}(x_1)\mathrm{MCMC}(x_0|x_1,\pi_{\alpha_j})\,.$$

Then

$$\begin{aligned}
&\pi(x^{(t)})\mathrm{MCMC}(x_1^{(t)}|x^{(t)},\pi_{\alpha_1})\cdots\mathrm{MCMC}(x_{2p}^{(t)}|x_{2p-1}^{(t)},\pi_{\alpha_1})\\
&\quad\times\min\left\{1,\frac{\pi_{\alpha_1}(x^{(t)})}{\pi(x^{(t)})}\cdots\frac{\pi_{\alpha_p}(x_{p-1}^{(t)})}{\pi_{\alpha_{p-1}}(x_{p-1}^{(t)})}\frac{\pi_{\alpha_{p-1}}(x_p^{(t)})}{\pi_{\alpha_p}(x_p^{(t)})}\cdots\frac{\pi(x_{2p-1}^{(t)})}{\pi_{\alpha_1}(x_{2p-1}^{(t)})}\right\}\\
&\quad=\pi_{\alpha_1}(x^{(t)})\mathrm{MCMC}(x_1^{(t)}|x^{(t)},\pi_{\alpha_1})\cdots\mathrm{MCMC}(x_{2p}^{(t)}|x_{2p-1}^{(t)},\pi_{\alpha_1})\\
&\quad\times\min\left\{\frac{\pi(x^{(t)})}{\pi_{\alpha_1}(x^{(t)})},\frac{\pi_{\alpha_2}(x_1^{(t)})}{\pi_{\alpha_1}(x_1^{(t)})}\cdots\frac{\pi(x_{2p-1}^{(t)})}{\pi_{\alpha_1}(x_{2p-1}^{(t)})}\right\}\\
&\quad=\mathrm{MCMC}(x^{(t)}|x_1^{(t)},\pi_{\alpha_1})\pi_{\alpha_1}(x_1^{(t)})\mathrm{MCMC}(x_2^{(t)}|x_1^{(t)},\pi_{\alpha_2})\cdots\\
&\quad\times\mathrm{MCMC}(x_{2p}^{(t)}|x_{2p-1}^{(t)},\pi_{\alpha_1})\min\left\{\frac{\pi(x^{(t)})}{\pi_{\alpha_1}(x^{(t)})},\frac{\pi_{\alpha_2}(x_1^{(t)})}{\pi_{\alpha_1}(x_1^{(t)})}\cdots\frac{\pi(x_{2p-1}^{(t)})}{\pi_{\alpha_1}(x_{2p-1}^{(t)})}\right\}\\
&\quad=\mathrm{MCMC}(x^{(t)}|x_1^{(t)},\pi_{\alpha_1})\mathrm{MCMC}(x_1^{(t)}|x_2^{(t)},\pi_{\alpha_2})\pi_{\alpha_2}(x_2^{(t)})\cdots\\
&\quad\times\mathrm{MCMC}(x_{2p}^{(t)}|x_{2p-1}^{(t)},\pi_{\alpha_1})\min\left\{\frac{\pi(x^{(t)})\pi_{\alpha_1}(x_1^{(t)})}{\pi_{\alpha_1}(x^{(t)})\pi_{\alpha_2}(x_1^{(t)})},\frac{\pi_{\alpha_3}(x_2^{(t)})}{\pi_{\alpha_2}(x_2^{(t)})}\cdots\right\}\\
&\quad=\cdots\\
&\quad=\mathrm{MCMC}(x^{(t)}|x_1^{(t)},\pi_{\alpha_1})\cdots\mathrm{MCMC}(x_{2p-1}^{(t)}|x_{2p}^{(t)},\pi_{\alpha_1})\pi(x_{2p}^{(t)})\\
&\quad\times\min\left\{\frac{\pi(x^{(t)})}{\pi_{\alpha_1}(x^{(t)})}\cdots\frac{\pi_{\alpha_1}(x_{2p}^{(t)})}{\pi(x_{2p}^{(t)})},1\right\}
\end{aligned}$$

by a "domino effect" resulting from the individual detailed balance conditions. This generalised detailed balance condition then ensures that π is the stationary distribution of the chain $(x^{(t)})_t$.

Exercise 6.15 Show that the decomposition (6.9) is correct by representing the generic parameter θ as (k,θ_k) and by introducing the submodel marginals, $m_k(x_1,\ldots,x_n) = \int f_k(x_1,\ldots,x_n|\theta_k)\pi_k(\theta_k)\,\mathrm{d}\theta_k$.

In a variable dimension model, the sampling distribution is

$$f(x_1,\ldots,x_n|\theta) = f(x_1,\ldots,x_n|(k,\theta_k)) = f_k(x_1,\ldots,x_n|\theta_k).$$

Decomposing the prior distribution as

$$\pi(\theta) = \pi((k,\theta_k)) = \mathrm{P}(\mathfrak{M}_k)\pi_k(\theta_k)\,,$$

the joint distribution of $(x_1,\ldots,x_n)$ and of θ is

$$f(x_1,\ldots,x_n|\theta)\pi(\theta) = \mathrm{P}(\mathfrak{M}_k)\pi_k(\theta_k)f_k(x_1,\ldots,x_n|\theta_k)$$

and the marginal distribution of $(x_1,\ldots,x_n)$ is therefore derived by

$$\begin{aligned} m(x_1,\ldots,x_n) &= \sum_k \int \mathrm{P}(\mathfrak{M}_k)\pi_k(\theta_k)f_k(x_1,\ldots,x_n|\theta_k)\mathrm{d}\theta_k \\ &= \sum_k \mathrm{P}(\mathfrak{M}_k)m_k(x_1,\ldots,x_n)\,. \end{aligned}$$

Similarly, the predictive distribution $f(x|x_1,\ldots,x_n)$ can be expressed as

$$\begin{aligned} f(x|x_1,\ldots,x_n) &= \int f(x|\theta)\pi(\theta|x_1,\ldots,x_n)\,\mathrm{d}\theta \\ &= \sum_k \int f_k(x|\theta_k)\frac{\mathrm{P}(\mathfrak{M}_k)\pi_k(\theta_k)}{m(x_1,\ldots,x_n)}\,\mathrm{d}\theta_k \\ &= \sum_k \frac{\mathrm{P}(\mathfrak{M}_k)m_k(x_1,\ldots,x_n)}{m(x_1,\ldots,x_n)} \int f_k(x|\theta_k)\frac{\pi_k(\theta_k)}{m_k(x_1,\ldots,x_n)}\,\mathrm{d}\theta_k \\ &= \sum_k \mathrm{P}(\mathfrak{M}_k|x_1,\ldots,x_n)\int f_k(x|\theta_k)\pi_k(\theta_k|x_1,\ldots,x_n)\,\mathrm{d}\theta_k\,. \end{aligned}$$

Therefore,

$$\mathbb{E}[x|x_1,\ldots,x_n] = \sum_k \mathrm{P}(\mathfrak{M}_k|x_1,\ldots,x_n)\iint f_k(x|\theta_k)\pi_k(\theta_k|x_1,\ldots,x_n)\,\mathrm{d}\theta_k\,.$$

Exercise 6.16 For a finite collection of submodels $\mathfrak{M}_k$ $(k = 1,\ldots,K)$, with respective priors $\pi_k(\theta_k)$ and weights ϱ_k, write a generic importance sampling algorithm that approximates the posterior distribution.

The formal definition of an importance sampler in this setting is straightforward: all that is needed is a probability distribution $(\omega_1,\ldots,\omega_K)$ and a collection of importance distributions η_k with supports at least as large as $\mathrm{supp}(\pi_k)$. The corresponding importance algorithm is then made of the three following steps:

Algorithm **Importance Sampled Model Choice**

1. Generate $k \sim (\omega_1, \ldots, \omega_K)$;
2. Generate $\theta_k \sim \eta_k(\theta_k)$;
3. Compute the importance weight $\varrho_k \pi_k(\theta_k)/\omega_k \eta_k(\theta_k)$.

Obviously, the difficulty in practice is to come up with weights ω_k that are not too different from the ϱ_k's [for efficiency reasons] while selecting pertinent importance distributions η_k. This is most often impossible, hence the call to reversible jump techniques that are more local and thus require less information about the target.

Exercise 6.17 Show that, if we define the acceptance probability

$$\varrho = \frac{\pi_2(x')}{\pi_1(x)} \frac{q(x|x')}{q(x'|x)} \wedge 1$$

for moving from x to x' and

$$\varrho' = \frac{\pi_1(x)}{\pi_2(x')} \frac{q(x'|x)}{q(x|x')} \wedge 1$$

for the reverse move, the detailed balance condition is modified in such a way that, if $X_t \sim \pi_1(x)$ and if a proposal is made based on $q(x|x_t)$, X_{t+1} is distributed from $\pi_2(x)$. Relate this property to Algorithm 6.5 and its acceptance probability.

