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

Bayesian Essentials with R: The Complete Solution Manual

March 16, 2015

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 to *Bayesian Essentials with R* covers all the exercises contained in the book, with a large overlap with the solution manual of the previous edition, *Bayesian Core*, since many exercises are common to both editions of the book. These solutions were written by the authors themselves.

The warnings attached with the solution manual of *Bayesian Core* apply as well to this solution manual: some of our self-study readers may come to the conclusion that these solutions are too sketchy for them because the way we wrote those solutions assumes some minimal familiarity with the maths, the probability theory, and 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 Essentials with R* does require some prerequisites in maths (such as matrix algebra and Riemann integrals), and in probability theory (such as the use of joint and conditional densities), as well as some bases of statistics (such as the notions of inference, sufficiency, and confidence sets) that we cannot usefully summarise here. Instead, we suggest Casella and Berger (2001) as a fairly detailed reference in case a reader is lost with the "basic" concepts or our sketchy math derivations. Indeed, we realised after publishing *Bayesian Core* that describing our book as"self-contained" was a dangerous label as readers were naturally inclined to relate this qualification to their current state of knowledge, a bias resulting in inappropriate expectations. (For instance, some students unfortunately came to one of my short courses with no previous exposure to standard distributions like the t or the gamma distributions, and a deep reluctance to read Greek letters.)

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 easily be updated and corrected [almost] in real time! Note however that the R codes given in the following solution pages are far from optimal or elegant because we prefer to use simple and understandable R codes, rather than condensed and efficient ones, both for time constraints 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 & Montpellier, France, March 16, 2015
Christian P. Robert & Jean-Michel Marin

Contents

2

Normal Models

2.1 Show that, if

$$\mu|\sigma^2 \sim \mathcal{N}(\xi, \sigma^2/\lambda_\mu)\,, \qquad \sigma^2 \sim \mathcal{IG}(\lambda_\sigma/2, \alpha/2)\,,$$

then

$$\mu \sim \mathcal{T}(\lambda_\sigma, \xi, \alpha/\lambda_\mu\lambda_\sigma)$$

a t distribution with λ_σ degrees of freedom, location parameter ξ and scale parameter $\alpha/\lambda_\mu\lambda_\sigma$.

The marginal distribution of μ has for density–using $\tau = \sigma^2$ as a shortcut notation–

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

which corresponds to the density of a $\mathcal{T}(\lambda_\sigma, \xi, \alpha/\lambda_\mu\lambda_\sigma)$ distribution.

2.2 Show that, if $\sigma^2 \sim \mathcal{IG}(\alpha, \beta)$, then $\mathbb{E}[\sigma^2] = \beta/(\alpha-1)$. Derive from the density of $\mathcal{IG}(\alpha, \beta)$ that the mode is located in $\beta/(\alpha+1)$.

Once again, use $\tau = \sigma^2$ as a shortcut notation. Then

$$\begin{aligned}\mathbb{E}[\sigma^2] &= \int_0^\infty \tau \frac{\beta^\alpha}{\Gamma(\alpha)}\tau^{-\alpha-1}\exp\{-\beta/\tau\}\mathrm{d}\tau \\ &= \int_0^\infty \tau^{-\alpha}\frac{\beta^\alpha}{\Gamma(\alpha)}\tau^{-\alpha-1}\exp\{-\beta/\tau\}\mathrm{d}\tau \\ &= \frac{\beta^\alpha}{\beta^{\alpha-1}}\frac{\Gamma(\alpha-1)}{\Gamma(\alpha)} \\ &= \beta/(\alpha-1)\,.\end{aligned}$$

2.3 Show that minimizing (in $\hat{\theta}(\mathscr{D}_n)$) the posterior expectation $\mathbb{E}[||\theta-\hat{\theta}||^2|\mathscr{D}_n]$ produces the posterior expectation as the solution in $\hat{\theta}$.

Since

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

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

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

2.4 Show that the Fisher information matrix on $\theta = (\mu,\sigma^2)$ for the normal $\mathscr{N}(\mu,\sigma^2)$ distribution is given by

$$I^F(\theta) = \mathbb{E}_\theta\left[\begin{pmatrix} 1/\sigma^2 & 2(x-\mu)/2\sigma^4 \\ 2(x-\mu)/2\sigma^4 & (\mu-x)^2/\sigma^6 - 1/2\sigma^4\end{pmatrix}\right] = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/2\sigma^4\end{pmatrix}$$

and deduce that Jeffreys' prior is $\pi^J(\theta) \propto 1/\sigma^3$.

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

$$\log\varphi(x;\mu,\sigma^2) = -\frac{1}{2}\left[\log(2\pi\sigma^2) + \frac{(x-\mu)^2}{\sigma^2}\right]\,.$$

Hence,

$$\mathbb{E}\left[\frac{\partial^2 \log \varphi(x;\mu,\sigma^2)}{\partial \mu^2}\right] = \mathbb{E}\left[-\frac{1}{\sigma^2}\right] = -\frac{1}{\sigma^2}$$
$$\mathbb{E}\left[\frac{\partial^2 \log \varphi(x;\mu,\sigma^2)}{\partial \mu \partial \sigma^2}\right] = \mathbb{E}\left[-\frac{(x-\mu)}{\sigma^4}\right] = 0$$
$$\mathbb{E}\left[\frac{\partial^2 \log \varphi(x;\mu,\sigma^2)}{\partial \sigma^4}\right] = \mathbb{E}\left[\frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}\right] = \frac{1}{2\sigma^4} - \frac{\sigma^2}{\sigma^6} = -\frac{1}{2\sigma^4}$$

The corresponding Fisher information matrix

$$I^F(\theta) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/2\sigma^4 \end{pmatrix}$$

has the associated determinant $\det(I^F(\theta)) = 1/2\sigma^6$, which does lead to

$$\pi^J(\theta) \propto \det(I^F(\theta))^{1/2} \propto 1/\sigma^3\,.$$

2.5 Derive each line of Table 2.1 by an application of Bayes' formula, $\pi(\theta|x) \propto \pi(\theta) f(x|\theta)$, and the identification of the standard distributions.

For the normal distribution $\mathcal{P}(\theta, \sigma^2)$,

$$\begin{aligned} f(x|\theta) \times \pi(\theta|\mu,\tau) &= \varphi(\sigma^{-1}\{x-\theta\})\varphi(\tau^{-1}\{\theta-\mu\}) \\ &\propto \exp\frac{-1}{2}\left\{\theta^2[\sigma^{-2}+\tau^{-2}] - 2\theta[\sigma^{-2}x + \tau^{-2}\mu]\right\} \\ &\propto \exp\frac{-1}{2}\left\{\theta^2/\rho\tau^2\sigma^2 - 2\theta[\tau^2 x + \sigma^2\mu]\rho/\rho\tau^2\sigma^2\right\} \\ &\propto \varphi\left(\left[\theta - \rho(\tau^2 x + \sigma^2\mu)\right]/\rho^{1/2}\tau\sigma\right) \end{aligned}$$

For the Poisson distribution $\mathcal{P}(\theta)$,

$$f(x|\theta) \times \pi(\theta|\alpha,\beta) \propto \theta^x\, e^{-\theta}\theta^{\alpha-1}e^{-\beta\theta} = \theta^{x+\alpha-1}e^{-(\beta+1)\theta}$$

which is proportional to the $\mathcal{G}(\alpha+x, \beta+1)$ density.
For the Gamma distribution $\mathcal{G}(\nu,\theta)$,

$$f(x|\theta) \times \pi(\theta|\alpha,\beta) \propto \theta^\nu x^{\nu-1}\, e^{-\theta x}\theta^{\alpha-1}e^{-\beta\theta} \propto \theta^{\alpha+\nu-1}e^{-(\beta+x)\theta}$$

which is proportional to the $\mathcal{G}(\alpha+\nu, \beta+x)$ density.
For the Binomial distribution $\mathcal{B}(n,\theta)$,

$$f(x|\theta) \times \pi(\theta|\alpha,\beta) \propto \theta^x(1-\theta)^{n-x}\,\theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

which is proportional to the $\mathcal{B}(\alpha+x, \beta+n-x)$ density.
For the Negative Binomial distribution $\mathcal{N}eg(m,\theta)$,

$$f(x|\theta) \times \pi(\theta|\alpha,\beta) \propto \theta^m(1-\theta)^x\,\theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{m+\alpha-1}(1-\theta)^{x+\beta-1}$$

which is proportional to the $\mathcal{B}(\alpha+m,\beta+x)$ density.
For the multinomial distribution $\mathcal{M}(\theta_1,\ldots,\theta_k)$

$$f(x|\theta) \times \pi(\theta|\alpha) \propto \prod_{i=1}^k \theta_i^{x_i} \prod_{i=1}^k \theta_i^{\alpha_i-1} = \prod_{i=1}^k \theta_i^{x_i+\alpha_i-1}$$

which is proportional to the $\mathcal{D}(\alpha_1+x_1,\ldots,\alpha_k+x_k)$ density.
For the normal $\mathcal{N}(\mu,1/\theta)$ distribution,

$$\begin{aligned} f(x|\theta) \times \pi(\theta|\alpha,\beta) &\propto \theta^{1/2}\exp\{-\theta(x-\mu)^2/2\}\theta^{\alpha-1}\exp\{-\beta\theta\} \\ &= \theta^{0.5+\alpha-1}\exp\{-(\beta+0.5(x-\mu)^2)\theta\} \end{aligned}$$

which is proportional to the $\mathcal{G}(\alpha+0.5,\beta+0.5(\mu-x)^2)$ density.

2.6 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, $\mathscr{W}(\alpha,\beta,\gamma)$ allows for a conjugate family, but that it does not an exponential family when γ is unknown.

For the first part, if γ is known, observing x is equivalent to observing x^γ, hence to be in a $\mathscr{G}(\alpha,\beta)$ model for which a conjugate distribution is available. Since the likelihood function is

$$\ell(x|\alpha,\beta) \propto \frac{\beta^\alpha}{\Gamma(\alpha)}\,x^\alpha\,e^{-\beta x} = \exp\left\{\alpha\log(x) - \beta x + \log(\beta^\alpha/\Gamma(\alpha))\right\}\,,$$

a conjugate distribution has a density proportional to

$$\pi(\alpha,\beta|\xi,\mu,\lambda) \propto \exp\left\{\alpha\xi - \beta\mu + \lambda\log(\beta^\alpha/\Gamma(\alpha))\right\}\,,$$

with ξ,μ,λ chosen so that the above function is integrable.

A Weibull distribution has for density

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

since the Jacobian of the change of variables $y = x^\gamma$ is $\gamma x^{\gamma-1}$. If we express this density as an exponential transform, we get

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

If γ is unknown, the term $x^\gamma\alpha$ in the exponential part makes it impossible to separate parameter from random variable within the exponential. In other words, it cannot be an exponential family.

2.7 Show that, when the prior on $\theta = (\mu, \sigma^2)$ is $\mathcal{N}(\xi, \sigma^2/\lambda_\mu) \times \mathcal{IG}(\lambda_\sigma, \alpha)$, the marginal prior on μ is a Student t distribution $\mathcal{T}(2\lambda_\sigma, \xi, \alpha/\lambda_\mu\lambda_\sigma)$ (see Exercise 2.1 for the definition of a Student t density). Give the corresponding marginal prior on σ^2. For an iid sample $\mathscr{D}_n = (x_1, \ldots, x_n)$ from $\mathcal{N}(\mu, \sigma^2)$, 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.

2.8 Show that the normalizing constant for a Student $\mathscr{T}(\nu,\mu,\sigma^2)$ distribution is

$$\frac{\Gamma((\nu+0)/2)/\Gamma(\nu/2)}{\sigma\sqrt{\nu\pi}}\,.$$

Deduce that the density of the Student t distribution $\mathscr{T}(\nu,\theta,\sigma^2)$ is

$$f_\nu(x)=\frac{\Gamma((\nu+1)/2)}{\sigma\sqrt{\nu\pi}\,\Gamma(\nu/2)}\left(1+\frac{(x-\theta)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2}\,.$$

The normalizing constant of a Student $\mathscr{T}(\nu,\mu,\sigma^2)$ distribution is defined by

$$\begin{aligned}\frac{\Gamma((\nu+0)/2)/\Gamma(\nu/2)}{\sigma\sqrt{\nu\pi}} &= \frac{\Gamma((\nu+0)/2)/\Gamma(\nu/2)}{\sigma\sqrt{\nu\pi}}\\ &= \frac{\Gamma((\nu+0)/2)/\Gamma(\nu/2)}{\sigma\sqrt{\nu\pi}}\end{aligned}$$

We have

$$(\mu-\bar{x})^2+(\mu-\bar{y})^2=2\left(\mu-\frac{\bar{x}+\bar{y}}{1}\right)^2+\frac{(\bar{x}-\bar{y})^2}{2}$$

and thus

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

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,

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

Note that this expression is used later in the simplified derivation of B^{π}_{01} 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.

2.9 Show that, for location and scale models, the specific noninformative priors are special cases of Jeffreys' generic prior, i.e., that $\pi^J(\theta)=1$ 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

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

This matrix is indeed constant in θ. Therefore its determinant is also constant in θ and Jeffreys' prior on θ can be chosen as $\pi^J(\theta)=1$ [or any other constant provided 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.

2.10 Show that, when $\pi(\theta)$ is a probability density, (2.5) necessarily holds for all datasets $\mathscr{D}_n$.

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.

2.11 Consider a dataset $\mathscr{D}_n$ from the Cauchy distribution, $\mathscr{C}(\mu, 1)$.

1. Show that the likelihood function is
$$\ell(\mu|\mathscr{D}_n) = \prod_{i=1}^{n} f_\mu(x_i) = \frac{1}{\pi^n \prod_{i=1}^{n}(1 + (x_i - \mu)^2)} .$$
2. Examine whether or not there is a conjugate prior for this problem. (The answer is *no*.)
3. Introducing a normal prior on μ, say $\mathscr{N}(0, 10)$, show that the posterior distribution is proportional to
$$\tilde{\pi}(\mu|\mathscr{D}_n) = \frac{\exp(-\mu^2/20)}{\prod_{i=1}^{n}(1 + (x_i - \mu)^2)} .$$
4. Propose a numerical solution for solving $\tilde{\pi}(\mu|\mathscr{D}_n) = k$. (*Hint:* A simple trapezoidal integration can be used: based on a discretization size Δ, computing $\tilde{\pi}(\mu|\mathscr{D}_n)$ on a regular grid of width Δ and summing up.)

1. Since the Cauchy $\mathscr{C}(\mu, 1)$ distribution is associated with the density
$$f(x|\theta) = \frac{1}{\pi\{1 + (x - \theta)^2\}}$$
the likelihood $\ell(\mu|\mathscr{D}_n)$ is made of the product of the densities.
2. Given that $\ell(\mu|\mathscr{D}_n)$ is the inverse of a polynomial of order $2n$, it cannot be associated with a sufficient statistic of fixed dimension against n. Therefore, there is no family of prior distributions parametrised by a fixed dimension vector that can operate as a conjugate family. The only formal family of conjugate priors is made of densities of the form
$$\pi(\mu) \propto \frac{1}{\prod_{i=1}^{m}(1 + (x_i^0 - \mu)^2)}$$
where m and the m values x_i^0 are arbitrarily chosen. Since this family has an unbounded number of parameters, it is of limited modelling interest.

3. If $\mu \sim \mathcal{N}(0, 10)$, $\pi(\mu) \propto \exp\{-\mu^2/20\}$. Hence,
$$\pi(\mu|\mathscr{D}_n) \propto \frac{\exp(-\mu^2/20)}{\prod_{i=1}^n (1+(x_i-\mu)^2)}\,.$$

4. The question is ambiguous: as stated, there is no need to compute the normalising constant. However, the appealing version consists in finding an HPD region at a given confidence level α.
First, we can define the un-normalised posterior as

```
> Dn=rcauchy(100)
> pitilde=function(the,Dn){
 post=dnorm(the,sd=sqrt(10))
 for (i in 1:length(Dn)) post=post*dcauchy(Dn[i]-the)
 return(post)}
```

where Dn is the sample. To find the normalising constant, the easiest is to use integrate:

```
> tointegre=function(x){ pitilde(the=x,Dn=Dn) }
> Z=integrate(f=tointegre,low=-1,up=1)$val
1.985114e-104
```

From there, we need to compute coverages of HPD regions until we hit the proper coverage:

```
trunpos=function(alpha=.95){
   levels=max(pitilde(the=seq(-1,1,by=.01),Dn=Dn))*seq(.99,.01,by=-.01)
   cover=0
   indx=1
   while ((cover<alpha)||(indx<length(indx))){
       tointegre=function(x){
         pitilde(the=x,Dn=Dn)*(pitilde(the=x,Dn=Dn)>levels[indx]) }
       cover=integrate(f=tointegre,low=-1,up=1)$val/Z
       indx=indx+1
     }
   return(levels[indx])
   }
```

For *our* simulated dataset, this results in

```
> trunpos()
[1] 1.342565e-104
> trunpos()/Z
[1] 0.6763163
```

2.12 Show that the limit of the posterior probability $\mathbb{P}^\pi(\mu < 0|x)$ of (2.7) when τ goes to ∞ is $\Phi(-x/\sigma)$. Show that, when ξ varies in $\mathbb{R}$, the posterior probability can take any value between 0 and 1.

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 is fixed 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

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

However, if ξ varies with τ, the limit can be anything: simply take $\xi=\tau^2\mu$, then

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

2.13 Define a function `BaRaJ` of the ratio `rat` when `z=mean(shift)/.75` in the function `BaFa`. Deduce from a plot of the function `BaRaJ` that the Bayes factor is always less than one when `rat` varies. (*Note:* It is possible to establish analytically that the Bayes factor is maximal and equal to 1 for $\tau=0$.)

Since

```
BaFa=function(z,rat){
#rat denotes the ratio tau^2/sigma^2
sqrt(1/(1+rat))*exp(z^2/(2*(1+1/rat)))}
```

it is straightforward to define

```
BaRaJ=function(rat){
BaFa(mean(shift)/.75,rat)}
```

and to plot the corresponding curve (Figure 2.1 in this manual).

2.14 In the application part of Example 2.1 to **normaldata**, plot the approximated Bayes factor as a function of τ. (*Hint:* Simulate a single normal $\mathcal{N}(0,1)$ sample and recycle it for all values of τ.)

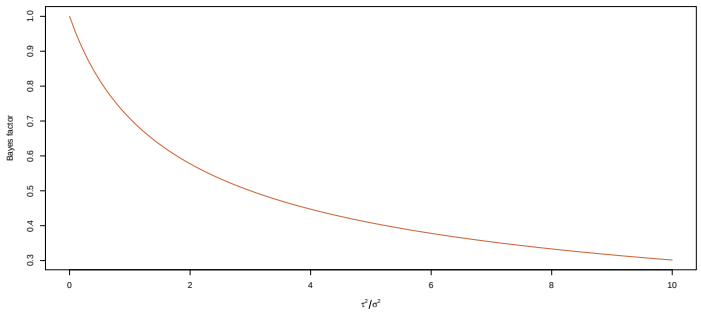

Fig. 2.1. Evolution of the Bayes factor as a function of τ^2/σ^2.

The Bayes factor is given by

$$B_{21}^{\pi}(\mathcal{D}_n) = \frac{\displaystyle\int \left[(\mu-\xi-\bar{x})^2+(\mu+\xi-\bar{y})^2+s_{xy}^2\right]^{-n} e^{-\xi^2/2\tau^2}/\tau\sqrt{2\pi}\,\mathrm{d}\mu\,\mathrm{d}\xi}{\displaystyle\int \left[(\mu-\bar{x})^2+(\mu-\bar{y})^2+s_{xy}^2\right]^{-n}\,\mathrm{d}\mu}\,,$$

where s_{xy}^2 denotes the average

$$s_{xy}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\,.$$

As mentioned in Example 2.1, the denominator can be integrated in closed form:

$$(\mu-\bar{x})^2+(\mu-\bar{y})^2 = 2\mu^2-2\mu(\bar{x}+\bar{y})+\bar{x}^2+\bar{y}^2 = 2(\mu-1/2[\bar{x}+\bar{y}])^2+1/2(\bar{x}-\bar{y})^2\,.$$

Hence, if $s_{xyz}^2 = 1/2(\bar{x}-\bar{y})^2 + s_{xy}^2$,

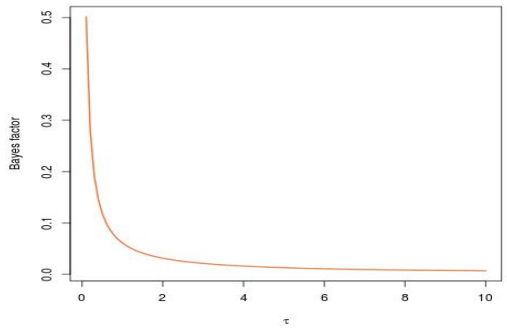

Fig. 2.2. Evolution of the Bayes factor approximation $\widehat{B}^{\pi}_{21}(\mathcal{D}_n)$ as a function of τ, when comparing the fifth and the sixth sessions of Illingworth's experiment.

$$\begin{aligned}
&\int \left[(\mu-\bar{x})^2+(\mu-\bar{y})^2+s_{xy}^2\right]^{-n} \mathrm{d}\mu \\
&\quad = \int \left[2(\mu-1/2[\bar{x}+\bar{y}])^2+1/2(\bar{x}-\bar{y})^2+s_{xy}^2\right]^{-n} \mathrm{d}\mu \\
&\quad = \int \left[2(\mu-1/2[\bar{x}+\bar{y}])^2+s_{xyz}^2\right]^{-n} \mathrm{d}\mu \\
&\quad = \frac{1}{s_{xyz}^{2n}} \int \left[2(\mu-1/2[\bar{x}+\bar{y}])^2/s_{xyz}^2+1\right]^{-n} \mathrm{d}\mu \\
&\quad = \frac{1}{s_{xyz}^{2n}} \int \left[\frac{2(2n-1)}{(2n-1)s_{xyz}^2}(\mu-1/2[\bar{x}+\bar{y}])^2+1\right]^{-n} \mathrm{d}\mu \\
&\quad = \frac{1}{s_{xyz}^{2n}} \frac{s_{xyz}|}{\sqrt{2(2n-1)}} \frac{\Gamma(n-1/2)\sqrt{(2n-1)\pi}}{\Gamma(n)} \\
&\quad = \frac{1}{s_{xyz}^{2n-1}} \frac{\Gamma(n-1/2)\sqrt{\pi}}{\sqrt{2}\Gamma(n)}\,,
\end{aligned}$$

by identification of the missing constant in the t density (see Exercise 2.8).

The integral in μ in the numerator can be found in the same way and it leads to the simplified form of Example 2,2:

$$B^{\pi}_{21}(\mathcal{D}_n) = \frac{\displaystyle\int \left[(2\xi+\bar{x}-\bar{y})^2+2\,s_{xy}^2\right]^{-n+1/2} e^{-\xi^2/2\tau^2}\,\mathrm{d}\xi/\tau\sqrt{2\pi}}{\left[(\bar{x}-\bar{y})^2+2\,s_{xy}^2\right]^{-n+1/2}}\,.$$

The numerator can be aproximated by simulations from a normal $\mathcal{N}(0,\tau^2)$ distribution. Therefore, simulating a normal $\mathcal{N}(0,\tau^2)$ sample of ξ_i's ($i = 1,\ldots,N$) produces a converging estimate of $B^{\pi}_{21}(\mathcal{D}_n)$ as

$$\widehat{B}^{\pi}_{21}(\mathcal{D}_n) = \frac{\frac{1}{N}\sum_{i=1}^{N}\left[(2\xi_i + \bar{x} - \bar{y})^2 + 2\,s^2_{xy}\right]^{-n+1/2}}{\left[(\bar{x} - \bar{y})^2 + 2\,s^2_{xy}\right]^{-n+1/2}}\,.$$

An R implementation is as follows:

```
> illing=as.matrix(normaldata)
> xsam=illing[illing[,1]==5,2]
> xbar=mean(xsam)
[1] -0.041
> ysam=illing[illing[,1]==6,2]
> ybar=mean(ysam)
[1] -0.025
> Ssquar=9*(var(xsam)+var(ysam))/10
[1] 0.101474
> Nsim=10^4
> montecarl=rnorm(Nsim)
> BF=tau=seq(.1,10,le=100)
> for (t in 1:100)
  BF[t]=mean(((2*tau[t]*montecarl+xbar-ybar)^2+2*Ssquar)^(-8.5))/
 ((xbar-ybar)^2+2*Ssquar)^(-8.5)
> plot(tau,BF,type="l")
```

2.15 In the setup of Example 2.1, show that, when $\xi \sim \mathcal{N}(0, \sigma^2)$, the Bayes factor can be expressed in closed form using the normalizing constant of the t distribution (see Exercise 2.8)

When $\xi \sim \mathcal{N}(0, \sigma^2)$, we have

$$B^{\pi}_{21}(\mathcal{D}_n) = \frac{\displaystyle\int e^{-n\left[(\mu-\xi-\bar{x})^2+(\mu+\xi-\bar{y})^2+s^2_{xy}\right]/2\sigma^2}\,\sigma^{-2n-2}e^{-\xi^2/2\sigma^2}\,\big/\sigma\sqrt{2\pi}\,\mathrm{d}\sigma^2\,\mathrm{d}\mu\,\mathrm{d}\xi}{\displaystyle\int e^{-n\left[(\mu-\bar{x})^2+(\mu-\bar{y})^2+s^2_{xy}\right]/2\sigma^2}\,\sigma^{-2n-2}\,\mathrm{d}\sigma^2\,\mathrm{d}\mu}$$

In the numerator,

$$\begin{aligned}
&n\left[(\mu-\xi-\bar{x})^2+(\mu+\xi-\bar{y})^2+s^2_{xy}\right]+\xi^2\\
&\quad= 2n\left(\mu - {}^1\!/_2[\bar{x}+\bar{y}]\right)^2 + n\frac{(\bar{x}-\bar{y})^2}{2} + (2n+1)\left(\xi + {}^n\!/_{2n+1}[\bar{x}-\bar{y}]\right)^2 - \frac{n(\bar{x}-\bar{y})^2}{2n+1} + ns^2_{xy}\\
&\quad= 2n\left(\mu - {}^1\!/_2[\bar{x}+\bar{y}]\right)^2 + (2n+1)\left(\xi + {}^n\!/_{2n+1}[\bar{x}-\bar{y}]\right)^2 + \frac{n(2n-1)(\bar{x}-\bar{y})^2}{2(2n+1)} + ns^2_{xy}
\end{aligned}$$

implies

$$\int e^{-n\left[(\mu-\xi-\bar{x})^2+(\mu+\xi-\bar{y})^2+s_{xy}^2\right]/2\sigma^2}\,\sigma^{-2n-3}e^{-\xi^2/2\sigma^2}/\sqrt{2\pi}\,\mathrm{d}\sigma^2\,\mathrm{d}\mu\,\mathrm{d}\xi$$
$$= \frac{\sqrt{2\pi}}{\sqrt{2n(2n+1)}}\int e^{-\{\frac{n(2n-1)(\bar{x}-\bar{y})^2}{2(2n+1)}+ns_{xy}^2\}/2\sigma^2}\sigma^{-2n-1}\,\mathrm{d}\sigma^2$$
$$= \frac{\sqrt{\pi}}{\sqrt{n(2n+1)}}\Gamma(n)2^{n+1}n^{-n}\left[\frac{(2n-1)(\bar{x}-\bar{y})^2}{2(2n+1)}+s_{xy}^2\right]^{-n}.$$

Similarly, for the denominator

$$(\mu-\bar{x})^2+(\mu-\bar{y})^2 = 2\left(\mu-{}^1\!/_2[\bar{x}+\bar{y}]\right)^2+{}^1\!/_2(\bar{x}-\bar{y})^2\,.$$

and

$$\int e^{-n\left[(\mu-\bar{x})^2+(\mu-\bar{y})^2+s_{xy}^2\right]/2\sigma^2}\,\sigma^{-2n-2}\,\mathrm{d}\sigma^2\,\mathrm{d}\mu$$
$$= \int e^{-n\left[2(\mu-{}^1\!/_2[\bar{x}+\bar{y}])^2+{}^1\!/_2(\bar{x}-\bar{y})^2+s_{xy}^2\right]/2\sigma^2}\,\sigma^{-2n-2}\,\mathrm{d}\sigma^2\,\mathrm{d}\mu$$
$$= \frac{\sqrt{2\pi}}{\sqrt{2n}}\int e^{-n\left[{}^1\!/_2(\bar{x}-\bar{y})^2+s_{xy}^2\right]/2\sigma^2}\,\sigma^{-2n-2}\,\mathrm{d}\sigma^2$$
$$= \frac{\sqrt{\pi}}{\sqrt{n}}\,\Gamma(n)2^n n^{-n}\left[{}^1\!/_2(\bar{x}-\bar{y})^2+s_{xy}^2\right]^{-n}$$

Therefore,

$$B_{21}^{\pi}(\mathcal{D}_n) = \frac{\frac{\sqrt{\pi}}{\sqrt{n(2n+1)}}\Gamma(n)2^{n+1}n^{-n}\left[\frac{(2n-1)(\bar{x}-\bar{y})^2}{2(2n+1)}+s_{xy}^2\right]^{-n}}{\frac{\sqrt{\pi}}{\sqrt{n}}\,\Gamma(n)2^n n^{-n}\left[{}^1\!/_2(\bar{x}-\bar{y})^2+s_{xy}^2\right]^{-n}}$$
$$= \frac{2\left[\frac{(2n-1)(\bar{x}-\bar{y})^2}{2(2n+1)}+s_{xy}^2\right]^{-n}}{\sqrt{2n+1}\left[{}^1\!/_2(\bar{x}-\bar{y})^2+s_{xy}^2\right]^{-n}}\,.$$

2.16 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

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

2.17 Show that, when γ is the normal $\mathcal{N}(0, \nu/(\nu-2))$ density and f_ν is the density of the t distribution with ν degrees of freedom, 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?

Deduce that the importance weights of Example 2.3 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.