If K denotes the associated Markov kernel, we have that

$$\begin{aligned}
\pi_1(x)K(x,x') &= \pi_1(x)\left\{q(x'|x)\varrho(x,x') + \delta_x(x')\int q(z|x)[1-\varrho(x,z)]\mathrm{d}z\right\} \\
&= \min\left\{\pi_1(x)q(x'|x), \pi_1(x')q(x|x')\right\} \\
&\quad + \delta_x(x')\int \max\left\{0, q(z|x)\pi_1(x) - q(x|z)\pi_2(x)\right\}\mathrm{d}z \\
&= \pi_2(x)\widetilde{K}(x',x)
\end{aligned}$$

under the assumption that the reverse acceptance probability for the reverse move is as proposed [a rather delicate assumption that makes the whole exercise definitely less than rigorous].

In Algorithm 6.5, the derivation is perfectly sound because the kernels are used twice, once forward and once backward.

Exercise 6.18 Show that the marginal distribution of p_1 when $(p_1, \ldots, p_k) \sim \mathscr{D}_k(a_1, \ldots, a_k)$ is a $\mathscr{B}e(a_1, a_2 + \ldots + a_k)$ distribution.

This result relies on the same representation as Exercise 4.19: if $(p_1, \ldots, p_k) \sim \mathscr{D}_k(a_1, \ldots, a_k)$, p_1 is distributed identically to

$$\frac{\xi_1}{\xi_1 + \ldots + \xi_k}\,, \quad \xi_j \sim \mathscr{G}a(a_j)\,.$$

Since $\xi_2 + \ldots + \xi_k \sim \mathscr{G}a(a_2 + \cdots + a_k)$, this truly corresponds to a $\mathscr{B}e(a_1, a_2 + \ldots + a_k)$ distribution.

7

Dynamic Models

Exercise 7.1 Consider the process $(x_t)_{t\in\mathbb{Z}}$ defined by

$$x_t = a + bt + y_t\,,$$

where $(y_t)_{t\in\mathbb{Z}}$ is an iid sequence of random variables with mean 0 and variance σ^2, and where a and b are constants. Define

$$w_t = (2q+1)^{-1}\sum_{j=-q}^{q} x_{t+j}\,.$$

Compute the mean and the autocovariance function of $(w_t)_{t\in\mathbb{Z}}$. Show that $(w_t)_{t\in\mathbb{Z}}$ is not stationary but that its autocovariance function $\gamma_w(t+h,t)$ does not depend on t.

We have

$$\begin{aligned}\mathbb{E}[w_t] &= \mathbb{E}\left[(2q+1)^{-1}\sum_{j=-q}^{q} x_{t+j}\right]\\ &= (2q+1)^{-1}\sum_{j=-q}^{q}\mathbb{E}\left[a+b(t+j)+y_t\right]\\ &= a+bt\,.\end{aligned}$$

The process $(w_t)_{t\in\mathbb{Z}}$ is therefore not stationary. Moreover

$$\begin{aligned}\mathbb{E}[w_t w_{t+h}] &= \mathbb{E}\left[\left(a+bt+\frac{1}{2q+1}\sum_{j=-q}^{q} y_{t+j}\right)\left(a+bt+bh+\sum_{j=-q}^{q} y_{t+h+j}\right)\right] \\ &= (a+bt)(a+bt+bh)+\mathbb{E}\left[\sum_{j=-q}^{q} y_{t+j}\sum_{j=-q}^{q} y_{t+h+j}\right] \\ &= (a+bt)(a+bt+bh)+\mathbb{I}_{|h|\le q}(q+1-|h|)\sigma^2\,.\end{aligned}$$

Then,

$$\operatorname{cov}(w_t, w_{t+h}) = \mathbb{I}_{|h|\le q}(q+1-|h|)\sigma^2$$

and,

$$\gamma_w(t+h,t) = \mathbb{I}_{|h|\le q}(q+1-|h|)\sigma^2\,.$$

Exercise 7.2 Suppose that the process $(x_t)_{t\in\mathbb{N}}$ is such that $x_0 \sim \mathcal{N}(0,\tau^2)$ and, for all $t \in \mathbb{N}$,

$$x_{t+1}|\mathbf{x}_{0:t} \sim \mathcal{N}(x_t/2, \sigma^2)\,, \qquad \sigma > 0\,.$$

Give a necessary condition on τ^2 for $(x_t)_{t\in\mathbb{N}}$ to be a (strictly) stationary process.

We have

$$\mathbb{E}[x_1] = \mathbb{E}[\mathbb{E}[x_1|x_0]] = \mathbb{E}[x_0/2] = 0\,.$$

Moreover,

$$\mathbb{V}(x_1) = \mathbb{V}(\mathbb{E}[x_1|x_0]) + \mathbb{E}[\mathbb{V}(x_1|x_0)] = \tau^2/4 + \sigma^2\,.$$

Marginaly, x_1 is then distributed as a $\mathcal{N}(0, \tau^2/4+\sigma^2)$ variable, with the same distribution as x_0 only if $\tau^2/4 + \sigma^2 = \tau^2$, i.e. if $\tau^2 = 4\sigma^2/3$.

Exercise 7.3 Suppose that $(x_t)_{t\in\mathbb{N}}$ is a *Gaussian random walk* on $\mathbb{R}$: $x_0 \sim \mathcal{N}(0,\tau^2)$ and, for all $t \in \mathbb{N}$,

$$x_{t+1}|\mathbf{x}_{0:t} \sim \mathcal{N}(x_t, \sigma^2)\,, \qquad \sigma > 0\,.$$

Show that, whatever the value of τ^2 is, $(x_t)_{t\in\mathbb{N}}$ is not a (strictly) stationary process.

We have

$$\mathbb{E}[x_1] = \mathbb{E}[\mathbb{E}[x_1|x_0]] = \mathbb{E}[x_0] = 0\,.$$

Moreover,

$$\mathbb{V}(x_1) = \mathbb{V}(\mathbb{E}[x_1|x_0]) + \mathbb{E}[\mathbb{V}(x_1|x_0)] = \tau^2 + \sigma^2\,.$$

The marginal distribution of x_1 is then a $\mathcal{N}(0, \tau^2 + \sigma^2)$ distribution which cannot be equal to a $\mathcal{N}(0, \tau^2)$ distribution.

Exercise 7.4 Consider the process $(x_t)_{t\in\mathbb{N}}$ such that $x_0 = 0$ and, for all $t \in \mathbb{N}$,

$$x_{t+1}|\mathbf{x}_{0:t} \sim \mathcal{N}(\varrho\, x_t, \sigma^2)\,.$$

Suppose that $\pi(\varrho, \sigma) = 1/\sigma$ and that there is no constraint on ϱ. Show that the conditional posterior distribution of ϱ, conditional on the observations $\mathbf{x}_{0:T}$ and on σ^2 is a $\mathcal{N}(\mu_T, \omega_T^2)$ distribution, with

$$\mu_T = \sum_{t=1}^{T} x_{t-1}x_t \Big/ \sum_{t=1}^{T} x_{t-1}^2 \quad \text{and} \quad \omega_T^2 = \sigma^2 \Big/ \sum_{t=1}^{T} x_{t-1}^2\,.$$

Show that the marginal posterior distribution of ϱ is a Student $\mathcal{T}(T-1, \mu_T, \nu_T^2)$ distribution, with

$$\nu_T^2 = \frac{1}{T-1}\left(\sum_{t=1}^{T} x_t^2 \Big/ \sum_{t=0}^{T-1} x_t^2 - \mu_T^2\right).$$

Apply this modeling to the AEGON series in Eurostoxx 50 and evaluate its predictive abilities.

Warning! The text above replaces the text of Exercise 7.4 in the first printing of the book, with $T-1$ degrees of freedom instead of T and a new expression for ν_T^2.

The posterior conditional density of ϱ is proportional to

$$\prod_{t=1}^{T} \exp\left\{-(x_t - \varrho\, x_{t-1})^2/2\sigma^2\right\}$$
$$\propto \exp\left\{\left[-\varrho^2 \sum_{t=0}^{T-1} x_t^2 + 2\varrho \sum_{t=0}^{T-1} x_t x_{t+1}\right]/2\sigma^2\right\},$$

which indeed leads to a $\mathcal{N}(\mu_T, \omega_T^2)$ conditional distribution as indicated above.