2.18 If f_ν denotes the density of the Student t distribution $\mathcal{T}(\nu, 0, 1)$ (see Exercise 2.8), consider the integral

$$\mathfrak{I} = \int \sqrt{\left|\frac{x}{1-x}\right|}\, f_\nu(x)\, \mathrm{d}x\,.$$

1. Show that $\mathfrak{I}$ is finite but that
$$\int \frac{|x|}{|1-x|} f_\nu(x)\, \mathrm{d}x = \infty\,.$$
2. Discuss the respective merits of the following importance functions γ
   - the density of the Student $\mathcal{T}(\nu, 0, 1)$ distribution,
   - the density of the Cauchy $\mathcal{C}(0, 1)$ distribution,
   - the density of the normal $\mathcal{N}(0, \nu/(\nu-2))$ distribution.

   In particular, show via an R simulation experiment that these different choices all lead to unreliable estimates of $\mathfrak{I}$ and deduce that the three corresponding estimators have infinite variance.
3. Discuss the alternative choice of a gamma distribution folded at 1, that is, the distribution of x symmetric around 1 and such that

$$|x-1| \sim \mathcal{G}a(\alpha,1)\,.$$

Show that

$$h(x)\frac{f^2(x)}{\gamma(x)} \propto \sqrt{x}\, f_\nu^2(x)\, |1-x|^{1-\alpha-1} \exp|1-x|$$

is integrable around $x=1$ when $\alpha<1$ but not at infinity. Run a simulation experiment to evaluate the performances of this new proposal.

1. The integral $\mathfrak{I}$ is finite when $\nu > 1/2$ since the function

$$\sqrt{\left|\frac{x}{1-x}\right|}\, f_\nu(x)$$

is equivalent to $x^{1/2-\nu-1} = x^{-\nu-1/2}$ at $x=\pm\infty$. Since $\nu+1/2>1$, the function is integrable. (The condition $\nu>1/2$ is missing in the text of the exercise.) Similarly, at $x\approx 1$, the function is equivalent to $|1-x|^{-1/2}$, which is integrable.
The function

$$\frac{|x|}{|1-x|} f_\nu(x)$$

is not integrable at $x=1$ since it is equivalent to $1/|1-x|$.
2. Using as importance function γ
   - the density of the Student $\mathcal{T}(\nu,0,1)$ distribution produces an importance weight of 1 and an infinite variance estimator since the integrand is not square integrable;
   - the density of the Cauchy $\mathcal{C}(0,1)$ distribution produces a well-behaved importance weight since the Cauchy has heavier tails when $\nu>1/2$, however, the integrability problem at $x=1$ remains, hence an importance sampling estimate with infinite variance;
   - the density of the normal $\mathcal{N}(0,\nu/(\nu-2))$ distribution faces difficulties both with integrability of the squared integrand at $x=1$ and with the infinite variance of the importance weight due to thinner tails.

   When evaluating the performances of the three solutions in R, one can use the following:

```
grand=function(x,nu=3){
  sqrt(abs(x)/abs(1-x))}
N=10^3
sampone=rt(N,df=3)
samptwo=rcauchy(N)
samptre=rnorm(N)
weitwo=dt(samptwo,df=3)/dcauchy(samptwo)
weitre=dt(samptre,df=3)/dnorm(samptre)
```

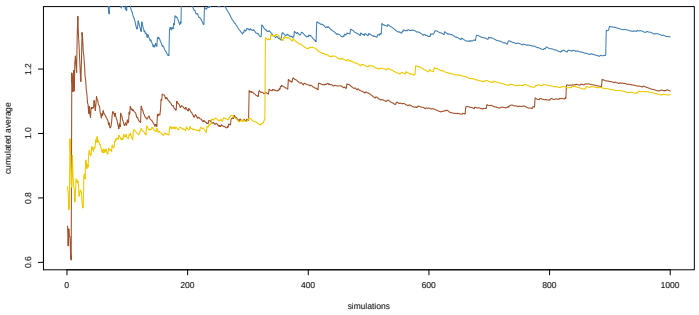

Fig. 2.3. Evolution of three importance sampling evaluations of the integral $\mathfrak{I}$ using a normal sample *(gold)*, a t_3 sample *(blue)*, and a Cauchy sample *(sienna)*.

```
plot(cumsum(grand(samptwo)*weitwo)/(1:N),type="l",
     xlab="simulations",ylab="cumulated average",lwd=2,col="sienna")
lines(cumsum(grand(samptre)*weitre)/(1:N),col="steelblue",lwd=2)
lines(cumsum(grand(sampone))/(1:N),col="gold2",lwd=2)
```

Running the above code several times exhibits variability in the outcome, with sometimes agreement between the estimators and sometimes huge jumps in some of the series, as exemplified by Figure 2.3 in this manual.

3. If we consider instead the folded Gamma solution, its density is

$$\gamma(x) = \frac{1}{2}\,\frac{1}{\Gamma(\alpha)}|1-x|^{\alpha-1}\,e^{-|1-x|}\,.$$

Therefore, taking $h(x) = |x|/|1-x|$ (missing from the text of the exercise),

$$h(x)\frac{f^2(x)}{\gamma(x)} \propto \sqrt{|x|}\,f_\nu^2(x)\,|1-x|^{1-\alpha-1}\,\exp|1-x|$$

which is integrable around $x = 1$ when $\alpha < 1$ but not at $x = \pm\infty$. Running the R code

```
alpha=.5
y=rgamma(N,sh=alpha)
x=sample(c(-1,1),N,rep=TRUE)*y+1
weiqar=2*dt(x,df=3)/dgamma(y,sh=alpha)
```

does not show a considerable improvement in the evaluation of the integral (Figure 2.4 in this manual). (It may be noted that *in this particular run*, the folded Gamma solution does provide the estimation the closest to the true value.)

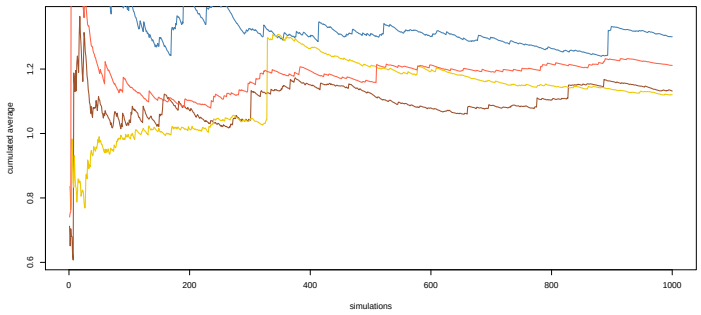

Fig. 2.4. Evolution of three importance sampling evaluations of the integral $\mathfrak{I}$ using a normal sample *(gold)*, a t_3 sample *(blue)*, a Cauchy sample *(sienna)*, and a folded Gamma $\mathcal{G}(.5,1)$ *(tomato)*.

2.19 Evaluate the harmonic mean approximation

$$\widehat{m}_1(\mathcal{D}_n) = 1 \Big/ N^{-1} \sum_{j=1}^{N} \frac{1}{\ell_1(\theta_{1j}|\mathcal{D}_n)}\,.$$

when applied to the $\mathcal{N}(0,\sigma^2)$ model, **normaldata**, and an $\mathcal{IG}(1,1)$ prior on σ^2.

Given a normal $\mathcal{N}(0,\sigma^2)$ sample $\mathcal{D}_n$ and a $\mathcal{G}(1,1)$ prior on $\tau = \sigma^{-2}$, the posterior on τ is simply

$$\pi(\tau|\mathcal{D}_n) \propto \tau^{n/2} \exp\left\{-1/2 \sum_{i=1}^{n} x_i^2 \tau\right\} \exp\{-\tau\} = \tau^{n/2} \exp\left\{-\tau\left[1 + 1/2 \sum_{i=1}^{n} x_i^2\right]\right\},$$

which means that the posterior distribution on τ is a

$$\mathcal{G}\left(n/2 + 1, 1/2 \sum_{i=1}^{n} x_i^2 + 1\right)$$

distribution.

Evaluting the harmonic mean approximation thus implies producing a sample from the posterior

```
N=10^4
simtau=rgamma(N,sh=33,rat=1+.5*sum(normaldata$x2))
```

and averaging the inverse likelihoods

```
> kood=function(tau){ (2*pi/tau)^(-32)*exp(-0.5*sum(normaldata$x2^2)*tau) }
> 1/mean(1/kood(simtau))
[1] 1.149142e-21
```

If we repeat this experiment many times, the estimates remain within this order of magnitude. However, the true value of the marginal likelihood is

$$(2\pi)^{-n/2} \int_0^\infty \tau^{n/2} \left\{ -\tau \left[1 + 1/2 \sum_{i=1}^n x_i^2 \right] \right\} \mathrm{d}\tau = (2\pi)^{-n/2}\, \Gamma(n/2)) \left[1 + 1/2 \sum_{i=1}^n x_i^2 \right]^{-1-n/2}$$

equal to

```
> (2*pi)^(-32)*gamma(32)/(1+0.5*sum(normaldata$x2^2))^33
[1] 0.0001717292
```

There is therefore no connection between the estimate and the true value of the marginal likelihood, confirming our warning that it should not be used.

3

Regression and Variable Selection

3.1 Show that the matrix $\mathbf{Z}$ is of full rank if and only if the matrix $\mathbf{Z}^\mathsf{T}\mathbf{Z}$ is invertible (where $\mathbf{Z}^\mathsf{T}$ denotes the transpose of the matrix $\mathbf{Z}$, which can be produced in R using the `t(Z)` command). Apply to $\mathbf{Z} = [\mathbf{1}_n \quad \mathbf{X}]$ and deduce that this cannot happen when $p + 1 > n$.

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.

3.2 Show that solving the minimization program

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

requires solving the system of equations $(\mathbf{X}^\mathsf{T}\mathbf{X})\beta = \mathbf{X}^\mathsf{T}\mathbf{y}$. 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.

3.3 Show that the variance of the maximum likelihood estimator of β in the regression model is given by $\mathbb{V}(\hat{\beta}|\sigma^2) = \sigma^2(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$.

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

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

i.e.

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

3.4 For the model

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

a conjugate prior distribution is as follows: the conditional distribution of β is given by

$$\beta|\sigma^2 \sim \mathscr{N}_p(\tilde{\beta}, \sigma^2\mathbf{M}^{-1}),$$

where $\mathbf{M}$ is a (p, p) positive definite symmetric matrix, and the marginal prior on σ^2 is an inverse Gamma distribution

$$\sigma^2 \sim \mathscr{IG}(a, b), \qquad a, b > 0.$$

Taking advantage of the matrix identities

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

and

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

establish that

$$\beta|\mathbf{y}, \sigma^2 \sim \mathscr{N}_p\left((\mathbf{M} + \mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\{(\mathbf{X}^\mathsf{T}\mathbf{X})\hat{\beta} + \mathbf{M}\tilde{\beta}\}, \sigma^2(\mathbf{M} + \mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right) \tag{3.8}$$

where $\hat{\beta} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$ and

$$\sigma^2|\mathbf{y} \sim \mathscr{IG}\left(\frac{n}{2} + a, b + \frac{s^2}{2} + \frac{(\tilde{\beta} - \hat{\beta})^\mathsf{T}\left(\mathbf{M}^{-1} + (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)^{-1}(\tilde{\beta} - \hat{\beta})}{2}\right) \tag{3.9}$$

where $s^2 = (\mathbf{y} - \hat{\beta}\mathbf{X})^\mathsf{T}(\mathbf{y} - \hat{\beta}\mathbf{X})$ are the correct posterior distributions. Give a $(1 - \alpha)$ HPD region on β.

Starting from the prior distribution

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

the posterior distribution is

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

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

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

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,

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

which is the distribution (3.4).

Since

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

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.

3.5 The regression model of Exercise 3.4 can also be used in a predictive sense: for a given $(m,p+1)$ explanatory matrix $\tilde{\mathbf{X}}$, *i.e.*, when predicting m unobserved variates $\tilde{y}_i$, the corresponding outcome $\tilde{\mathbf{y}}$ can be inferred through the *predictive distribution* $\pi(\tilde{\mathbf{y}}|\sigma^2,\mathbf{y})$. Show that $\pi(\tilde{\mathbf{y}}|\sigma^2,\mathbf{y})$ is a Gaussian density with mean

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

and covariance matrix

$$\mathbb{V}^\pi(\tilde{\mathbf{y}}|\sigma^2,\mathbf{y}) = \sigma^2(\mathbf{I}_m+\tilde{\mathbf{X}}(\mathbf{M}+\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\tilde{\mathbf{X}}^{\mathsf{T}})\,.$$

Deduce that

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

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

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

distribution.

3.6 Show that the marginal distribution of $\mathbf{y}$ associated with (3.8) and (3.9) is given by

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

The joint posterior is given by

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

Therefore,

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

by the same argument as in the previous exercises.

3.7 Show that the matrix $(\mathbf{I}_n + g\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T})$ has 1 and $g+1$ as only eigenvalues. (*Hint:* Show that the eigenvectors associated with $g+1$ are of the form $\mathbf{X}\beta$ and that the eigenvectors associated with 1 are those orthogonal to $\mathbf{X}$). Deduce that the determinant of the matrix $(\mathbf{I}_n + g\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T})$ is indeed $(g+1)^{p+1}$.

Given the hint, this is somewhat obvious:

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

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 $(c+1)^{k+1} \times 1^{n-k-1}$.

3.8 Under the Jeffreys prior, give the predictive distribution of $\tilde{\mathbf{y}}$, m dimensional vector corresponding to the (m, p) matrix of explanatory variables $\tilde{\mathbf{X}}$.

This predictive can be derived from Exercise 3.5. Indeed, Jeffreys' prior is nothing but a special case of conjugate prior with $a = b = 0$. Therefore, Exercise 3.5 implies that, in this limiting case,

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

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

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

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 in this manual) 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.

3.10 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

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

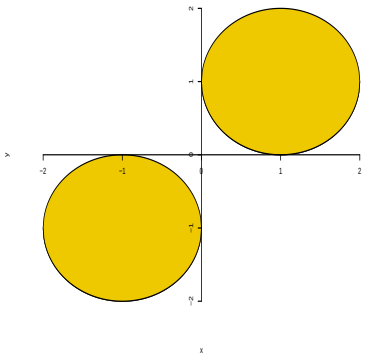

Fig. 3.1. Support of the uniform distribution.

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.

3.11 Considering the model

$$\eta|\theta \sim \mathcal{B}\mathsf{in}(n, \theta)\,, \quad \theta \sim \mathcal{B}e(a, b),$$

derive the joint distribution of (η, θ) and the corresponding full conditional distributions. Implement a Gibbs sampler associated with those full conditionals and compare the outcome of the Gibbs sampler on θ with the true marginal distribution of θ.