Given that the joint posterior density of (ϱ, σ) is proportional to

$$\sigma^{-T-1} \prod_{t=1}^{T} \exp\left\{-(x_t - \varrho\, x_{t-1})^2/2\sigma^2\right\}$$

integrating out σ leads to a density proportional to

$$\int \left(\sigma^2\right)^{-T/2-1/2} \exp\left(\sum_{t=1}^{T}(x_t-\rho x_{t-1})^2/(2\sigma^2)\right) \mathrm{d}\sigma$$
$$= \int \left(\sigma^2\right)^{-T/2-1} \exp\left(\sum_{t=1}^{T}(x_t-\rho x_{t-1})^2/(2\sigma^2)\right) \mathrm{d}\sigma^2$$
$$= \left\{\sum_{t=1}^{T}(x_t-\varrho\, x_{t-1})^2\right\}^{-T/2}$$

when taking into account the Jacobian. We thus get a Student $\mathscr{T}(T-1, \mu_T, \nu_T^2)$ distribution and the parameters can be derived from expanding the sum of squares:

$$\sum_{t=1}^{T}(x_t-\varrho\, x_{t-1})^2 = \sum_{t=0}^{T-1} x_t^2\left(\varrho^2-2\varrho\mu_T\right) + \sum_{t=1}^{T} x_t^2$$

into

$$\sum_{t=0}^{T-1} x_t^2(\varrho-\mu_T)^2 + \sum_{t=1}^{T} x_t^2 - \sum_{t=0}^{T-1} x_t^2\mu_T^2$$
$$\propto \frac{(\varrho-\mu_T)^2}{T-1} + \frac{1}{T-1}\left(\frac{\sum_{t=1}^{T} x_t^2}{\sum_{t=0}^{T-1} x_t^2} - \mu_T^2\right)$$
$$= \frac{(\varrho-\mu_T)^2}{T-1} + \nu_T^2\,.$$

The main point with this example is that, when ϱ is unconstrained, the joint posterior distribution of (ϱ, σ) is completely closed-form. Therefore, the predictive distribution of x_{T+1} is given by

$$\int \frac{1}{\sqrt{2\pi}\sigma} \exp\{-(x_{T+1}-\varrho x_T)^2/2\sigma^2\}\, \pi(\sigma, \varrho|\mathbf{x}_{0:T})\mathrm{d}\sigma\mathrm{d}\varrho$$

which has again a closed-form expression:

$$\begin{aligned}
&\int \frac{1}{\sqrt{2\pi}\sigma} \exp\{-(x_{T+1}-\varrho x_T)^2/2\sigma^2\}\,\pi(\sigma,\varrho|\mathbf{x}_{0:T})\mathrm{d}\sigma\mathrm{d}\varrho \\
&\qquad \propto \int \sigma^{-T-2}\exp\{-\sum_{t=0}^{T}(x_{t+1}-\varrho x_t)^2/2\sigma^2\}\mathrm{d}\sigma\mathrm{d}\varrho \\
&\qquad \propto \int \left\{\sum_{t=0}^{T}(x_{t+1}-\varrho\, x_t)^2\right\}^{-(T+1)/2} \mathrm{d}\varrho \\
&\qquad \propto \left(\sum_{t=0}^{T} x_t^2\right)^{-(T+1)/2} \int \left\{\frac{(\varrho-\mu_{T+1})^2}{T}+\nu_{T+1}^2\right\}^{-(T+2)/2} \mathrm{d}\varrho \\
&\qquad \propto \left(\sum_{t=0}^{T} x_t^2\right)^{-(T+1)/2} \nu_T^{-T-1} \\
&\qquad \propto \left(\sum_{t=0}^{T} x_t^2 \sum_{t=0}^{T} x_{t+1}^2 - \left\{\sum_{t=0}^{T} x_t x_{t+1}\right\}^2\right)^{(T+1)/2} .
\end{aligned}$$

This is a Student $\mathcal{T}(T,\delta_T,\omega_T)$ distribution, with

$$\delta_T = x_T \sum_{t=0}^{T-1} x_t x_{t+1} \Big/ \sum_{t=0}^{T-1} x_t^2 = \hat{\rho}_T x_T$$

and

$$\omega_T = \left\{\sum_{t=0}^{T} x_t^2 \sum_{t=0}^{T} x_t^2 - \left(\sum_{t=0}^{T} x_t x_{t+1}\right)^2\right\} \Big/ T\sum_{t=0}^{T-1} x_t^2 .$$

The predictive abilities of the model are thus in providing a point estimate for the next observation $\hat{x}_{T+1} = \hat{\rho}_T x_T$, and a confidence band around this value.

Exercise 7.5 Give the necessary and sufficient condition under which an AR(2) process with autoregressive polynomial $\mathcal{P}(u) = 1 - \varrho_1 u - \varrho_2 u^2$ (with $\varrho_2 \neq 0$) is causal.

The AR(2) process with autoregressive polynomial $\mathcal{P}$ is causal if and only if the roots of $\mathcal{P}$ are outside the unit circle in the complex plane. The roots of $\mathcal{P}$ are given by

$$u^- = \frac{-\varrho_1 - \sqrt{\varrho_1^2+4\varrho_2}}{-2\varrho_2} \quad \text{and} \quad u^+ = \frac{-\varrho_1 + \sqrt{\varrho_1^2+4\varrho_2}}{-2\varrho_2}$$

with the convention that $\sqrt{x} = \iota\sqrt{-x} \in \mathbb{C}$ if $x < 0$. (Because of the symmetry of the roots wrt ρ_1, the causality region will be symmetric in ρ_1.)

A first empirical approach based on simulation is to produce a sample of (ϱ_1, ϱ_2)'s over the sphere of radius 6 (6 is chosen arbitrarily and could be changed if this is too small a bound) and to plot only those (ϱ_1, ϱ_2)'s for which the roots u^- and u^+ are outside the unit circle.

```
# Number of points
N=10^4

# Sample of rho's
rho1=rnorm(N)
rho2=rnorm(N)
rad=6*runif(N)/sqrt(rho1^2+rho2^2)
rho1=rad*rho1
rho2=rad*rho2
R=matrix(1,ncol=3,nrow=N)
R[,2]=-rho1
R[,3]=-rho2

roots=apply(R,1,polyroot)
indx=(1:N)[(Mod(roots[1,])>1)]
indx=indx[(Mod(roots[2,indx])>1)]
plot(rho1[indx],rho2[indx],col="grey",cex=.4,
  xlab=expression(rho[1]),ylab=expression(rho[2]))
```

The output of this program is given on Figure 7.1 but, while it looks like a triangular shape, this does not define an analytical version of the restricted parameter space.

If we now look at the analytical solution, there are two cases to consider: either $\varrho_1^2 + 4\varrho_2 < 0$ and the roots are then complex numbers, or $\varrho_1^2 + 4\varrho_2 > 0$ and the roots are then real numbers.

If $\varrho_1^2 + 4\varrho_2 < 0$, then

$$\left| \frac{-\varrho_1 \pm i\sqrt{-(\varrho_1^2 + 4\varrho_2)}}{2\varrho_2} \right|^2 > 1$$

implies that $-1 < \varrho_2$, which, together with the constraint $\varrho_1^2 + 4\varrho_2 < 0$, provides a first region of values for (ϱ_1, ϱ_2):

$$\mathcal{C}_1 = \left\{(\varrho_1, \varrho_2);\ |\varrho_1| \leq 4, \varrho_2 < -\varrho_1^2/4\right\} .$$

If $\varrho_1^2 + 4\varrho_2 > 0$, then the condition $|u^\pm| > 1$ turns into

$$\sqrt{\varrho_1^2 + 4\varrho_2} - |\varrho_1| > 2\varrho_2 \quad \text{if} \quad \varrho_2 > 0 \qquad \text{and} |\varrho_1| - \sqrt{\varrho_1^2 + 4\varrho_2} > -2\varrho_2 \quad \text{if} \quad \varrho_2 < 0\,.$$

Thus, this amounts to compare $\sqrt{\varrho_1^2 + 4\varrho_2}$ with $2|\varrho_2 + |\varrho_1|$ or, equivalently, $4\varrho_2$ with $4\varrho_2^2 + 4|\varrho_1\varrho_2|$, ending up with the global condition $\varrho_2 < 1 - |\varrho_1$. The second part of the causality set is thus

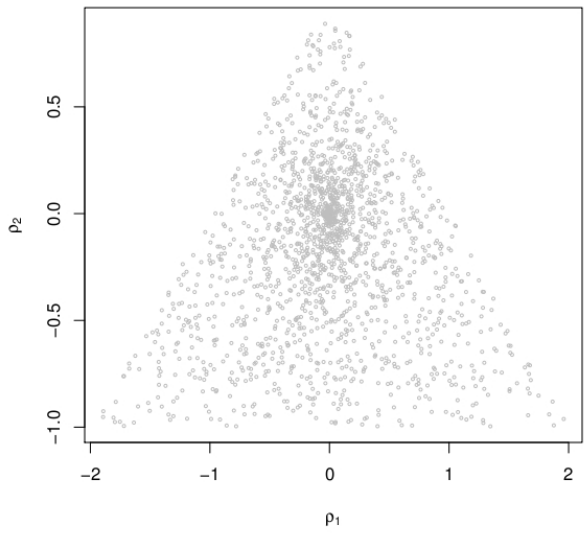

Fig. 7.1. Acceptable values of (ρ_1, ρ_2) for the $AR(2)$ model obtained by simulation.

$$\mathcal{C}_2 = \left\{(\varrho_1, \varrho_2);\, |\varrho_1| \leq 4, \varrho_2 < -\varrho_1^2/4\right\} ,$$

and cumulating both regions leads to the triangle

$$\mathcal{T} = \{(\varrho_1, \varrho_2);\, \varrho_2 > -1, \varrho_2 < 1 - |\varrho_1|\} .$$

Exercise 7.6 Show that the stationary distribution of $\mathbf{x}_{-p:-1}$ is a $\mathcal{N}_p(\mu\mathbf{1}_p, \mathbf{A})$ distribution, and give a fixed point equation satisfied by the covariance matrix $\mathbf{A}$.