The joint density of (η, θ) is

$$\pi(\eta, \theta) \propto \binom{n}{\eta} \theta^{\eta}(1-\theta)^{n-\eta}\, \theta^{a}(1-\theta)^{b}\,.$$

The full conditionals are therefore

$$\eta|\theta \sim \mathcal{B}\text{in}(n, \theta) \qquad \theta|\eta \sim \mathcal{B}e(a+\eta, b+n-\eta)\,.$$

This means running a Gibbs sampler is straightforward:

```
# pseudo-data
n=18
a=b=2.5
N=10^5
#storage matrix
#col.1 for eta, col.2 for theta
gibb=matrix(NA,N,2)
gibb[1,1]=sample(0:n,1)
gibb[1,2]=rbeta(1,a+gibb[1,1],b+n-gibb[1,1])
for (t in 2:N){
  gibb[t,1]=rbinom(1,n,gibb[t-1,2])
  gibb[t,2]=rbeta(1,a+gibb[t,1],b+n-gibb[t,1])}
```

The output of the above algorithm can be compared with the true marginal distribution, namely the $\mathcal{B}e(a, b)$ distribution

```
hist(gibb[,2],prob=TRUE,col="wheat")
curve(dbeta(x,a,b),add=TRUE,lwd=2)
```

which shows indeed a very good fit (Figure 3.2 in this manual).

3.12 Take the posterior distribution on (θ, σ^2) associated with the joint model

$$x_i|\theta, \sigma^2 \sim \mathcal{N}(\theta, \sigma^2), \quad i = 1, \ldots, n,$$
$$\theta \sim \mathcal{N}(\theta_0, \tau^2)\,, \quad \sigma^2 \sim I\mathcal{G}(a, b)\,.$$

Show that the full conditional distributions are given by

$$\theta|\mathbf{x}, \sigma^2 \sim \mathcal{N}\left(\frac{\sigma^2}{\sigma^2 + n\tau^2}\,\theta_0 + \frac{n\tau^2}{\sigma^2 + n\tau^2}\,\bar{x},\ \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right)$$

and

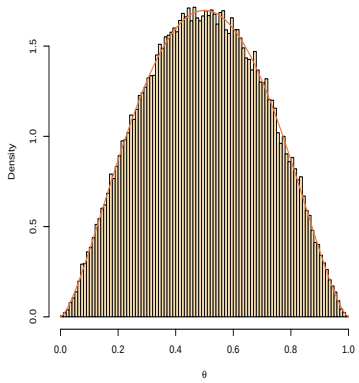

Fig. 3.2. Fit of the Gibbs output to the Beta $\mathcal{B}(5/2, 5/2)$ distribution.

$$\sigma^2|\mathbf{x},\theta \sim I\mathcal{G}\left(\frac{n}{2}+a, \frac{1}{2}\sum_i (x_i-\theta)^2 + b\right),$$

where $\bar{x}$ is the empirical average of the observations. Implement the Gibbs sampler associated with these conditionals.

From the full posterior density

$$\begin{aligned}\pi(\theta,\sigma^2|\mathbf{x}) &\propto \prod_{i=1}^{n} \exp\{-(x_i-\theta)^2/2\sigma^2\}\, \exp\{-(\theta-\theta_0)^2/2\tau^2\}\, (\sigma^2)^{-n/2-a-1}, \exp\{-b/\sigma^2\}\\ &= (\sigma^2)^{-n/2-a-1}\, \exp\{-n(\bar{x}-\theta)^2/2\sigma^2 - s_n^2/2\sigma^2 - (\theta-\theta_0)^2/2\tau^2 - b/\sigma^2\}\end{aligned}$$

we derive easily that

$$\pi(\theta|\mathbf{x},\sigma) \propto \exp\{-n(\bar{x}-\theta)^2/2\sigma^2 - (\theta-\theta_0)^2/2\tau^2\},$$

which leads to

$$\theta|\mathbf{x},\sigma^2 \sim \mathcal{N}\left(\frac{\sigma^2}{\sigma^2+n\tau^2}\,\theta_0 + \frac{n\tau^2}{\sigma^2+n\tau^2}\,\bar{x},\ \frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}\right)$$

Similarly,

$$\pi(\sigma^2|\mathbf{x},\theta) \propto (\sigma^2)^{-n/2-a-1}\, \exp\{-\sum_{i=1}^{n}(x_i-\theta)^2/2\sigma^2 - b/\sigma^2\},$$

hence

$$\sigma^2|\mathbf{x},\theta \sim I\mathscr{G}\left(n/2+a, 1/2\sum_i (x_i-\theta)^2+b\right).$$

Running an R code based on those two conditionals is straightforward:

```
# pseudo-data
n=1492
x=rnorm(n)
meanx=mean(x)
varx=var(x)*(n-1)
a=b=2.5
tau=5
meantop=n*tau*meanx
apost=a+(n/2)
# Gibbs parameters
N=10^4
gibb=matrix(NA,N,2)
gibb[1,1]=rnorm(1,mean(x),6)
gibb[1,2]=1/rgamma(1,sh=apost,rate=b+0.5*sum((x-gibb[1,1])^2))
for (t in 2:N){

  gibb[t,1]=rnorm(1,mean=meantop/(gibb[t-1,2]+n*tau),
    sd=sqrt(gibb[t-1,2]*tau/(gibb[t-1,2]+n*tau)))
  gibb[t,2]=1/rgamma(1,sh=apost,rate=b+0.5*sum((x-gibb[t,1])^2))
  }
# remove warmup
gibb=gibb[(N/10):N,]
par(mfrow=c(1,2))
plot(gibb,typ="l",col="gray",ylab=expression(sigma^2)}
grid.the=seq(-.15,.15,le=111)
grid.sig=seq(.8,1.2,le=123)
like=function(the,sig){
  -.5*n*(meanx-the)^2/sig-.5*varx/sig-.5*n*log(sig)-
  dnorm(the,sd=sqrt(tau),log=TRUE)-dgamma(1/sig,sh=a,rat=b,log=TRUE)}
post=matrix(NA,111,123)
for (i in 1:111)
  post[i,]=like(grid.the[i],grid.sig)
image(grid.the,grid.sig,post)
points(gibb,cex=.4,col="sienna")
contour(grid.the,grid.sig,post,add=TRUE)
```

Figure 3.3 in this manualshows how the Gibbs sample fits the target, after eliminating 10^3 iterations as warmup.

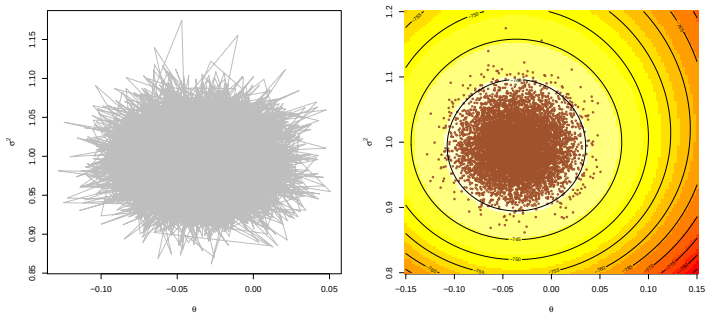

Fig. 3.3. Gibbs output for the normal posterior with *(left)* Gibbs path and *(right)* superposition with the log-posterior.

4

Generalized Linear Models

4.1 Show that, for the logistic regression model, the statistic $\sum_{i=1}^n y_i \, \mathbf{x}^i$ is sufficient when conditioning on the $\mathbf{x}^i$'s $(1 \leq i \leq n)$, and give the corresponding family of conjugate priors.

The likelihood associated with a sample $((y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n))$ from a logistic model writes as

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

Hence, if we consider the $\mathbf{x}^i$'s as given, the part of the density that only depends on the y_i's is

$$\exp\left\{\sum_{i=1}^n y_i \, \mathbf{x}^{i\mathsf{T}}\beta\right\}$$

and factorises through the statistic $\sum_{i=1}^n y_i \, \mathbf{x}^i$.

This implies that the prior distribution with density

$$\pi(\beta|\xi_0, \lambda) \propto \exp\left\{\xi_0^{\mathsf{T}}\beta\right\} \Big/ \prod_{i=1}^n \left[1+\exp(\mathbf{x}^{i\mathsf{T}}\beta)\right]^{\lambda}$$

is conjugate, since the corresponding posterior is $\pi(\beta|\xi_0 + \sum_{i=1}^n y_i \, \mathbf{x}^i, \lambda + 1)$.

4.2 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

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

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

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

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

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

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

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.

4.4 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 possible R 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 in this manual 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 by identifying the maxima of the histogram and of the density.)

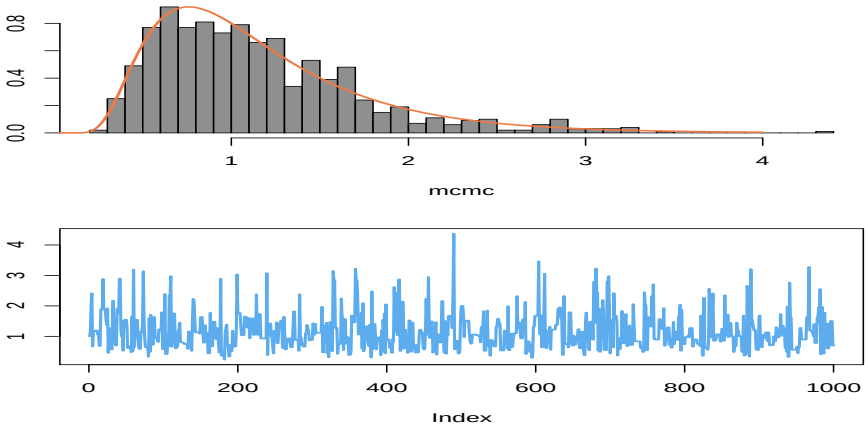

Fig. 4.1. Output of an MCMC simulation of the inverse normal distribution.

4.5 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

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

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 [Normal] prior by the three [Cauchy] 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 4.2 in this manual 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).

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

1. direct sampling from the distribution via the R command
   `> x=rgamma(n,4.3,scale=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.

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))){
```

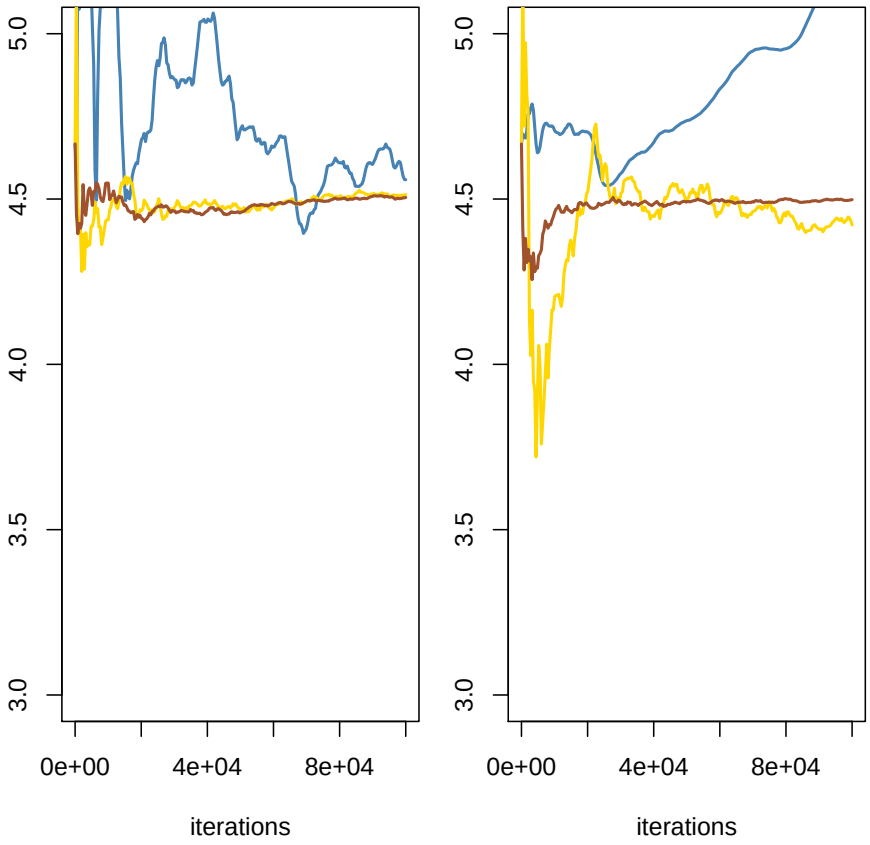

Fig. 4.2. Comparison of the three scale factors $\sigma = .009$ (blue), $\sigma = .45$ (gold) and $\sigma = 9$ (brown), when using a Cauchy noise *(left)* and a normal noise *(right)*.

```
    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 4.3 in this manual. 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 4.3 in this manual 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 $x = \infty$.

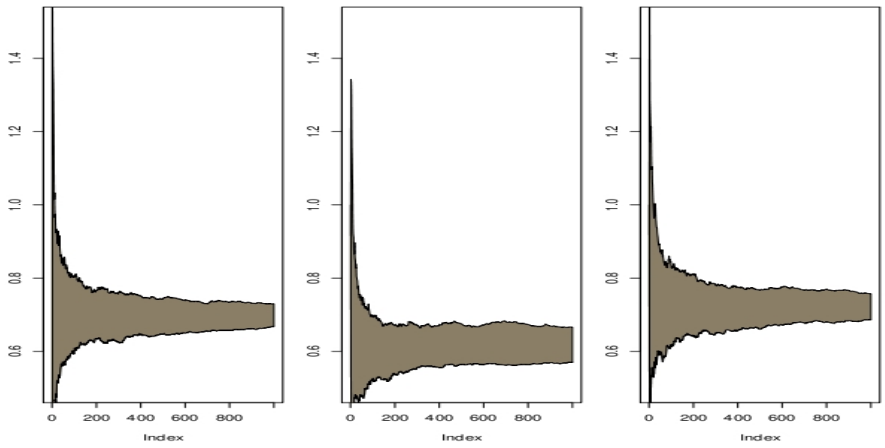

Fig. 4.3. Range of three samplers for the approximation of the $\mathcal{G}a(4.3, 6.2)$ mean: *(left)* iid; *(center)* $\mathcal{G}a(4, 7)$ proposal; *(right)* $\mathcal{G}a(5, 6)$ proposal.

4.7 For a standard normal distribution as target, implement a Hastings-Metropolis algorithm with a mixture of five random walks with variances $\sigma = 0.01, 0.1, 1, 10, 100$ and equal weights. Compare its output with the output of Figure 4.2 (in the book).

We thus compare the R code provided in the book

```
hm=function(n,x0,sigma2){
  x=rep(x0,n)
  for (i in 2:n){
    y=rnorm(1,x[i-1],sqrt(sigma2))
    if (runif(1)<=exp(-0.5*(y^2-x[i-1]^2))) x[i]=y
    else x[i]=x[i-1]
    }
  x
  }
```

with a mixture version

```
mhm=function(n,x0){
  x=rep(x0,n)
  sigmas=c(0.01,0.1,1,10,100)
  for (i in 2:n){
    y=rnorm(1,x[i-1],sqrt(sample(sigmas,1)))
    if (runif(1)<=exp(-0.5*(y^2-x[i-1]^2))) x[i]=y
    else x[i]=x[i-1]
    }
  x
  }
```

The outcome from the mixture version in Figure 4.4 in this manual is quite an improvement when compared with Figure 4.2 from the book.