If we denote

$$\mathbf{z}_t = (x_t, x_{t-1}, \ldots, x_{t+1-p}) ,$$

then

$$\mathbf{z}_{t+1} = \mu\mathbf{1}_p + B\left(\mathbf{z}_t - \mu\mathbf{1}_p\right) + \epsilon_{t+1} .$$

Therefore,

$$\mathbb{E}\left[\mathbf{z}_{t+1}|\mathbf{z}_t\right] = \mu\mathbf{1}_p + B\left(\mathbf{z}_t - \mu\mathbf{1}_p\right)$$

and

$$\mathbb{V}\left(\mathbf{z}_{t+1}|\mathbf{z}_t\right) = \mathbb{V}\left(\epsilon_{t+1}\right) = \begin{bmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 \end{bmatrix} = V .$$

Then,

$$\mathbf{z}_{t+1}|\mathbf{z}_t \sim \mathcal{N}_p\left(\mu\mathbf{1}_p + B\left(\mathbf{z}_t - \mu\mathbf{1}_p\right), V\right) .$$

Therefore, if $\mathbf{z}_{-1} = \mathbf{x}_{-p:-1} \sim \mathcal{N}_p\left(\mu\mathbf{1}_p, A\right)$ is Gaussian, then $\mathbf{z}_t$ is Gaussian. Suppose that $\mathbf{z}_t \sim \mathcal{N}_p(M, A)$, we get

$$\mathbb{E}\left[\mathbf{z}_{t+1}\right) = \mu\mathbf{1}_p + B\left(M - \mu\mathbf{1}_p\right]$$

and $\mathbb{E}\left[\mathbf{z}_{t+1}\right] = \mathbb{E}\left[\mathbf{z}_t\right]$ if

$$\mu\mathbf{1}_p + B\left(M - \mu\mathbf{1}_p\right) = M ,$$

which means that $M = \mu\mathbf{1}_p$. Similarly, $\mathbb{V}\left(\mathbf{z}_{t+1}\right) = \mathbb{V}\left(\mathbf{z}_t\right)$ if and only if

$$BAB' + V = A ,$$

which is the "fixed point" equation satisfied by A.

Exercise 7.7 Show that the posterior distribution on $\boldsymbol{\theta}$ associated with the prior $\pi(\boldsymbol{\theta}) = 1/\sigma^2$ is well-defined for $T > p$ observations.

Warning: The prior $\pi(\boldsymbol{\theta}) = 1/\sigma$ was wrongly used in the first printing of the book.

The likelihood conditional on the initial values $\mathbf{x}_{0:(p-1)}$ is proportional to

$$\sigma^{-T+p-1} \prod_{t=p}^{T} \exp\left\{ -\left(x_t - \mu - \sum_{i=1}^{p} \varrho_i (x_{t-i} - \mu) \right)^2 / 2\sigma^2 \right\} .$$

A traditional noninformative prior is $\pi(\mu, \varrho_1, \ldots, \varrho_p, \sigma^2) = 1/\sigma^2$. In that case, the probability density of the posterior distribution is proportional to

$$\sigma^{-T+p-3} \prod_{t=p}^{T} \exp\left\{ -\left(x_t - \mu - \sum_{i=1}^{p} \varrho_i (x_{t-i} - \mu) \right)^2 / 2\sigma^2 \right\} .$$

And

$$\int (\sigma^2)^{-(T-p+3)/2} \prod_{t=p}^{T} \exp\left\{ -\left(x_t - \mu - \sum_{i=1}^{p} \varrho_i (x_{t-i} - \mu) \right)^2 / 2\sigma^2 \right\} \mathrm{d}\sigma^2 < \infty$$

holds for $T - p + 1 > 0$, i.e., $T > p - 1$. This integral is equal to

$$\left\{ -\left(x_t - \mu - \sum_{i=1}^{p} \varrho_i (x_{t-i} - \mu) \right)^2 / 2\sigma^2 \right\}^{(p-T-1)/2} ,$$

which is integrable in μ for $T - p > 0$, i.e. $T > p$. The other parameters ϱ_j $(j = 1, \ldots, p0$ being bounded, the remaining integrand is clearly integrable in $\boldsymbol{\varrho}$.

Exercise 7.8 Show that the coefficients of the polynomial $\mathcal{P}$ can be derived in $\mathrm{O}(p^2)$ time from the inverse roots λ_i using the recurrence relations $(i = 1, \ldots, p, j = 0, \ldots, p)$

$$\psi_0^i = 1\,, \qquad \psi_j^i = \psi_j^{i-1} - \lambda_i \psi_{j-1}^{i-1}\,,$$

where $\psi_0^0 = 1$ and $\psi_j^i = 0$ for $j > i$, and setting $\varrho_j = -\psi_j^p$ $(j = 1, \ldots, p)$.

Warning: The useless sentence *"Deduce that the likelihood is computable in $O(Tp^2)$ time"* found in the first printing of the book has been removed.

Since

$$\prod_{i=1}^{p}(1 - \lambda_i x) = 1 - \sum_{j=1}^{j} \varrho_j x^j\,,$$

we can expand the lhs one root at a time. If we set

$$\prod_{j=1}^{i}(1 - \lambda_j x) = \sum_{j=0}^{i} \psi_j^i x^j\,,$$

‘ then

$$\begin{aligned}\prod_{j=1}^{i+1}(1 - \lambda_j x) &= (1 - \lambda_{i+1}x)\prod_{j=1}^{i}(1 - \lambda_j x)\\ &= (1 - \lambda_{i+1}x)\sum_{j=0}^{i}\psi_j^i x^j\\ &= 1 + \sum_{j=1}^{i}(\psi_j^i - \lambda_{i+1}\psi_{j-1}^i)x^j - \lambda_{i+1}\psi_i^i x^{i+1}\,,\end{aligned}$$

which establishes the $\psi_j^{i+1} = \psi_j^i - \lambda_{i+1}\psi_{j-1}^i$ recurrence relation.

This recursive process requires the allocation of i variables at the ith stage; the coefficients of $\mathcal{P}$ can thus be derived with a complexity of $\mathrm{O}(p^2)$.

Exercise 7.9 Show that, if the proposal on σ^2 is a log-normal distribution $\mathcal{LN}(\log(\sigma_{t-1}^2), \tau^2)$ and if the prior distribution on σ^2 is the noninformative prior $\pi(\sigma^2) = 1/\sigma^2$, the acceptance ratio also reduces to the likelihood ratio because of the Jacobian.

Warning: In the first printing of the book, there is a log missing in the mean of the log-normal distribution.

If we write the Metropolis–Hastings ratio for a current value σ_0^2 and a proposed value σ_1^2, we get

$$\frac{\pi(\sigma_1^2)\ell(\sigma_1^2)}{\pi(\sigma_0^2)\ell(\sigma_0^2)}\,\frac{\exp\left(-(\log(\sigma_0^2-\log(\sigma_1^2))^2/2\tau^2\right)/\sigma_0^2}{\exp\left(-(\log(\sigma_0^2-\log(\sigma_1^2))^2/2\tau^2\right)/\sigma_1^2}=\frac{\ell(\sigma_1^2)}{\ell(\sigma_0^2)}\,,$$

as indicated.

Exercise 7.10 Write an R program that extends the reversible jump algorithm 7.1 to the case when the order p is unknown and apply it to the same Ahold Kon. series of Eurostoxx 50.

The modification is rather straightforward if one only considers birth and death moves, adding and removing real or complex roots in the polynomial. When the new values are generated from the prior, as in the program provided by #7.txt on the Webpage, the acceptance probability remains equal to the likelihood ratio (with the usual modifications at the boundaries).

Exercise 7.11 For an MA(q) process, show that ($s \le q$)

$$\gamma_x(s)=\sigma^2\sum_{i=0}^{q-|s|}\vartheta_i\vartheta_{i+|s|}\,.$$

We have

$$\begin{aligned}\gamma_x(s)&=\mathbb{E}\left[x_t x_{t-s}\right]\\&=\mathbb{E}\left[\left[\epsilon_t+\vartheta_1\epsilon_{t-1}+\ldots+\vartheta_q\epsilon_{t-q}\right]\left[\epsilon_{t-s}+\vartheta_1\epsilon_{t-s-1}+\ldots+\vartheta_q\epsilon_{t-s-q}\right]\right]\,.\end{aligned}$$

Then, if $1 \le s \le q$,

$$\gamma_x(s)=\left[\vartheta_s+\vartheta_{s+1}\vartheta_1+\ldots+\vartheta_q\vartheta_{q-s}\right]\sigma^2$$

and

$$\gamma_x(0)=\left[1+\vartheta_1^2+\ldots+\vartheta_q^2\right]\sigma^2\,.$$

Therefore, if ($0 \le s \le q$) with the convention that $\vartheta_0 = 1$

$$\gamma_x(s)=\sigma^2\sum_{i=0}^{q-s}\vartheta_i\vartheta_{i+s}\,.$$

The fact that $\gamma_x(s)=\gamma_x(-s)$ concludes the proof.

Exercise 7.12 Show that the conditional distribution of $(\epsilon_0, \ldots, \epsilon_{-q+1})$ given both $\mathbf{x}_{1:T}$ and the parameters is a normal distribution. Evaluate the complexity of computing the mean and covariance matrix of this distribution.

The distribution of $\mathbf{x}_{1:T}$ conditional on $(\epsilon_0, \ldots, \epsilon_{-q+1})$ is proportional to

$$\sigma^{-T} \prod_{t=1}^{T} \exp\left\{ -\left(x_t - \mu + \sum_{j=1}^{q} \vartheta_j \widehat{\epsilon}_{t-j} \right)^2 \Big/ 2\sigma^2 \right\} ,$$

Take

$$(\epsilon_0, \ldots, \epsilon_{-q+1}) \sim \mathcal{N}_q\left(0_q, \sigma^2 I_q\right) .$$

In that case, the conditional distribution of $(\epsilon_0, \ldots, \epsilon_{-q+1})$ given $\mathbf{x}_{1:T}$ is proportional to

$$\prod_{i=-q+1}^{0} \exp\left\{-\epsilon_i^2/2\sigma^2\right\} \prod_{t=1}^{T} \exp\left\{-\widehat{\epsilon}_t^2/2\sigma^2\right\} .$$

Due to the recursive definition of $\hat{\epsilon}_t$, the computation of the mean and the covariance matrix of this distribution is too costly to be available for realistic values of T. For instance, getting the conditional mean of ϵ_i requires deriving the coefficients of ϵ_i from all terms

$$\left(x_t - \mu + \sum_{j=1}^{q} \vartheta_j \widehat{\epsilon}_{t-j} \right)^2$$

by exploiting the recursive relation

$$\widehat{\epsilon}_t = x_t - \mu + \sum_{j=1}^{q} \vartheta_j \widehat{\epsilon}_{t-j} .$$

If we write $\widehat{\epsilon}_1 = \delta_1 + \beta_1 \epsilon_i$ and $\widehat{\epsilon}_t = \delta_t + \beta_t \epsilon_i$, then we need to use the recursive formula

$$\delta_t = x_t - \mu + \sum_{j=1}^{q} \vartheta_j \delta_{t-j} , \qquad \beta_t = \sum_{j=1}^{q} \beta_{t-j} ,$$

before constructing the conditional mean of ϵ_i. The corresponding cost for this single step is therefore $\mathrm{O}(Tq)$ and therefore $\mathrm{O}(qT^2)$ for the whole series of ϵ_i's. Similar arguments can be used for computing the conditional variances.

Exercise 7.13 Give the conditional distribution of ϵ_{-t} given the other ϵ_{-i}'s, $\mathbf{x}_{1:T}$, and the $\widehat{\epsilon}_i$'s. Show that it only depends on the other ϵ_{-i}'s, $\mathbf{x}_{1:q-t+1}$, and $\widehat{\epsilon}_{1:q-t+1}$.

The formulation of the exercise is slightly too vague in that the $\widehat{\epsilon}_i$'s are deterministic quantities based on the ϵ_{-i}'s and $\mathbf{x}_{1:T}$. Thus, from a probabilistic point of view, the conditional distribution of ϵ_{-t} only depends on the other ϵ_{-i}'s and $\mathbf{x}_{1:T}$. However, from an algorithmic point of view, if we take the $\widehat{\epsilon}_i$'s as additional observables in (7.11), spotting $\epsilon_{-\ell}$ in

$$\sum_{t=1}^{T}\left(x_t-\mu+\sum_{j=1}^{q}\vartheta_j\widehat{\epsilon}_{t-j}\right)^2$$

leads to keep only

$$\sum_{t=1}^{q-\ell}\left(x_t-\mu+\sum_{j=1}^{t-1}\vartheta_j\widehat{\epsilon}_{t-j}+\sum_{j=t}^{q}\vartheta_j\epsilon_{t-j}\right)^2$$

in the sum since, for $t-q>-\ell$, i.e. for $t>q-\ell$, $\epsilon_{-\ell}$ does not appear in the distribution. (Again, this is a formal construct that does not account for the deterministic derivation of the $\widehat{\epsilon}_i$'s.) The conditional distribution of $\epsilon_{-\ell}$ is then obviously a normal distribution.

Exercise 7.14 Show that the predictive horizon for the MA(q) model is restricted to the first q future observations x_{t+i}.

Obviously, due to the lack of correlation between x_{T+q+j} $(j>0)$ and $\mathbf{x}_{1:T}$ we have

$$\mathbb{E}\left[x_{T+q+1}|\mathbf{x}_{1:T}\right]=\mathbb{E}\left[x_{T+q+1}\right]=0$$

and therefore the $MA(q)$ model has no predictive ability further than horizon q.

Exercise 7.15 Show that, when the support $\mathcal{Y}$ is finite and when $(y_t)_{t\in\mathbb{N}}$ is stationary, the marginal distribution of x_t is the same mixture distribution for all t's. Deduce that the same identifiability problem as in mixture models occurs in this setting.

Since the marginal distribution of x_t is given by

$$\int f(x_t|y_t)\pi(y_t)\,\mathrm{d}y_t=\sum_{y\in\mathcal{Y}}\pi(y)f(x_t|y)\,,$$

where π is the stationary distribution of (y_t), this is indeed a mixture distribution. Although this is not the fundamental reason for the unidentifiability

of hidden Markov models, there exists an issue of label switching similar to the case of standard mixtures.

Exercise 7.16 Write down the joint distribution of $(y_t, x_t)_{t\in\mathbb{N}}$ in (7.19) and deduce that the (observed) likelihood is not available in closed form.

Recall that $y_0 \sim \mathcal{N}(0, \sigma^2)$ and, for $t = 1, \ldots, T$,

$$\begin{cases} y_t = \varphi y_{t-1} + \sigma\epsilon^*_{t-1}\,, \\ x_t = \beta e^{y_t/2}\epsilon_t\,, \end{cases}$$

where both ϵ_t and ϵ^*_t are iid $\mathcal{N}(0,1)$ random variables. The joint distribution of $(\mathbf{x}_{1:T}, \mathbf{y}_{0:T})$ is therefore

$$\begin{aligned} f\left(\mathbf{x}_{1:T}, \mathbf{y}_{0:T}\right) &= f\left(\mathbf{x}_{1:T}|\mathbf{y}_{0:T}\right) f\left(\mathbf{y}_{0:T}\right) \\ &= \left(\prod_{i=1}^{T} f(x_i|y_i)\right) f(y_0) f(y_1|y_0)\ldots f(y_T|y_{T-1}) \\ &= \frac{1}{\left(2\pi\beta^2\right)^{T/2}} \exp\left\{-\sum_{t=1}^{T} y_t/2\right) \exp\left(-\frac{1}{2\beta^2}\sum_{t=1}^{T} x_t^2 \exp(-y_t)\right) \\ &\quad \times \frac{1}{\left(2\pi\sigma^2\right)^{(T+1)/2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_0^2 + \sum_{t=1}^{T}\left(y_t - \varphi y_{t-1}\right)^2\right)\right\}\,. \end{aligned}$$

Due to the double exponential term $\exp\left(-\frac{1}{2\beta^2}\sum_{t=1}^{T} x_t^2\exp(-y_t)\right)$, it is impossible to find a closed-form of the integral in $\mathbf{y}_{0:T}$.

Exercise 7.17 Show that the counterpart of the prediction filter in the Markov-switching case is given by

$$\log p(\mathbf{x}_{1:t}) = \sum_{r=1}^{t} \log\left[\sum_{i=1}^{\kappa} f(x_r|x_{r-1}, y_r = i)\varphi_r(i)\right]\,,$$

where $\varphi_r(i) = \mathbb{P}(y_r = i|\mathbf{x}_{1:r-1})$ is given by the recursive formula

$$\varphi_r(i) \propto \sum_{j=1}^{\kappa} p_{ji} f(x_{r-1}|x_{r-2}, y_{r-1} = j)\varphi_{r-1}(j)\,.$$

Warning! There is a typo in the first printing of the book where φ_r is defined conditional on $\mathbf{x}_{1:t-1}$ instead of $\mathbf{x}_{1:r-1}$.