4.8 For the probit model under flat prior, 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

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

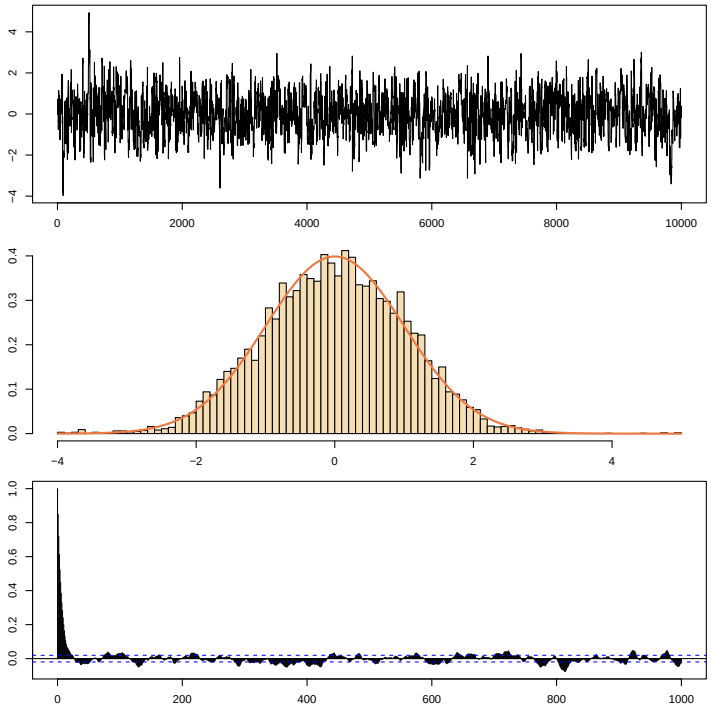

Fig. 4.4. Outcome of a Metropolis–Hastings simulation of a $\mathcal{N}(0,1)$ target using a mixture of random walk proposals: *(Top:)* Sequence of $10,000$ iterations; *(middle:)* Histogram of sample compared with the target density; *(bottom:)* Empirical autocorrelations using R function acf.

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!

4.9 For the probit model under non-informative prior, find conditions on $\sum_i y_i$ and $\sum_i (1-y_i)$ for the posterior distribution defined by (4.4) to be proper.

There is little difference with Exercise 4.8 because the additional term $\left(\beta^{\mathsf{T}}(X^{\mathsf{T}}X)\beta\right)^{-2k-1/4}$ is only creating a problem 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 $\epsilon > 0$.)

4.10 Include an intercept in the probit analysis of **bank** and run the corresponding version of Algorithm 4.7 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 function `hmflatprobit`, 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 in this manual 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.

4.11 Using the latent variable representation of the probit model, introduce $z_i|\beta \sim \mathcal{N}\left(\mathbf{x}^{i\mathsf{T}}\beta, 1\right)$ $(1 \le i \le n)$ such that $y_i = \mathbb{I}_{z_i \le 0}$. Deduce that

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

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

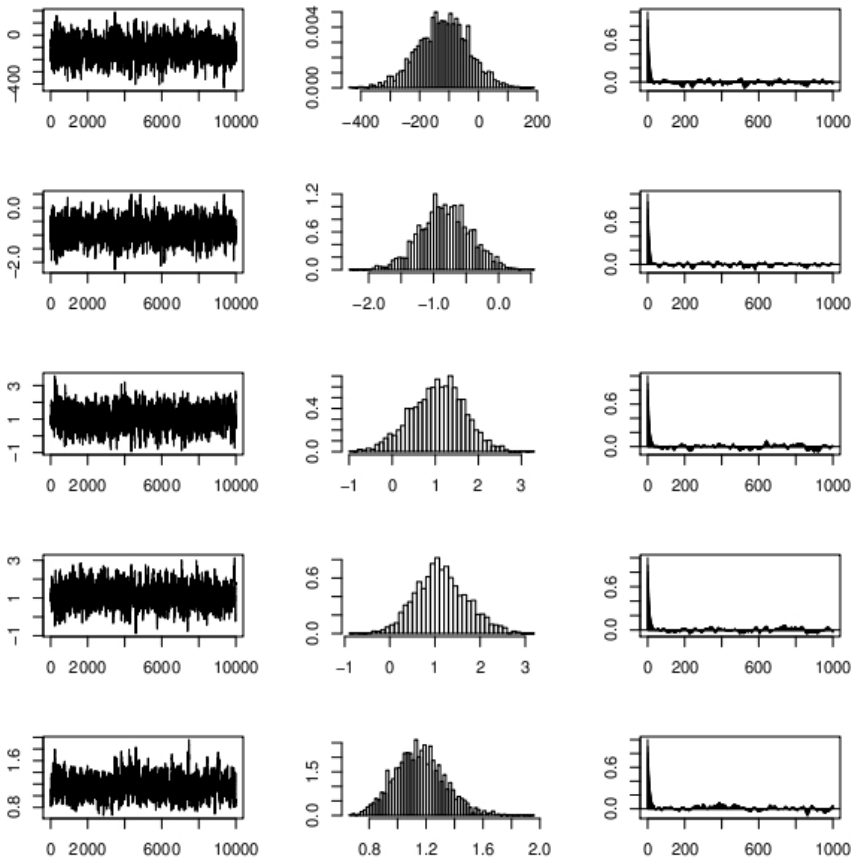

Fig. 4.5. bank: estimation of the probit coefficients [including one intercept β_0] via Algorithm 4.2 and a flat prior. *Left:* β_i's $(i = 0, \ldots, 4)$; *center:* histogram over the last $9,000$ iterations; *right:* auto-correlation over the last $9,000$ iterations.

```
> 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

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

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 in this manual. (*Note*: Account for differences in computing time.)

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

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

since $P(y_i = 1) = P(z_i \ge 0) = 1 - \Phi\left(-\mathbf{x}^{i\mathsf{T}}\beta\right) = \Phi\left(\mathbf{x}^{i\mathsf{T}}\beta\right)$. The conditional distribution of z_i given y_i is then a constrained normal distribution: if $y_i = 1$, $z_i \le 0$ 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, e.g., in Robert and Casella (2004): the cdf of the $\mathcal{N}_+(\mu, 1, 0)$ distribution is

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

(An alternative is to call the R library truncnorm.) 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

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

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.

A R code of this sampler is available as follows (based on a call to the R library truncnorm):

```
gibbsprobit=function(niter,y,X){
  p=dim(X)[2]
  beta=matrix(0,niter,p)
  z=rep(0,length(y))
  mod=summary(glm(y~-1+X,family=binomial(link="probit")))
  beta[1,]=as.vector(mod$coefficient[,1])
  Sigma2=solve(t(X)%*%X)
  for (i in 2:niter){
    mean=X%*%beta[i-1,]
    z[y==1]=rtruncnorm(sum(y==1),a=0,b=Inf,mean[y==1],sd=1)
    z[y==0]==rtruncnorm(sum(y==0),a=-Inf,b=0,mean[y==0],sd=1)
    Mu=Sigma2%*%t(X)%*%z
```

```
    beta[i,]=rmvn(1,Mu,Sigma2)
    }
  beta
  }
```

The output of this function is represented on Figure 4.6 in this manual. Note that the output is somehow smoother than on Figure 4.5 in this manual. (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 above codes require $27s$ and $26s$ over $10,000$ iterations for `hmflatprobit` and `gibbsprobit`. respectively.

4.12 For the **bank** dataset and the probit model, 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 `hmnoinfprobit` 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)])-
```

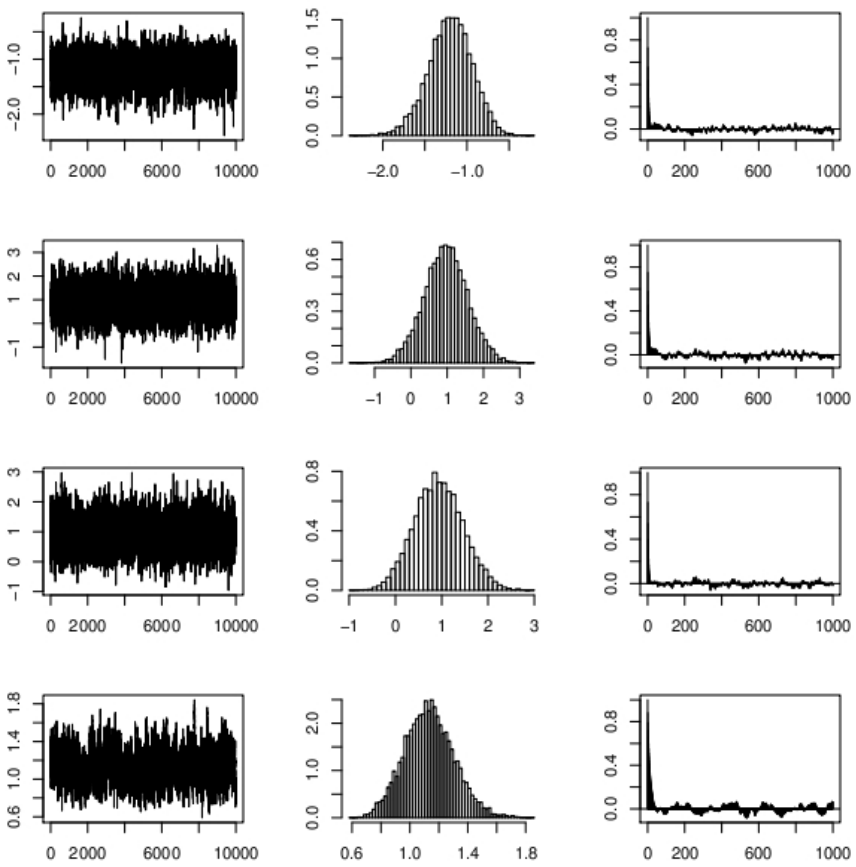

Fig. 4.6. bank: estimation of the probit coefficients [including one intercept β_0] by a Gibbs sampler 4.2 under a flat prior. *Left:* β_i's ($i = 0, \ldots, 4$); *center:* histogram over the last 9,000 iterations; *right:* auto-correlation over the last 9,000 iterations.

```
  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 `dt` 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.

4.13 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 4.6 on $(p_1, \ldots, p_k)$.

The only difference with Exercise 4.11 is in the use of a logistic density, hence both the Jacobian and the probabilities are modified:

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

4.14 Examine whether or not the sufficient conditions for propriety of the posterior distribution found in Exercise 4.9 for the probit model are the same for the logit model.

There is little difference with Exercise 4.8 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.

4.15 For the **bank** dataset 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 4.12.

This is very similar to Exercise 4.12, except that the parameters are now estimated for the logit model. The code is provided in `bayess` 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 4.12, 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.

4.16 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.

4.17 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 $\mathscr{D}_4(1/2, \ldots, 1/2)$, give the corresponding marginal distributions of $\alpha = \theta_{11} + \theta_{12}$ and $\beta = \theta_{11} + \theta_{21}$. 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 that

$$\frac{(\xi_1,\ldots,\xi_k)}{\xi_1+\ldots+\xi_k)} \sim \mathcal{D}_k(\delta_1,\ldots,\delta_k)$$

when

$$\xi_i \sim \mathcal{G}a(\delta_i,1)\,,\ i=1,\ldots,k\,.$$

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 $m_0(\mathcal{T})/m(\mathcal{T})$.

5

Capture–Recapture Experiments

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 $2(n^+ \vee 1) - 1$. 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.

$$\sum_{i=N^\star}^{\infty} 1/i(i+1) = 1/2\, 1/n^+ \vee 1 = 1/N^\star\,,$$

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.

5.2 Under the same prior as in Section 5.2.1, derive the marginal posterior density of N in the case where $n_1^+ \sim \mathscr{B}(N,p)$ and

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

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

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

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 $\hat{N} = 256$.

5.3 Show that the conditional distribution of m_2 conditional on both sample sizes n_1 and n_2 is given by (5.2) and does not depend on p. Deduce the expectation $\mathbb{E}^\pi[m_2|n_1, n_2, N]$.

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

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

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

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

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

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

with $n^+ = 45$ and $n^c = 50$. For $n^+ = 45$ and $n^c = 50$, the posterior mean is equal to 130.91.

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

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

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.

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.)

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

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

the likelihood is increasing for $N \leq n_1 n_2/m2$ and decreasing for $N \geq n_1 n_2/m2$. Thus $\hat{N} = n_1 n_2/m2$ 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.

5.7 Give both conditional posterior distributions involved in Algorithm 5.8 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.

5.8 Show that, for the two-stage capture model with probability p of capture, 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 $N - n_+$.

5.9 Reproduce the analysis of **eurodip** summarized by Figure 5.1 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 provided in the book (p.151) and in the demo for Chapter 5, 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.

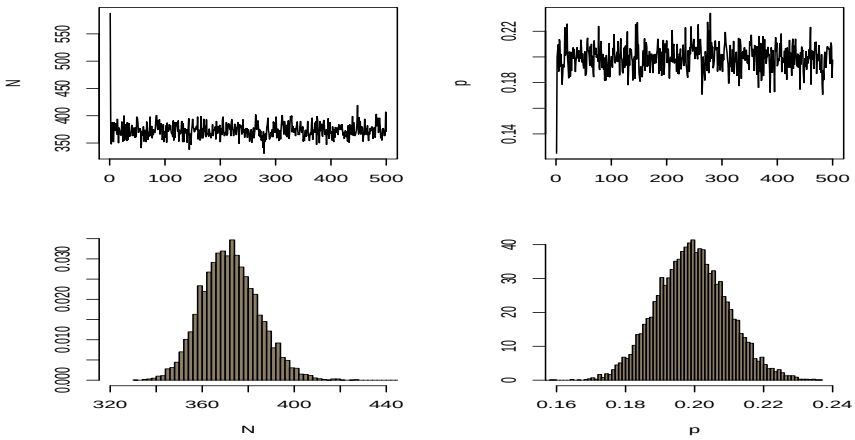

Fig. 5.1. eurodip: MCMC simulation under the prior $\pi(N,p) \propto N^{-1}$.

5.10 An extension of the T-stage capture–recapture model of Section 5.2.3 is to consider that the capture of an individual modifies its probability of being captured from p to q for future recaptures. 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,

$$n^* = Tn_1 + \sum_{j=2}^{T}(T-j+1)(n_j - m_j)$$

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.

5.11 Another extension of the 2-stage capture–recapture model is to allow for mark loss. 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

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

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 last two categories. Since z is not known, the [observed] likelihood is obtained by summation over z:

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

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.

5.12 Show that the conditional distribution of r_1 in the open population model of Section 5.3 is 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 $r_1 \le \min(n_1, n_1-r_2, n_1-r_2-c_3, n_1-c_2) = \min(n_1-r_2-c_3, n_1-c_2)$.

5.13 Show that the distribution of r_2 in the open population model of Section 5.3 can be integrated out from the joint distribution and that this leads to the following distribution on r_1:

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

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

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

5.14 Show that the likelihood associated with an open population as in Section 5.3 can be written as

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

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 $1\le i\le N$ 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 $O(3^{NT})$ [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.

5.15 In connection with the presentation of the accept-reject algorithm in Section 5.4, 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.

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 set $\mathscr{S}$ associated with the accept-reject algorithm in Section 5.4.

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.

5.17 Show that the probability of acceptance in Step 2 of Algorithm 5.9 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.

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 $N = 532$, $n_t = 118$, $\mu_t = -0.5$, and $\sigma^2 = 3$.

That the target is only known up to a constant is not a problem, as demonstrated in Exercise 5.20. 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

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

The upper bound thus involves a normal $\mathscr{N}(\mu_t-\sigma^2(N-n_t),\sigma^2)$ 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.

5.19 When uniform simulation on the accept-reject set $\mathscr{S}$ of Section 5.4 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.

5.20 Show that the normalizing constant M of a target density f can be deduced from the acceptance rate in the accept-reject algorithm (Algorithm 5.9 under the assumption that g is properly normalized.

This exercise generalises Exercise 5.17 where the target f is already normalised.

If $f(x) = M\tilde{f}(x)$ is a density to be simulated by Algorithm 5.9 and if g is a density such that

$$\tilde{f}(x) \le \tilde{M} g(x)$$

on the support of the density g, then running Algorithm 5.9 with an acceptance probability of $g(x)/\tilde{M}\tilde{f}(x)$ produces simulations from f since the accepted values have the marginal density proportional to

$$\int_0^1 \mathbb{I}_{[0,\tilde{f}(x)/\tilde{M}g(x)]}(u)\,\mathrm{d}u\, g(x) = \frac{\tilde{f}(x)}{\tilde{M}} \propto f(x)\,.$$

In that case, the average probability of acceptance is

$$\int_{\mathcal{X}} \frac{\tilde{f}(x)}{\tilde{M}}\,\mathrm{d}x = \int_{\mathcal{X}} \frac{f(x)}{M\tilde{M}}\,\mathrm{d}x = \frac{1}{M\tilde{M}}\,.$$

Since the value of $\tilde{M}$ is known, the average acceptance rate over simulations, $\hat{\varrho}$, leads to estimate M as

$$\hat{M} = \frac{1}{\hat{\varrho}\tilde{M}}\,.$$

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

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

5.22 Modify the function `ardipper` used in Section 5.4 to return the acceptance rate as well as a sample from the target distribution.

As provided in Section 5.4, the function `ardipper` is defined by