This exercise is more or less obvious given the developments provided in the book. The distribution of y_r given the past values $\mathbf{x}_{1:r-1}$ is the marginal of (y_r, y_{r-1}) given the past values $\mathbf{x}_{1:r-1}$:

$$\begin{aligned}
\mathbb{P}(y_r = i|\mathbf{x}_{1:t-1}) &= \sum_{j=1}^{\kappa} \mathbb{P}(y_r = i, y_{r-1} = j|\mathbf{x}_{1:r-1}) \\
&= \sum_{j=1}^{\kappa} \mathbb{P}(y_{r-1} = j|\mathbf{x}_{1:r-1})\, \mathbb{P}(y_r = i|y_{r-1} = j) \\
&\propto \sum_{j=1}^{\kappa} p_{ji} \mathbb{P}(y_{r-1} = j, x_{r-1}|\mathbf{x}_{1:r-2}) \\
&= \sum_{j=1}^{\kappa} p_{ji} \mathbb{P}(y_{r-1} = j, |\mathbf{x}_{1:r-2}) f(x_{r-1}|x_{r-2}, y_{r-1} = j)\,,
\end{aligned}$$

which leads to the update formula for the $\varphi_r(i)$'. The marginal distribution $\mathbf{x}_{1:t}$ is then derived by

$$\begin{aligned}
p(\mathbf{x}_{1:t}) &= \prod_{r=1}^{t} p(x_r|\mathbf{x}_{1:(r-1)}) \\
&= \prod_{r=1}^{t} \sum_{j=1}^{\kappa} \mathbb{P}(y_{r-1} = j, x_r|\mathbf{x}_{1:r-1}) \\
&= \prod_{r=1}^{t} \sum_{j=1}^{\kappa} f(x_r|x_{r-1}, y_r = i)\varphi_r(i)\,,
\end{aligned}$$

with the obvious convention $\varphi_1(i) = \pi_i$, if $(\pi_1, \ldots, \pi_\kappa)$ is the stationary distribution associated with $\mathbb{P} = (p_{ij})$.

8

Image Analysis

Exercise 8.1 Draw an example with $n = 5$ points in $\mathbb{R}^2$ such that the k-nearest-neighbor relation is not symmetric.

The first quadrant of Figure 8.2 in the book is already an illustration of an assymmetric neighborhood relation. Once two points are drawn, it is sufficient to find sequentialy new points that are closer to the latest than the earlier points. Figure 8.1 illustrates this case in dimension one, with decreasing-radius circles to emphasize the assymetry: each point on the line is such that its right neighbor is its nearest neighbor and that it is the nearest neighbor of its left neighbor.

Exercise 8.2 For a given pair (k, n) and a uniform distribution of $\mathbf{x}$ in $[0, 1]^3$, design a Monte Carlo experiment that evaluates the distribution of the size of the symmetrized k-nearest-neighborhood.

For a given pair (k, n), the Monte Carlo experiment produces N random samples on $[0, 1]^3$, for instance as

```
samp=matrix(runif(3*n),n,3)
```

compute the k-nearest-neighborhood matrix for this sample, by

```
disamp=as.matrix(dist(samp,up=T,diag=T))   #distance matrix

neibr=t(apply(disamp,1,order))[,-1]        #k nearest neighbours
knnbr=neibr[,1:k]
```

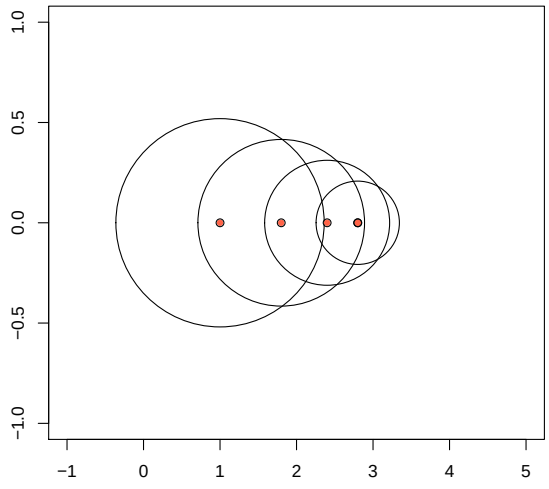

Fig. 8.1. Sequence of five points with assymetry in the nearest-neighbor relations.

```
newnbr=matrix(0,n,n)                          #k nearest neighbours
for (i in 1:n)                                #indicator matrix
  newnbr[i,knnbr[i,]]=1
```

and compute the sizes of the symmetrized k-nearest-neighborhoods for those samples, by

```
size[t,]=apply(newnbr+t(newnbr)>0,1,sum)
```

ending up with an approximation to the distribution of the size over the samples. It is then possible to summarize the matrix `size` on an histogram as in Figure 8.2.

Exercise 8.3 When $\mathbf{y} = (y_1, y_2)$, show that the joint pdf of $\mathbf{y}$ is given by

$$f(\mathbf{y}|\mathbf{X}, \beta, k) = f(y_1|y_2, \mathbf{X}, \beta, k) \Big/ \sum_{g=1}^{G} \frac{f(C_g|y_2, \mathbf{X}, \beta, k)}{f(y_2|C_g, \mathbf{X}, \beta, k)}\,.$$

Discuss the extension to the general case. (*Indication*: The extension is solved via the Hammersley–Clifford theorem, given in Section 8.3.1.)

This exercice is a consequence of Exercice 3.21: we have that, for all y_2's,

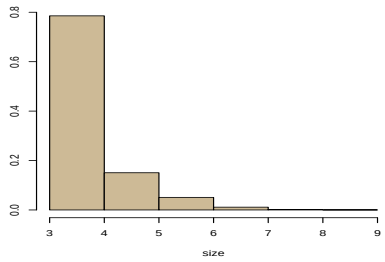

Fig. 8.2. Histogram of the size of the symmetrized k-nearest-neighborhoods when $k = 3$ and $n = 200$ based on $N = 1000$ simulations.

$$f(y_1|\mathbf{X},\beta,k) = \frac{f(y_1|y_2,\mathbf{X},\beta,k)/f(y_2|y_1,\mathbf{X},\beta,k)}{\int f(y_1|y_2,\mathbf{X},\beta,k)/f(y_2|y_1,\mathbf{X},\beta,k)\mathrm{d}y_1} = \frac{f(y_1|y_2,\mathbf{X},\beta,k)/f(y_2|y_1,\mathbf{X},\beta,k)}{\sum_{g=1}^{G} f(C_g|y_2,\mathbf{X},\beta,k)/f(y_2|C_g,\mathbf{X},\beta,k)},$$

since the support of y_1 is finite. We can therefore conclude that

$$f(\mathbf{y}|\mathbf{X},\beta,k) = \frac{f(y_1|y_2,\mathbf{X},\beta,k)}{\sum_{g=1}^{G} f(C_g|y_2,\mathbf{X},\beta,k)/f(y_2|C_g,\mathbf{X},\beta,k)}.$$

As suggested, the extension is solved via the Hammersley–Clifford theorem, given in (8.4). See Section 8.3.1 for details.

Exercise 8.4 Find two conditional distributions $f(x|y)$ and $g(y|x)$ such that there is no joint distribution corresponding to both f and g. Find a necessary condition for f and g to be compatible in that respect, i.e. to correspond to a joint distribution on (x, y).

As stated, this is a rather obvious question: if $f(x|y) = 4y\exp(-4yx)$ and if $g(y|x) = 6x\exp(-6xy)$, there cannot be a joint distribution inducing these two conditionals. What is more interesting is that, if $f(x|y) = 4y\exp(-4yx)$ and $g(y|x) = 4x\exp(-4yx)$, there still is no joint distribution, despite the formal agreement between both conditionals: the only joint that would work has the major drawback that it has an infinite mass!

Exercise 8.5 Using the Hammersley-Clifford theorem, show that the full conditional distributions given by (8.1) are compatible with a joint distribution.

Note: In order to expose the error made in the first printing in using the size of the symmetrized neighborhood, $N_k(i)$, we will compute the potential joint distribution based on the pseudo-conditional

$$\mathbb{P}(y_i = C_j|\mathbf{y}_{-i}, \mathbf{X}, \beta, k) \propto \exp\left(\beta \sum_{\ell \sim_k i} \mathbb{I}_{C_j}(y_\ell) \Big/ N_k(i)\right) ,$$

even though it is defined for a fixed $N_k(i) = N_k$ in the book.

It follows from (8.4) that, if there exists a joint distribution, it satisfies

$$\mathbb{P}(\mathbf{y}|\mathbf{X}, \beta, k) \propto \prod_{i=0}^{n-1} \frac{\mathbb{P}(y_{i+1}|y_1^*, \ldots, y_i^*, y_{i+2}, \ldots, y_n, \mathbf{X}, \beta, k)}{\mathbb{P}(y_{i+1}^*|y_1^*, \ldots, y_i^*, y_{i+2}, \ldots, y_n, \mathbf{X}, \beta, k)} .$$

Therefore,

$$\mathbb{P}(\mathbf{y}|\mathbf{X}, \beta, k) \propto \exp\left\{\beta \sum_{i=1}^{n} \frac{1}{N_k(i)} \left(\sum_{\ell<i, \ell \sim_k i} \left[\mathbb{I}_{y_\ell^*}(y_i) - \mathbb{I}_{y_\ell^*}(y_i^*)\right] + \sum_{\ell>i, \ell \sim_k i} \left[\mathbb{I}_{y_\ell}(y_i) - \mathbb{I}_{y_\ell}(y_i^*)\right]\right)\right\}$$

is the candidate joint distribution. Unfortunately, if we now try to derive the conditional distribution of y_j from this joint, we get

$$\mathbb{P}(y_i = C_j|\mathbf{y}_{-i}, \mathbf{X}, \beta, k) \propto \exp\beta \left\{\frac{1}{N_k(j)} \sum_{\ell>j, \ell \sim_k j} \mathbb{I}_{y_\ell}(y_j) + \sum_{\ell<j, \ell \sim_k j} \frac{\mathbb{I}_{y_\ell}(y_j)}{N_k(\ell)} + \frac{1}{N_k(j)} \sum_{\ell<j, \ell \sim_k j} \mathbb{I}_{y_\ell^*}(y_j) - \sum_{\ell<j, \ell \sim_k j} \frac{\mathbb{I}_{y_\ell^*}(y_j)}{N_k(\ell)}\right\}$$

which differs from the orginal conditional if the $N_k(j)$'s differ. In conclusion, there is no joint distribution if (8.1) is defined as in the first printing. Taking all the $N_k(j)$'s equal leads to a coherent joint distribution since the last line in the above equation cancels.

Exercise 8.6 If a joint density $\pi(y_1, ..., y_n)$ is such that the conditionals $\pi(y_{-i}|y_i)$ never cancel on the supports of the marginals $m_{-i}(y_{-i})$, show that the support of π is equal to the cartesian product of the supports of the marginals.

Warning! This exercise replaces the former version of Exercise 8.6: *"If $\pi(x_1, \ldots, x_n)$ is a density such that its full conditionals never cancel on its*

support, characterize the support of π in terms of the supports of the marginal distributions."

Let us suppose that the support of π is not equal to the product of the supports of the marginals. (This means that the support of π is smaller than this product.) Then the conditionals $\pi(\mathbf{y}_{-i}|y_i)$ cannot be positive everywhere on the support of $m(\mathbf{y}_{-i})$.

Exercise 8.7 Describe the collection of cliques $\mathcal{C}$ for an 8 neighbor neighborhood structure such as in Figure 8.7 on a regular $n \times m$ array. Compute the number of cliques.

If we draw a detailed graph of the connections on a regular grid as in Figure 8.3, then the maximal structure such that all members are neighbors is made of 4 points. Cliques are thus made of squares of 4 points and there are $(n-1) \times (m-1)$ cliques on a $n \times m$ array.

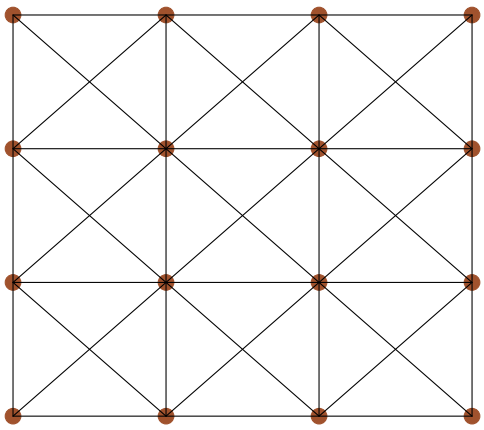

Fig. 8.3. Neighborhood relations between the points of a 4×4 regular grid for a 8 neighbor neighborhood structure.

Exercise 8.8 Use the Hammersley-Clifford theorem to establish that (8.6) is the joint distribution associated with the above conditionals. Deduce that the Ising model is a MRF.

Following the developments in Exercise 8.5, this is exactly the same problem as for the distribution (8.1) with a fixed neighborhood structure and the use of β instead of β/N_k.

Exercise 8.9 Draw the function $Z(\beta)$ for a 3×5 array. Determine the computational cost of the derivation of the normalizing constant $Z(\beta)$ of (8.6) for a $m \times n$ array.

The function $Z(\beta)$ is defined by

$$Z(\beta) = 1 \Big/ \sum_{\mathbf{x} \in \mathcal{X}} \exp\left(\beta \sum_{j \sim i} \mathbb{I}_{x_j = x_i}\right) ,$$

which involves a summation over the set $\mathcal{X}$ of size 2^{15}. The R code corresponding to this summation is

```
neigh=function(i,j){ #Neighbourhood indicator function

   (i==j+1)||(i==j-1)||(i==j+5)||(i==j-5)
}

zee=function(beta){

  val=0
  array=rep(0,15)

  for (i in 1:(2^15-1)){

    expterm=0
    for (j in 1:15)
      expterm=expterm+sum((array==array[j])*neigh(i=1:15,j=j))

    val=val+exp(beta*expterm)

    j=1
    while (array[j]==1){
```

```
        array[j]=0
        j=j+1
      }
      array[j]=1

    }

  expterm=0
  for (j in 1:15)
      expterm=expterm+sum((array==array[j])*neigh(i=1:15,j=j))

  val=val+exp(beta*expterm)

  1/val
}
```

It produces the (exact) curve given in Figure 8.4.

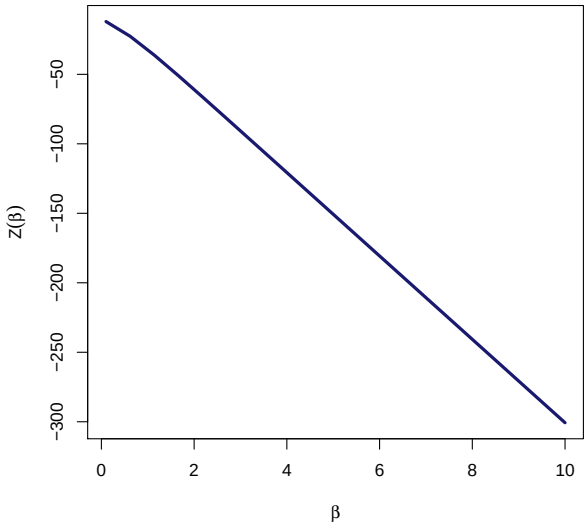

Fig. 8.4. Plot of the function $Z(\beta)$ for a 3×5 array with a four neighbor structure.

In the case of a $m \times n$ array, the summation involves $2^{m \times n}$ and each exponential term in the summation requires $(m \times n)^2$ evaluations, which leads to a $\mathrm{O}((m \times n)^2\, 2^{m \times n})$ overall cost.

Exercise 8.10 For an $n \times m$ array $\mathcal{I}$, if the neighborhood relation is based on the four nearest neighbors as in Figure 8.7, show that the $x_{i,j}$'s for which $(i+j) \equiv 0(2)$ are independent conditional on the $x_{i,j}$'s for which $(i+j) \equiv 1(2)$ $(1 \leq i \leq n, 1 \leq j \leq m)$. Deduce that the update of the whole image can be done in two steps by simulating the pixels with even sums of indices and then the pixels with odd sums of indices. (This modification of Algorithm 8.2 is a version of *the Swendsen–Wang* algorithm.)

Warning! This exercise replaces the former version of Exercise 8.10 *"For an $n \times m$ array $\mathcal{I}$, if the neighborhood relation is based on the four nearest neighbors, show that the $x_{2i,2j}$'s are independent conditional on the $x_{2i-1,2j-1}$'s $(1 \leq i \leq n, 1 \leq j \leq m)$. Deduce that the update of the whole image can be done in two steps by simulating the pixels with even indices and then the pixels with odd indices"*

This exercise is simply illustrating in the simplest case the improvement brought by the Swendsen-Wang algorithm upon the Gibbs sampler for image processing.

As should be obvious from Figure 8.7 in the book, the dependence graph between the nodes of the array is such that a given $x_{i,j}$ is independent from all the other nodes, conditional on its four neighbours. When $(i+j) \equiv 0(2)$, the neighbours have indices (i,j) such that $(i+j) \equiv 1(2)$, which establishes the first result.

Therefore, a radical alternative to the node-by-node update is to run a Gibbs sampler with two steps: a first step that updates the nodes $x_{i,j}$ with even $(i+j)$'s and a step that updates the nodes $x_{i,j}$ with odd $(i+j)$'s. This is quite a powerful solution in that it achieves the properties of two-stage Gibbs sampling, as for instance the Markovianity of the subchains generated at each step (see Robert and Casella, 2004, Chapter 9, for details).

Exercise 8.11 Determine the computational cost of the derivation of the normalizing constant of the distribution (8.7) for a $m \times n$ array and G different colors.

Just as in Exercise 8.9, finding the exact normalizing requires summing over all possible values of $\mathbf{x}$, which involves $G^{m \times n}$ terms. And each exponential term involves a sum over $(m \times n)^2$ terms, even though clever programing of the neighborhood system may reduce the computational cost down to $m \times n$. Overall, the normalizing constant faces a computing cost of at least $\mathrm{O}(m \times n \times G^{m \times n})$.

Exercise 8.12 Use the Hammersley-Clifford theorem to establish that (8.7) is the joint distribution associated with the above conditionals. Deduce that the Potts model is a MRF.