```
ardipper=function(nsimu=1,n1,c2,c3,r2,q2){

  barr=min(n1-c2,n1-r2-c3)
  boundM=thresh(0,n1,c2,c3,r2,barr)
```

```
  echan=1:nsimu
  for (i in 1:nsimu){
    test=TRUE
    while (test){
      y=rbinom(1,size=barr,prob=q2)
      test=(runif(1)>thresh(y,n1,c2,c3,r2,barr))
      }
    echan[i]=y
    }
  echan
}
```

The requested modification consists in monitoring the acceptance rate and returning a list with both items:

```
ardippest=function(nsimu=1,n1,c2,c3,r2,q2){

  barr=min(n1-c2,n1-r2-c3)
  boundM=thresh(0,n1,c2,c3,r2,barr)
  echan=1:nsimu
  acerate=-nsimu
  for (i in 1:nsimu){
    test=TRUE
    while (test){
      y=rbinom(1,size=barr,prob=q2)
      test=(runif(1)>thresh(y,n1,c2,c3,r2,barr))
      acerate=acerate+1
      }
    echan[i]=y
    }
  list(sample=echan,reject=acerate/nsimu)
}
```

5.23 Show that, given a mean and a 95% confidence interval in $[0,1]$, there exists at most one beta distribution $\mathscr{B}e(a,b)$ 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 4.17 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.

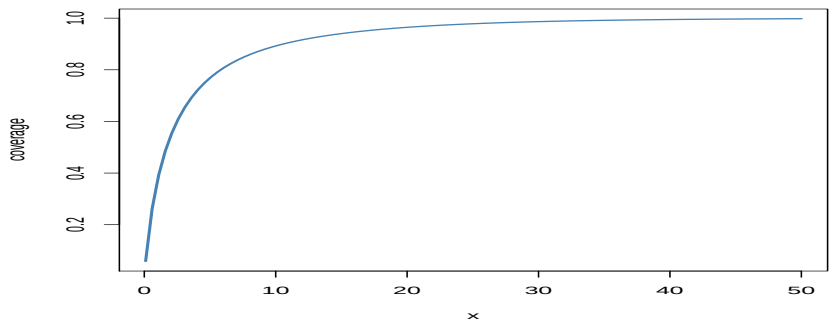

Fig. 5.2. Coverage of the interval $(\ell, u) = (0.1, 0.6)$ by a $\mathscr{B}e(0.4\alpha, 0.6\alpha)$ distribution when α varies.

5.24 Show that, for the Arnason–Schwarz model, 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

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.

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

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

Deduce that the number of partition sets is of order $\mathrm{O}(n^{k-1})$. (*Hint:* This is a classical combinatoric problem.)

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.

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 ($j = 1, 2$)

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

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 $\hat{s}_j^2(\mathbf{z})$ are the empirical mean and variance for the subsample with z_i equal to j, and

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

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

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

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., Exercises 2.7 and 2.15).

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.)

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

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

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

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

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 in this manual in this manual 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.

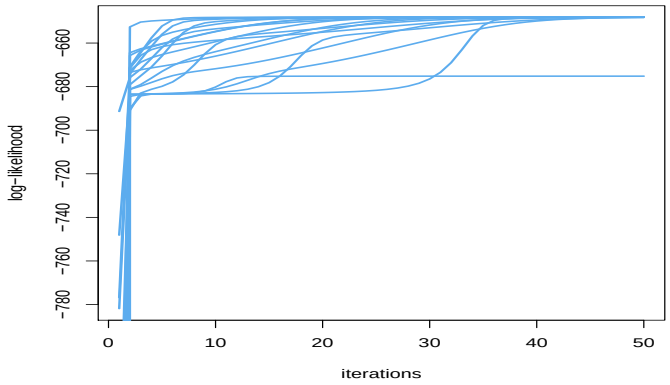

Fig. 6.1. Increase of the log-likelihood along EM iterations for 20 different starting points.

6.5 In the mixture model with independent priors on the θ_j's, show that the θ_j's are dependent on each other given (only) $\mathbf{x}$ by summing out the $\mathbf{z}$'s.

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.

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 `gibbsmean` in the `bayess`: 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 in this manual. This reparameterisation thus seems less robust than the original parameterisation.

6.7 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

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

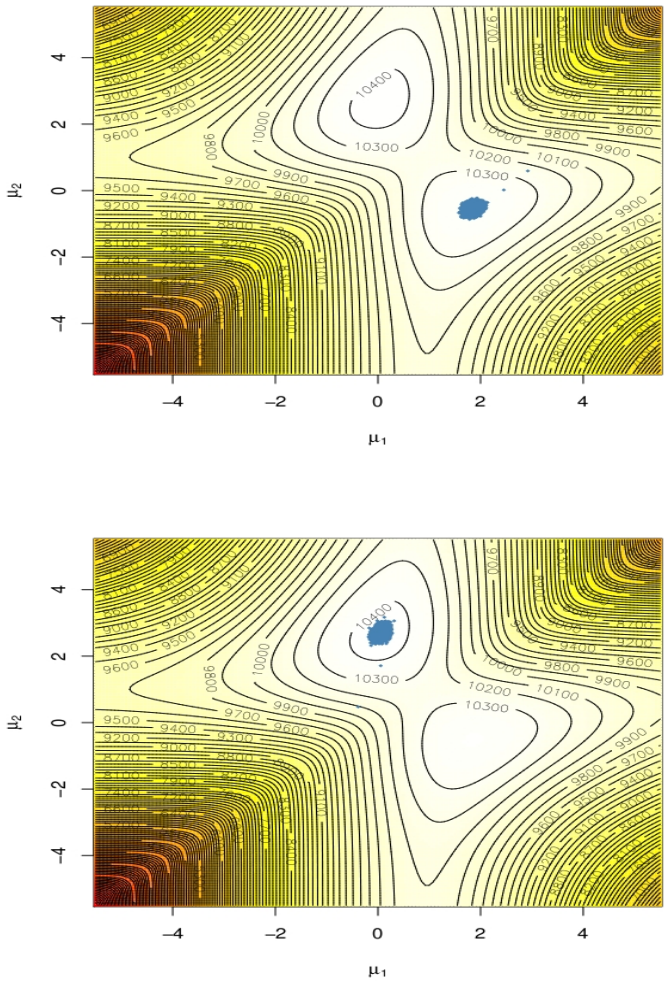

Fig. 6.2. Influence of the starting value on the convergence of the Gibbs sampler associated with the location parameterisation of the mean mixture (10,000 iterations).

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.

6.8 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. (*Note*: This is a crude version of the *simulated annealing* algorithm. See also Chapter 8.) Discuss the modifications required in Algorithm 6.11 to achieve simulation from $\pi(\theta|\mathbf{x})^{\gamma}$ when $\gamma \in \mathbb{N}^*$ is an integer.

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 6.1 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 $\mathcal{N}\left(\dfrac{\gamma\lambda\delta + \bar{x}_1\left(\mathbf{z}\right)}{\gamma\lambda + \ell}, \dfrac{1}{\gamma\lambda + \ell}\right)$
4. Generate $\mu_2^{(t)}$ from $\mathcal{N}\left(\dfrac{\gamma\lambda\delta + \bar{x}_2\left(\mathbf{z}\right)}{\gamma\lambda + \gamma n - \ell}, \dfrac{1}{\gamma\lambda + \gamma n - \ell}\right)$.

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.

6.9 Show that the ratio (6.7) 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.

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

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

6.10 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.

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

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 6.2 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 $\mathcal{N}\left(\dfrac{\gamma\lambda\delta + bar x_1(\mathbf{z})}{\gamma\lambda+\ell}, \dfrac{1}{\gamma\lambda+\ell}\right)$
4. Generate $\mu_2^{(t)}$ from $\mathcal{N}\left(\dfrac{\gamma\lambda\delta + \bar{x}_2(\mathbf{z})}{\gamma\lambda+\gamma n-\ell}, \dfrac{1}{\gamma\lambda+\gamma n-\ell}\right)$.

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.

6.12 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 in this manual exhibits the characteristic stripes of an explosive likelihood as σ approaches 0 for values of μ close to the values of the sample.

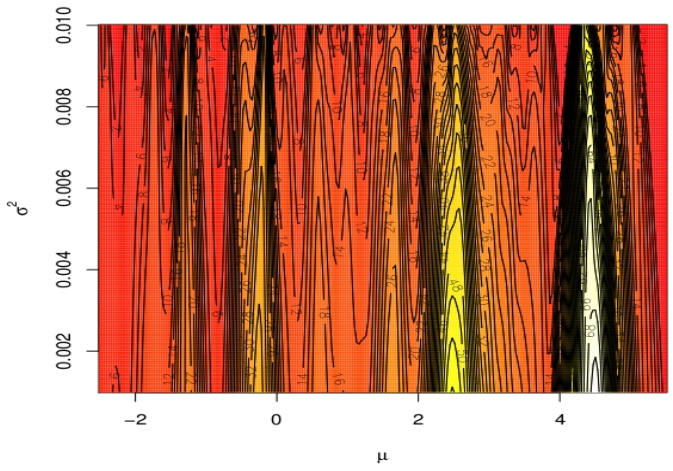

Fig. 6.3. Illustration of an unbounded mixture likelihood.

7

Dynamic Models

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

$$w_t = (2q+1)^{-1}\textstyle\sum_{j=-q}^{q} x_{t+j}\,.$$

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,

$$\gamma_w(t+h,t) = \mathbb{I}_{|h|\le q}(q+1-|h|)\sigma^2\,.$$

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 $\tau^2 = 4\sigma^2/3$.

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.

7.4 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.

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 $\tau^2 = 4\sigma^2/3$.

7.5 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 **Eurostoxx50** and evaluate its predictive abilities.

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

$$\begin{aligned}
&\int \left(\sigma^2\right)^{-T/2-1/2} \exp\left(\sum_{t=1}^{T}(x_t - \rho x_{t-1})^2/(2\sigma^2)\right) \mathrm{d}\sigma \\
&\quad = \int \left(\sigma^2\right)^{-T/2-1} \exp\left(\sum_{t=1}^{T}(x_t - \rho x_{t-1})^2/(2\sigma^2)\right) \mathrm{d}\sigma^2 \\
&\quad = \left\{\sum_{t=1}^{T}(x_t - \varrho\, x_{t-1})^2\right\}^{-T/2}
\end{aligned}$$

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

$$\begin{aligned}
&\sum_{t=0}^{T-1} x_t^2(\varrho - \mu_T)^2 + \sum_{t=1}^{T} x_t^2 - \sum_{t=0}^{T-1} x_t^2\mu_T^2 \\
&\qquad \propto \frac{(\varrho - \mu_T)^2}{T-1} + \frac{1}{T-1}\left(\frac{\sum_{t=1}^{T} x_t^2}{\sum_{t=0}^{T-1} x_t^2} - \mu_T^2\right) \\
&\qquad = \frac{(\varrho - \mu_T)^2}{T-1} + \nu_T^2\,.
\end{aligned}$$

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:

$$\begin{aligned}\int \frac{1}{\sqrt{2\pi}\sigma}\exp\{-(x_{T+1}-\varrho x_T)^2/2\sigma^2\}\,\pi(\sigma,\varrho|\mathbf{x}_{0:T})\mathrm{d}\sigma\mathrm{d}\varrho \\ &\propto \int \sigma^{-T-2}\exp\{-\sum_{t=0}^{T}(x_{t+1}-\varrho x_t)^2/2\sigma^2\}\mathrm{d}\sigma\mathrm{d}\varrho \\ &\propto \int \left\{\sum_{t=0}^{T}(x_{t+1}-\varrho\, x_t)^2\right\}^{-(T+1)/2}\mathrm{d}\varrho \\ &\propto \left(\sum_{t=0}^{T}x_t^2\right)^{-(T+1)/2}\int\left\{\frac{(\varrho-\mu_{T+1})^2}{T}+\nu_{T+1}^2\right\}^{-(T+2)/2}\mathrm{d}\varrho \\ &\propto \left(\sum_{t=0}^{T}x_t^2\right)^{-(T+1)/2}\nu_T^{-T-1} \\ &\propto \left(\sum_{t=0}^{T}x_t^2\sum_{t=0}^{T}x_{t+1}^2-\left\{\sum_{t=0}^{T}x_tx_{t+1}\right\}^2\right)^{(T+1)/2}.\end{aligned}$$

This is a Student $\mathcal{T}(T,\delta_T,\omega_T)$ distribution, with

$$\delta_T = x_T\sum_{t=0}^{T-1}x_tx_{t+1}\Big/\sum_{t=0}^{T-1}x_t^2 = \hat{\rho}_T x_T$$

and

$$\omega_T = \left\{\sum_{t=0}^{T}x_t^2\sum_{t=0}^{T}x_t^2-\left(\sum_{t=0}^{T}x_tx_{t+1}\right)^2\right\}\Big/T\sum_{t=0}^{T-1}x_t^2\,.$$

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.

7.6 For Algorithm 7.13, show that, if the proposal on σ^2 is a log-normal distribution $\mathcal{LN}(\log(\sigma^2_{t-1}),\tau^2)$ 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.

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.

7.7 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 $\mathbf{y}_{0:T}$.

7.8 Show that the stationary distribution of $\mathbf{x}_{-p:-1}$ in an AR(p) model 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

$$\mathbf{z}_t = \left(x_t, x_{t-1}, \ldots, x_{t+1-p}\right)\,,$$

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.

7.9 Show that the posterior distribution on $\boldsymbol{\theta}$ associated with the prior $\pi(\boldsymbol{\theta}) = 1/\sigma^2$ and an AR(p) model is well-defined for $T > p$ observations.

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 $\boldsymbol{\varrho}$.

7.10 Show that the coefficients of the polynomial $\mathcal{P}$ in (7.15) associated with an AR(p) model 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 $\varrho_j = -\psi_j^p$ $(j = 1, \ldots, p)$.

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 $\mathrm{O}(p^2)$.

7.11 Given the polynomial $\mathcal{P}$ in (7.5), the fact that all the roots are outside the unit circle can be determined without deriving the roots, thanks to the Schur–Cohn test. If $\mathcal{A}_p = \mathcal{P}$, a recursive definition of decreasing degree polynomials is $(k = p, \ldots, 1)$

$$u\mathcal{A}_{k-1}(u) = \mathcal{A}_{k-1}(u) - \varphi_k \mathcal{A}_k^\star(u)\,,$$

where $\mathcal{A}_k^\star$ denotes the reciprocal polynomial $\mathcal{A}_k^\star(u) = u^k \mathcal{A}_{k-1}(1/u)$.

1. Give the expression of φ_k in terms of the coefficients of $\mathcal{A}_k$.
2. Show that the degree of $\mathcal{A}_k$ is at most k.
3. If $a_{m,k}$ denotes the m-th degree coefficient in $\mathcal{A}_k$, show that $a_{k,k} \neq 0$ for $k = 0, \ldots, p$ if, and only if, $a_{0,k} \neq a_{k,k}$ for all k's.

4. Check by simulation that, in cases when $a_{k,k} \neq 0$ for $k = 0, \ldots, p$, the roots are outside the unit circle if, and only if, all the coefficients $a_{k,k}$ are positive.

Note: The above exercise is somewhat of a mystery (!) in that we cannot remember how it ended up in this exercise list, being incorrect and incomplete as stated. A proper substitute is given below:

7.11 Given a polynomial $\mathcal{P}$ of degree k, its reciprocal polynomial $\mathcal{P}_k^\star$ is defined as

$$\mathcal{P}^\star(u) = u^k \mathcal{P}_{k-1}(1/u)\,.$$

Assuming $\mathcal{P}(0) = 1$, the Schur transform of $\mathcal{P}$ is defined by

$$T\mathcal{P}(u) = \frac{\mathcal{P}(z) - \mathcal{P}^\star(0)\mathcal{P}^\star(z)}{1 - \mathcal{P}^\star(0)^2}\,.$$

1. Show that the roots of $\mathcal{P}$ and $\mathcal{P}_k^\star$ are inverses.
2. Show that the degree of $T\mathcal{P}$ is at most $k - 1$.
3. Show that $T\mathcal{P}(0) = 1$.
4. Check by a simulation experiment producing random polynomials the property that, when $T\mathcal{P}(0) > 1$, $T\mathcal{P}$ and $T\mathcal{P}$ have the same number of roots inside the unit circle.
5. Denote $T^n\mathcal{P} = T(T^{n-1}\mathcal{P})$, for $d \neq k$, and κ the first index with $T^\kappa\mathcal{P} = 0$. Deduce from the above property that, if $T^n\mathcal{P} > 0$ for $n = 1, \ldots, \kappa$, then $\mathcal{P}$ has no root inside the unit circle.

1. If we write the inverse root decomposition of $\mathcal{P}$ as

$$\mathcal{P}(u) = \prod_{i=1}^{k} (1 - \lambda_i u)\,,$$

since $\mathcal{P}(0) = 1$, we have

$$\mathcal{P}^\star(u) = u^k \prod_{i=1}^{k} (1 - \lambda_i u^{-1}) = \prod_{i=1}^{k} (u - \lambda_i) = \prod_{i=1}^{k} (1 - \lambda_i^{-1} u)\,.$$

2. By definition, if $\mathcal{P}(u) = \sum_{i=0}^{k} \alpha_i u^i$, then

$$\mathcal{P}^\star(u) = \sum_{i=0}^{k} \alpha_{k-i} u^i\,,$$

$\mathcal{P}^\star(0) = \alpha_k$, and

$$\mathcal{P}(u) - \mathcal{P}^\star(0)\mathcal{P}^\star(u) = \alpha_k u^k + \sum_{i=1}^{k-1} \alpha_i u^i - \alpha_k u^k - \alpha_k \sum_{i=1}^{k-1} \alpha_{k-i} u^i$$
$$= \sum_{i=1}^{k-1} [\alpha_i - \alpha_k \alpha_{k-i}] u^i$$

is at most of degree $k-1$.

3. Since
$$\mathcal{P}(0) - \mathcal{P}^\star(0)\mathcal{P}^\star(0) = 1 - \alpha_k^2 \,,$$
$T\mathcal{P}(0) = 1$.
4. A simulation experiment can be designed around the following code:

```
k=10
# random coefficients
Coef=c(1,runif(k,-1,1))
Schur=Coef-Coef[k]*rev(Coef)
print(sum(Mod(polyroot(Coef))<1)-sum(Mod(polyroot(Schur))<1))
```

Repeating this code a large number of times does not produce anything but zero's.

5. By virtue of the above result, $\mathcal{P}, T\mathcal{P}, \ldots, T^{\kappa-1}\mathcal{P}$ have the same number of roots inside the unit circle if $T^n\mathcal{P} > 0$ for $n = 1, \ldots, \kappa - 1$. Since
$$T^{\kappa-1}\mathcal{P} = 1 - \{\alpha_1^\kappa\}^2 = 1 - \lambda_1^2 \,,$$
the last root is outside the unit disk and hence so are the others.
6. Extending the above code leads to

```
k=10
# Schur sequence
Coef=matrix(0,nrow=k+1,ncol=k+1)
# initial polynomial
Coef[,k+1]=c(1,rnorm(k,sd=1/k))
for (t in k:1)
  Coef[1:t,t]=(Coef[1:(t+1),t+1]-Coef[t+1,t+1]*Coef[(t+1):1,
           t+1])/(1-Coef[t+1,t+1]^2)
while (prod(diag(Coef[1,]^2)<1)==0){
  Coef=matrix(0,nrow=k+1,ncol=k+1)
  Coef[,k+1]=c(1,rnorm(k,sd=1/k))
  for (t in k:1)
    Coef[1:t,t]=(Coef[1:(t+1),t+1]-Coef[t+1,t+1]*Coef[(t+1):1,
       t+1])/(1-Coef[t+1,t+1]^2)
  }
print(min(Mod(polyroot(Coef[,k+1]))))
```

Repeated calls to this code consistently exhibit root modules larger than 1.

7.12 For an MA(q) process, show that ($s \leq q$)

$$\gamma_x(s) = \sigma^2 \sum_{i=0}^{q-|s|} \vartheta_i \vartheta_{i+|s|} \,.$$

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 $1 \leq s \leq q$,

$$\gamma_x(s) = \left[\vartheta_s + \vartheta_{s+1}\vartheta_1 + \ldots + \vartheta_q\vartheta_{q-s}\right]\sigma^2$$

and

$$\gamma_x(0) = \left[1 + \vartheta_1^2 + \ldots + \vartheta_q^2\right]\sigma^2 \,.$$

Therefore, if ($0 \leq s \leq q$) 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.

7.13 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.

7.14 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 this distribution only depends on the other ϵ_{-i}'s, $\mathbf{x}_{1:q-t+1}$, and $\widehat{\epsilon}_{1:q-t+1}$.

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.

7.15 Show that the (useful) 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.

7.16 Show that the system of equations given by (7.13) and (7.14) induces a Markov chain on the completed variable $(\mathbf{x}_t, \mathbf{y}_t)$. Deduce that state-space models are special cases of hidden Markov models.

Given the time-dependence structure

$$\begin{aligned} \mathbf{x}_t &= G\mathbf{y}_t + \boldsymbol{\varepsilon}_t\,, \\ \mathbf{y}_{t+1} &= F\mathbf{y}_t + \boldsymbol{\xi}_t\,, \end{aligned}$$

we can write

$$\begin{pmatrix} \mathbf{x}_t \\ \mathbf{y}_{t+1} \end{pmatrix} = \begin{pmatrix} O & G \\ O & F \end{pmatrix} \begin{pmatrix} \mathbf{x}_{t-1} \\ \mathbf{y}_t \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_t \\ \boldsymbol{\xi}_t \end{pmatrix}\,.$$

Since the noises $\boldsymbol{\xi}_t$ and $\boldsymbol{\varepsilon}_t$ are independent, the full vector $(\mathbf{x}_t, \mathbf{y}_{t+1})$ is indeed a Markov chain. The subchain $(\mathbf{y}_t)$ is also a Markov chain on itd own. And observing *only* $\mathbf{x}_t$ means that we are observing a hidden Markov chain, in the sense of Figure 7.7 in the book.

7.17 Show that, for a hidden Markov model, 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.

7.18 Given a hidden Markov chain (x_t, y_t) with both x_t and y_t taking a finite number of possible values, k and κ, show that the time required for the simulation of T consecutive observations is in $\mathrm{O}(k\kappa T)$.

Note: The order indicated in the exercise should be $\mathrm{O}(\kappa^2 T)$, for the distribution conditional on the observed x_t's.

For direct simulation, given the hidden chain at time t, y_t, simulating y_{t+1} requires up to k comparisons with a uniform variate. Given y_{t+1}, simulating x_{t+1} involves another maximum of κ comparisons with a uniform variate. Repeating those steps T times leads to a $\mathrm{O}(\{k+\kappa\}T)$ time.

For inverse simulation, that is, after observing $(x_1,\ldots,x_T)$, the joint conditional distribution of $(y_1,\ldots,y_T)$ is given by

$$p(y_1,\ldots,y_T|x_1,\ldots,x_T) \propto p_0(y_1)p(y_2|y_1)\cdots p(y_T|y_{T-1})p(x_1|y_1)\cdots p(x_T|y_T)\,,$$

which takes κ^T values.

However, if we use the backward formula described in the book, we could gain some time. If we get back to the defintion of the backward formula, the distribution of y_T given the past being only conditional on y_{T-1}, $p(y_T|y_{T-1},\mathbf{x}_{0:T})$, takes κ^2 values. Then, for each previous hidden state, y_t, $p(y_t|y_{t-1},\mathbf{x}_{0:T})$ involves a summation of κ terms for all pairs (y_{t-1},y_t). But the summation

$$\sum_{i=1}^{\kappa} p^\star_{t+1}(i|y_t,\mathbf{x}_{1:T})$$

only depends on y_t, thus has to be computed κ times, to be later multiplied by $p_{y_{t-1}y_t}$. Therefore the cost of producing $p(y_t|y_{t-1},\mathbf{x}_{0:T})$ is again of order κ^2. At last, $p(y_0|\mathbf{x}_{0:T})$ requires κ summations of κ terms, thus is again of order

κ^2. This confirms that the overall cost is in $\mathrm{O}(\kappa^2 T)$ and that the number of possible values of the x_t's is irrelevant.

7.19 Implement Chib's method of Section 6.8 in the case of a doubly finite hidden Markov chain. First, show that an equivalent to the approximation (6.9) is available for the denominator of (6.8). Second, discuss whether or not the label switching issue also rises in this framework. Third, apply this approximation to **Dnadataset**.

In a hidden Markov model (x_t, y_t), y_t being the hidden part, when the parameters are unknown, it is usually the case that the full posterior distribution of the parameter $\pi(\mathbf{p}, \mathbf{q}|\mathbf{x}, \mathbf{y})$ is available in closed form. In particular, as shown in Algorithm 7.15, this full posterior distribution is a product of κ Beta distributions on the $p_{i\cdot}$'s and of κ Dirichlet distributions on the $q_{i\cdot}$'s $(i = 1, 2)$.

As alluded to in the book, it is also a setting where label switching occurs. Indeed, the introduction of states 1 and 2 in the hidden chain does not identify which state is which. The posteriors on $\mathbf{q}^1$ and $\mathbf{q}^2$ should therefore be the same. Since the Gibbs sampler does not produce such symmetry on Figure 7.9, it is quite likely that Chib's approximation will be biased in this setting.

The implementation for **Dnadataset** of the Chib involves picking the highest likelihood value for $\theta = (\mathbf{q}^1, \mathbf{q}^2, \mathbb{P})$ and averaging the full conditionals of θ given the hidden chain over the Gibbs iterations.

7.20 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

$$\varphi_r(i) \propto \sum_{j=1}^{\kappa} p_{ji} f(x_{r-1}|x_{r-2}, y_{r-1} = j)\varphi_{r-1}(j) .$$

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

8.1 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!

8.2 Using the Hammersley–Clifford theorem, show that the full conditional distributions given by (8.3) are compatible with a joint distribution. Deduce that the Ising model is a Markov random field.

Note: In order to expose the error made in the earlier printing of *Bayesian Core*, namely using the size of the symmetrized neighborhood, $N_k(i)$, in the full conditoinal, we will compute here 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 $N_k(i) = 1$ 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 (8.3) is defined as in the earlier edition. Taking all the $N_k(j)$'s equal to 1 leads to a coherent joint distribution since the last line in the above equation cancels.

8.3 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.

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 $m(\mathbf{y}_{-i})$.

8.4 Describe the collection of cliques $\mathcal{C}$ for an 8 neighbor neighborhood structure such as in Figure 8.2 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 8.1 in this manual, 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.

8.5 Draw the function $Z(\beta)$ for a 3×5 array. Determine the computational cost of the derivation of the normalizing constant $Z(\beta)$ of (8.4) for an $m \times n$ array.

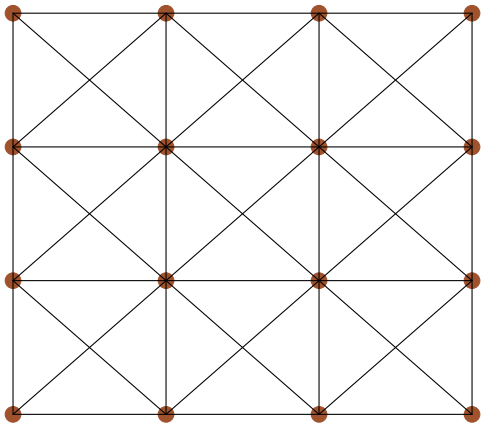

Fig. 8.1. Neighborhood relations between the points of a 4×4 regular grid for a 8 neighbor neighborhood structure.

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 8.2 in this manual.

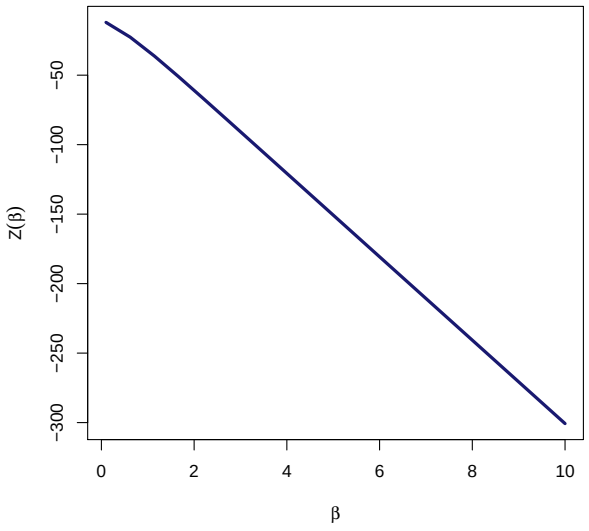

Fig. 8.2. Plot of the function $Z(\beta)$ for a 3×5 array with a four neighbor structure.

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.

8.6 Show that the joint distribution (8.5) is indeed compatible with the full conditionals of the Potts model. Can you derive this joint distribution from the Hammersley–Clifford representation (8.1)?

If we defined the joint distribution as

$$\pi(\mathbf{x}) \propto \exp\left(\beta \sum_{(i,j);\, j\sim i} \mathbb{I}_{x_j=x_i}\right). \tag{8.5}$$

the full conditional distribution of x_i is

$$\begin{aligned}
\pi(x_i = g|\mathbf{x}_{-i}) &\propto \pi((g, \mathbf{x}_{-i}) \\
&\propto \exp\left(\beta \sum_{\substack{(u,v);\, u\sim v \\ u,v\neq i}} \mathbb{I}_{x_u=x_v} + \sum_{u;\, i\sim u} \mathbb{I}_{x_u=g}\right) \\
&\propto \exp\left(\beta \sum_{u;\, i\sim u} \mathbb{I}_{x_u=g}\right) \\
&= \exp\left(\beta n_{i,g}\right)
\end{aligned}$$

Conversely, if we start from the full conditionals

$$\pi(x_i = g|\mathbf{x}_{-i}) \propto \exp(\beta n_{i,g})\,. \quad i \in \mathcal{I}\,,\, 1 \le g \le G\,,$$

and apply the Hammersley–Clifford representation (8.1)

$$\frac{\pi(\mathbf{x})}{\pi(\mathbf{x}^*)} = \prod_{i=0}^{n-1} \frac{\pi(x_{i+1}|x_1^*, \ldots, x_i^*, x_{i+2}, \ldots, x_n)}{\pi(x_{i+1}^*|x_1^*, \ldots, x_i^*, x_{i+2}, \ldots, x_n)}\,,$$

we have

$$\begin{aligned}