Similar to the resolution of Exercise 8.5, using the Hammersley-Clifford representation (8.4) and defining an arbitrary order on the set $\mathcal{I}$ leads to the joint distribution

$$\begin{aligned}\pi(\mathbf{x}) &\propto \frac{\exp\left\{\beta \sum_{i\in\mathcal{I}} \sum_{j<i, j\sim i} \mathbb{I}_{x_i=x_j} + \sum_{j>i, j\sim i} \mathbb{I}_{x_i=x_j^\star}\right\}}{\exp\left\{\beta \sum_{i\in\mathcal{I}} \sum_{j<i, j\sim i} \mathbb{I}_{x_i^\star=x_j} + \sum_{j>i, j\sim i} \mathbb{I}_{x_i^\star=x_j^\star}\right\}} \\ &\propto \exp\left\{\beta\left(\sum_{j\sim i, j<i} \mathbb{I}_{x_i=x_j} + \sum_{j\sim i, j>i} \mathbb{I}_{x_i=x_j^\star} - \sum_{j\sim i, j>i} \mathbb{I}_{x_j^\star=x_i}\right)\right\} \\ &= \exp\left\{\beta \sum_{j\sim i} \mathbb{I}_{x_i=x_j}\right\}.\end{aligned}$$

So we indeed recover a joint distribution that is compatible with the initial full conditionals of the Potts model. The fact that the Potts is a MRF is obvious when considering its conditional distributions.

Exercise 8.13 Derive an alternative to Algorithm 8.3 where the probabilities in the multinomial proposal are proportional to the numbers of neighbors $n_{u_\ell,g}$ and compare its performance with those of Algorithm 8.3.

In Step 2 of Algorithm 8.3, another possibility is to select the proposed value of x_{u_ℓ} from a multinomial distribution

$$\mathcal{M}_G\left(1; n_1^{(t)}(u_\ell), \ldots, n_G^{(t)}(u_\ell)\right)$$

where $n_g^{(t)}(u_\ell)$ denotes the number of neighbors of u_l that take the value g. This is likely to be more efficient than a purely random proposal, especially when the value of β is high.

Exercise 8.14 Show that the Swendsen-Wang improvement given in Exercise 8.10 also applies to the simulation of $\pi(\mathbf{x}|\mathbf{y}, \beta, \sigma^2, \boldsymbol{\mu})$.

This is kind of obvious when considering that taking into account the values of the y_i's does not modify the dependence structure of the Potts model. Therefore, if there is a decomposition of the grid $\mathcal{I}$ into a small number of sub-grids $\mathcal{I}_1, \ldots, \mathcal{I}_k$ such that all the points in $\mathcal{I}_j$ are independent from one another given the other $\mathcal{I}_\ell$'s, a k step Gibbs sampler can be proposed for the simulation of $\mathbf{x}$.

Exercise 8.15 Using a piecewise-linear interpolation of $f(\beta)$ based on the values $f(\beta^1), \ldots, f(\beta^M)$, with $0 < \beta_1 < \ldots < \beta_M = 2$, give the explicit value of the integral

$$\int_{\alpha_0}^{\alpha_1} \hat{f}(\beta)\, \mathrm{d}\beta$$

for any pair $0 \leq \alpha_0 < \alpha_1 \leq 2$.

This follows directly from the R program provided in `#8.txt`, with

$$\int_{\alpha_0}^{\alpha_1} \hat{f}(\beta)\, \mathrm{d}\beta \approx \sum_{i, \alpha_0 \leq \beta_i \leq \alpha_1} f(\beta_i)(\beta_{i+1} - \beta_i)\,,$$

with the appropriate corrections at the boundaries.

Exercise 8.16 Show that the estimators $\widehat{\mathbf{x}}$ that minimize the posterior expected losses $\mathbb{E}[L_1(\mathbf{x}, \widehat{\mathbf{x}})|\mathbf{y})]$ and $\mathbb{E}[L_2(\mathbf{x}, \widehat{\mathbf{x}})|\mathbf{y})]$ are $\widehat{\mathbf{x}}^{MPM}$ and $\widehat{\mathbf{x}}^{MAP}$, respectively.

Since

$$L_1(\mathbf{x}, \widehat{\mathbf{x}}) = \sum_{i \in \mathcal{I}} \mathbb{I}_{x_i \neq \hat{x}_i}\,,$$

the estimator $\widehat{\mathbf{x}}$ associated with L_1 is minimising

$$\mathbb{E}\left[\sum_{i \in \mathcal{I}} \mathbb{I}_{x_i \neq \hat{x}_i} \big| \mathbf{y}\right]$$

and therefore, for every $i \in \mathcal{I}$, $\hat{x}_i$ minimizes $\mathbb{P}(x_i \neq \hat{x}_i)$, which indeed gives the MPM as the solution. Similarly,

$$L_2(\mathbf{x}, \widehat{\mathbf{x}}) = \mathbb{I}_{\mathbf{x} \neq \widehat{\mathbf{x}}}$$

leads to $\widehat{\mathbf{x}}$ as the solution to

$$\min_{\widehat{\mathbf{x}}} \mathbb{E}\left[\mathbb{I}_{\mathbf{x} \neq \widehat{\mathbf{x}}} \big| \mathbf{y}\right] = \min_{\widehat{\mathbf{x}}} \mathbb{P}\left(\mathbf{x} \neq \widehat{\mathbf{x}} \big| \mathbf{y}\right)\,,$$

which means that $\widehat{\mathbf{x}}$ is the posterior mode.

Exercise 8.17 Determine the estimators $\widehat{\mathbf{x}}$ associated with two loss functions that penalize differently the classification errors,

$$L_3(\mathbf{x},\widehat{\mathbf{x}}) = \sum_{i,j\in\mathcal{I}} \mathbb{I}_{x_i=x_j}\,\mathbb{I}_{\hat{x}_i\neq\hat{x}_j} \quad \text{and} \quad L_4(\mathbf{x},\widehat{\mathbf{x}}) = \sum_{i,j\in\mathcal{I}} \mathbb{I}_{x_i\neq x_j}\,\mathbb{I}_{\hat{x}_i=\hat{x}_j}$$

Even though L_3 and L_4 are very similar, they enjoy completely different properties. In fact, L_3 is basically useless because $\widehat{\mathbf{x}} = (1,\cdots,1)$ is always an optimal solution!

If we now look at L_4, we first notice that this loss function is invariant by permutation of the classes in $\mathbf{x}$: all that matters are the groups of components of $\mathbf{x}$ taking the same value. Minimizing this loss function then amounts to finding a clustering algorithm. To achieve this goal, we first look at the difference in the risks when allocating an arbitrary $\hat{x}_i$ to the value a and when allocating $\hat{x}_i$ to the value b. This difference is equal to

$$\sum_{j,\hat{x}_j=a} \mathbb{P}(x_i = x_j) - \sum_{j,\hat{x}_j=b} \mathbb{P}(x_i = x_j)\,.$$

It is therefore obvious that, for a given configuration of the other x_j's, we should pick the value a that minimizes the sum $\sum_{j,\hat{x}_j=a} \mathbb{P}(x_i = x_j)$. Once x_i is allocated to this value, a new index ℓ is to be chosen for possible reallocation until the scheme has reached a fixed configuration, that is, no $\hat{x}_i$ need reallocation.

This scheme produces a smaller risk at each of its steps so it does necessarily converge to a fixed point. What is less clear is that this produces the global minimum of the risk. An experimental way of checking this is to run the scheme with different starting points and to compare the final values of the risk.

Exercise 8.18 Since the maximum of $\pi(\mathbf{x}|\mathbf{y})$ is the same as that of $\pi(\mathbf{x}|\mathbf{y})^\kappa$ for every $\kappa \in \mathbb{N}$, show that

$$\pi(\mathbf{x}|\mathbf{y})^\kappa = \int \pi(\mathbf{x},\theta_1|\mathbf{y})\,\mathrm{d}\theta_1 \times \cdots \times \int \pi(\mathbf{x},\theta_\kappa|\mathbf{y})\,\mathrm{d}\theta_\kappa \tag{8.10}$$

where $\theta_i = (\beta_i, \boldsymbol{\mu}_i, \sigma_i^2)$ $(1 \le i \le \kappa)$. Deduce from this representation an optimization scheme that slowly increases κ over iterations and that runs a Gibbs sampler for the integrand of (8.10) at each iteration.

The representation (8.10) is obvious since

$$\left(\int \pi(\mathbf{x},\theta|\mathbf{y})\,\mathrm{d}\theta\right)^{\kappa} = \int \pi(\mathbf{x},\theta|\mathbf{y})\,\mathrm{d}\theta \times \cdots \times \int \pi(\mathbf{x},\theta|\mathbf{y})\,\mathrm{d}\theta$$
$$= \int \pi(\mathbf{x},\theta_1|\mathbf{y})\,\mathrm{d}\theta_1 \times \cdots \times \int \pi(\mathbf{x},\theta_\kappa|\mathbf{y})\,\mathrm{d}\theta_\kappa$$

given that the symbols θ_i within the integrals are dummies.

This is however the basis for the so-called SAME algorithm of Doucet, Godsill and Robert (2001), described in detail in Robert and Casella (2004).