\frac{\pi(x_1|x_2, \ldots, x_n)}{\pi(x_1^*|x_2, \ldots, x_n)} &= \exp\left(\beta \sum_{u;\, 1\sim u} \left[\mathbb{I}_{x_u=x_1} - \mathbb{I}_{x_u=x_1^*}\right]\right) \\
\frac{\pi(x_2|x_1^*, x_3, \ldots, x_n)}{\pi(x_2^*|x_1^*, x_3, \ldots, x_n)} &= \exp\left(\beta \mathbb{I}_{1\sim 2}\left[\mathbb{I}_{x_1^*=x_2} - \mathbb{I}_{x_1^*=x_2^*}\right] + \sum_{u>1;\, 2\sim u} \left[\mathbb{I}_{x_u=x_2} - \mathbb{I}_{x_u=x_2^*}\right]\right) \\
&\vdots \\
\frac{\pi(x_n|x_1^*, \ldots, x_{n-1}^*)}{\pi(x_n^*|x_1^*, \ldots, x_{n-1}^*)} &= \exp\left(\beta \sum_{u;\, n\sim u} \left[\mathbb{I}_{x_u^*=x_n} - \mathbb{I}_{x_u^*=x_n^*}\right]\right)
\end{aligned}$$

which means that all terms involving both x_i and x_j^* cancel out and that

$$\pi(\mathbf{x}) \propto \exp\left(\beta \sum_{(i,j);\, j\sim i} \mathbb{I}_{x_j=x_i}\right). \tag{8.5}$$

This exercise is essentially the same as Exercise 8.9.

8.7 For an $n \times m$ array $\mathcal{I}$, if the neighbourhood relation is based on the four nearest neighbors, show that the $x_{i,j}$'s for which $(i+j) \equiv 0 (\text{mod } 2)$ are independent conditional on the $x_{i,j}$'s for which $(i+j) \equiv 1 (\text{mod } 2)$ $(1 \leq i \leq n, 1 \leq j \leq m)$. 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 8.16 is a version of *the Swendsen–Wang* algorithm.)

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).

8.8 Determine the computational cost of the derivation of the normalizing constant of the distribution (8.5) for an $n \times m$ array and G different colors.

Just as in Exercise 8.5, 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 $\mathrm{O}(m \times n \times G^{m \times n})$.

8.9 Use the Hammersley–Clifford theorem to establish that (8.5) is the joint distribution associated with the conditionals above. Deduce that the Potts model is an MRF.

Similar to the resolution of Exercise 8.2, using the Hammersley-Clifford representation (8.5) and defining an arbitrary order on the set $\mathcal{I}$ leads to the joint distribution

$$\pi(\mathbf{x}) \propto \frac{\exp\left\{\beta \sum_{i\in\mathcal{I}} \sum_{j<i,j\sim i} \mathbb{I}_{x_i=x_j} + \sum_{j>i,j\sim i} \mathbb{I}_{x_i=x_j^\star}\right\}}{\exp\left\{\beta \sum_{i\in\mathcal{I}} \sum_{j<i,j\sim i} \mathbb{I}_{x_i^\star=x_j} + \sum_{j>i,j\sim i} \mathbb{I}_{x_i^\star=x_j^\star}\right\}}$$
$$\propto \exp\left\{\beta\left(\sum_{j\sim i,j<i} \mathbb{I}_{x_i=x_j} + \sum_{j\sim i,j>i} \mathbb{I}_{x_i=x_j^\star} - \sum_{j\sim i,j>i} \mathbb{I}_{x_j^\star=x_i}\right)\right\}$$
$$= \exp\left\{\beta \sum_{j\sim i} \mathbb{I}_{x_i=x_j}\right\}.$$

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.

8.10 Derive an alternative to Algorithm 8.17 where the probabilities in the multinomial proposal are proportional to the numbers of neighbors $n_{u_\ell,g}$ and compare its performance with that of Algorithm 8.17.

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.

8.11 Show that the Swendsen–Wang improvement given in Exercise 8.7 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 $\mathbf{x}$.

8.12 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 $0 \le \alpha_0 < \alpha_1 \le 2$.

This follows directly from the R code in `demo/Chapter.8.R` as `sumising`, with

$$\int_{\alpha_0}^{\alpha_1} \hat{f}(\beta)\,\mathrm{d}\beta \approx \sum_{i,\alpha_0 \le \beta_i \le \alpha_1} f(\beta_i)(\beta_{i+1} - \beta_i)\,,$$

with the appropriate corrections at the boundaries.

8.13 Show that the estimators $\widehat{\mathbf{x}}$ that minimize the posterior expected losses $\mathbb{E}^\pi[L_1(\mathbf{x},\widehat{\mathbf{x}})|\mathbf{y})]$ and $\mathbb{E}^\pi[L_2(\mathbf{x},\widehat{\mathbf{x}})|\mathbf{y}]$ 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.

8.14 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.

8.15 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

$$\pi(\mathbf{x}|\mathbf{y})^\kappa = \int \pi(\mathbf{x}, \theta_1|\mathbf{y})\,\mathrm{d}\theta_1 \times \cdots \times \int \pi(\mathbf{x}, \theta_\kappa|\mathbf{y})\,\mathrm{d}\theta_\kappa\,, \tag{8.1}$$

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 (8.9) at each iteration.

The representation (8.10) is obvious since

$$\begin{aligned}\left(\int \pi(\mathbf{x}, \theta|\mathbf{y})\,\mathrm{d}\theta\right)^\kappa &= \int \pi(\mathbf{x}, \theta|\mathbf{y})\,\mathrm{d}\theta \times \cdots \times \int \pi(\mathbf{x}, \theta|\mathbf{y})\,\mathrm{d}\theta \\ &= \int \pi(\mathbf{x}, \theta_1|\mathbf{y})\,\mathrm{d}\theta_1 \times \cdots \times \int \pi(\mathbf{x}, \theta_\kappa|\mathbf{y})\,\mathrm{d}\theta_\kappa\end{aligned}$$

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).

8.16 For the Ising model, show that the distribution (8.4) can be also defined as

$$\pi(\mathbf{x}) \propto \exp\left(2\beta \sum_{j\sim i} \mathbb{I}_{x_j = x_i = 1}\right)$$

when the number of neighbors is constant.

Since

$$\pi(\mathbf{x}) \propto \exp\left(\beta \sum_{j\sim i} \mathbb{I}_{x_j=x_i}\right) ,$$

we have

$$\begin{aligned}\pi(\mathbf{x}) &\propto \exp\left(\beta \sum_{j\sim i} \mathbb{I}_{x_j=x_i=1} + \beta \sum_{j\sim i} \mathbb{I}_{x_j=x_i=-1}\right)\\ &= \exp\left(\beta \sum_{j\sim i} \mathbb{I}_{x_j=x_i=1} + \beta \left[N - \sum_{j\sim i} \mathbb{I}_{x_j=x_i=1}\right]\right)\\ &= \exp\left(2\beta \sum_{j\sim i} \mathbb{I}_{x_j=x_i=1}\right) \exp(N\beta)\end{aligned}$$

if N denotes the number of connected pairs $i \sim j$.

8.17 Show that the joint distribution (8.4) can be obtained from the full conditionals (8.3) by virtue of the Hammerseley-Clifford representation (8.1).

This is a special case of Exercise 8.9 since the Ising model is a Potts model with only two modalities.

8.18 Show that the Ising distribution is symmetric in that inverting the color of all pixels does not change the probability (8.4).

Given the definition of the Ising model as

$$\pi(\mathbf{x}) \propto \exp\left(\beta \sum_{j\sim i} \mathbb{I}_{x_j=x_i}\right) , \tag{8.3}$$

switching 1's and -1's does not modify the right hand side and hence does not change $\pi(\mathbf{x})$.

8.19 For the Ising model, run a simulation experiment that should locate the limiting value of β above which almost all pixels are of the same color. Same question for the (negative) limiting value of β below which the image is a perfect checkerboard.

A possible approach used in the following code is to resort to simulated annealing, increasing progressively β until all sites are of the same color. Opting for a four-neighbour structure, we slightly modify the R functions

```
xneig4=function(x,a,b,col){
n=dim(x)[1];m=dim(x)[2]
nei=c(x[a-1,b]==col,x[a,b-1]==col)
if (a!=n)
  nei=c(nei,x[a+1,b]==col)
if (b!=m)
  nei=c(nei,x[a,b+1]==col)
sum(nei)
}
```

and

```
isingibbs=function(niter=10^2,n,m=n,beta=1,
   x=matrix(sample(c(-1,1),n*m,rep=TRUE),n,m)){
  for (i in 1:niter){
    sampl1=sample(1:n)
    sampl2=sample(1:m)
    for (k in 1:n){
    for (l in 1:m){
     n0=xneig4(x,sampl1[k],sampl2[l],-1)
     n1=xneig4(x,sampl1[k],sampl2[l],1)
     x[sampl1[k],sampl2[l]]=sample(c(-1,1),1,
                 prob=exp(beta*c(n0,n1)))
     }}}
  x
  }
```

defined in the book. Then the function

```
isinganeal=function(niter=10^3,precis=.1,n,m=n){
  beta=precis
  simu=isingibbs(niter,n,m,beta)
  while (min(simu)<max(simu)){
   beta=beta+precis
   simu=isingibbs(niter,n,m,beta,x=simu)}
  return(beta)
}
```

increases the coefficient β until all simulated entries are of the same color.

Figure 8.3 in this manual provides an histogram of the β's returned by the above code in the case of a 5×5 grid. It gives indications on the zone to study more precisely the occurence of unicolor grids and the detection of the cutoff point.

For the opposite case, the coefficient β is decreased in `isinganeal` until

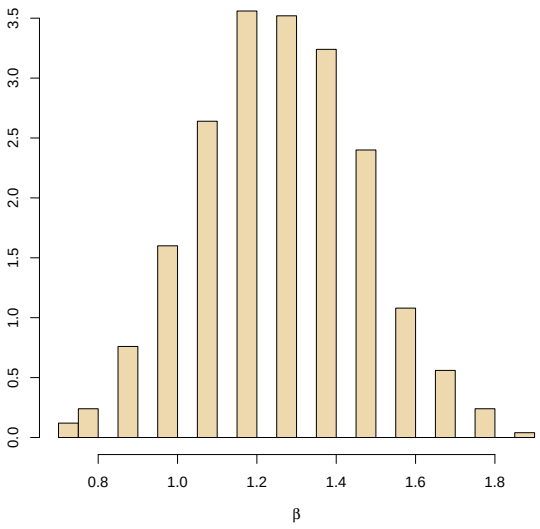

Fig. 8.3. Empirical distribution of the β's leading to a unicolor simulation of the Ising model, for a $(5, 5)$ grid, based on 250 replications and a precision of 0.1.

```
sum(abs(simu[,-1]+simu[,-m]))+sum(abs(simu[-1,]+simu[-n,]))==0
```

Figure 8.4 in this manual provides an histogram of the β's returned by the above code in the case of a 5×5 grid. As for Figure 8.3 in this manual, it only provide some indications on the zone of β's for producing checker grids almost surely.

8.20 Show that the ABC algorithm implemented with $\epsilon = 0$ and a distance between sufficient statistics is not approximate in that the output is truly simulated from the posterior distribution $\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$.

When the ABC algorithm is used with a tolerance $\epsilon = 0$, the probability of accepting $\theta \sim \pi(\theta)$ in Algorithm 8.18 is $\mathbb{P}_\theta(S(Y) = S(x)) = f^S(S(x)|\theta)$, the probability mass function of the statistic $S(X)$ when $X \sim f(x|\theta)$. Therefore the distribution of the accepted θ's is

$$\pi^{\text{ABC}}(\theta|x) \propto \pi(\theta) f^S(S(x)|\theta)$$

which is the *exact* posterior distribution of θ when observing $S(x)$. If $S(\cdot)$ is a sufficient statistic, this posterior is also equal to the posterior distribution of θ given the observation x. Therefore, an ABC simulation of the Potts model

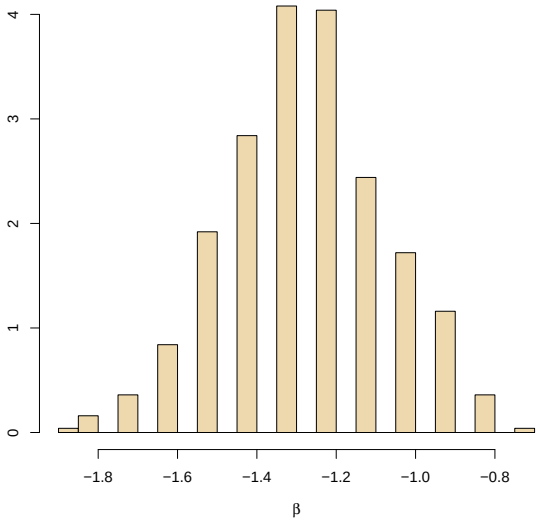

Fig. 8.4. Empirical distribution of the β's leading to a checkerboard simulation of the Ising model, for a $(5, 5)$ grid, based on 250 replications and a precision of 0.1.

posterior in Section 8.3.3 could be rerun with a tolerance of $\epsilon = 0$, albeit at a higher computational cost.

http://www.springer.com/978-1-4614-8686-2

Bayesian Essentials with R

Marin, J.-M.; Robert, C.

2014, XIV, 296 p. 75 illus., 38 illus. in color., Hardcover
ISBN: 978-1-4614-8686-2