Bayesian modelling

Léo Belzile

Table of contents

Welcome

This book is a web complement to MATH 80601A *Bayesian modelling*, a graduate course offered at HEC Montréal. Consult the course webpage for more details.

These notes are licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License and were last compiled on Friday, January 03 2025.

The objective of the course is to provide a hands on introduction to Bayesian data analysis. The course will cover the formulation, evaluation and comparison of Bayesian models through examples and real-data applications.

1 Introduction

This section review basic concepts in probability theory that will be used throughout the course. The overview begins with basic statistical concepts, random variables, their distribution and density, moments and likelihood derivations.

1.1 Random vectors

We begin with a characterization of random vectors and their marginal, conditional and joint distributions. A good reference for this material is Chapter 3 of McNeil, Frey, and Embrechts (2005).

Definition 1.1 (Density and distribution function)**.** Let $\boldsymbol{X}$ denote a d-dimensional vector with real entries in $\mathbb{R}^d$. The distribution function of $\boldsymbol{X}$ is

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \Pr(\boldsymbol{X} \leq \boldsymbol{x}) = \Pr(X_1 \leq x_1, \ldots, X_d \leq x_d).$$

If the distribution of $\boldsymbol{X}$ is absolutely continuous, we may write

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{-\infty}^{x_d} \cdots \int_{-\infty}^{x_1} f_{\boldsymbol{X}}(z_1, \ldots, z_d) \mathrm{d}z_1 \cdots \mathrm{d}z_d,$$

where $f_{\boldsymbol{X}}(\boldsymbol{x})$ is the joint **density function**. The density function can be obtained as the derivative of the distribution function with respect to all of it's arguments.

We use the same notation for the mass function in the discrete case where $f_{\boldsymbol{X}}(\boldsymbol{x}) = \Pr(X_1 = x_1, \ldots, X_d = x_d)$, where the integral is understood to mean a summation over all values lower or equal to $\boldsymbol{x}$ in the support. In the discrete case, $0 \leq f_{\boldsymbol{X}}(\boldsymbol{x}) \leq 1$ is a probability and the total probability over all points in the support sum to one, meaning $\sum_{\boldsymbol{x} \in \mathsf{supp}(\boldsymbol{X})} f_{\boldsymbol{X}}(\boldsymbol{x}) = 1$.

Definition 1.2 (Location and scale distribution)**.** A random variable Y is said to belong to a location scale family with location parameter b and scale $a > 0$ if it is equal in distribution

to a location and scale transformation of a standard variable X with location zero and unit scale, denoted $Y \stackrel{\mathrm{d}}{=} aX + b$ and meaning,

$$\Pr(Y \leq y) = \Pr(aX + b \leq y).$$

If the density exists, then $f_Y(y) = a^{-1} f_X\{(y-b)/a\}$.

We can extend this definition to the multivariate setting for location vector $\boldsymbol{b} \in \mathbb{R}^d$ and positive definite scale matrix $\mathbf{A}$, such that

$$\Pr(\boldsymbol{Y} \leq \boldsymbol{y}) = \Pr(\boldsymbol{b} + \mathbf{A}\boldsymbol{X} \leq \boldsymbol{y}).$$

Definition 1.3 (Exponential family)**.** A univariate distribution is an exponential family if it's density or mass function can be written for all $y \in \mathbb{R}$ as

$$f(y; \boldsymbol{\theta}) = \exp\left\{\sum_{k=1}^{K} Q_k(\boldsymbol{\theta}) t_k(y) + D(\boldsymbol{\theta}) + h(y)\right\},$$

where functions $Q_1(\cdot), \ldots, Q_K(\cdot)$ and $D(\cdot)$ depend only on $\boldsymbol{\theta}$ and not on the data, and conversely $t_1(\cdot), \ldots, t_K(\cdot)$ and $h(\cdot)$ do not depend on the vector of parameters $\boldsymbol{\theta}$.

The support of f must not depend on $\boldsymbol{\theta}$. The transformed parameters $Q_k(\boldsymbol{\theta})$ $(k = 1, \ldots, K)$ are termed canonical parameters.

1.1.1 Common distributions

Definition 1.4 (Gamma, chi-square and exponential distributions)**.** A random variable follows a gamma distribution with shape $\alpha > 0$ and rate $\beta > 0$, denoted $Y \sim \mathsf{gamma}(\alpha, \beta)$, if it's density is

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x), \qquad x \in (0, \infty),$$

where $\Gamma(\alpha) \coloneqq \int_0^\infty t^{\alpha-1} \exp(-t) \mathrm{d}t$ is the gamma function.

If $\alpha = 1$, the density simplifies to $\beta \exp(-\beta x)$ and we recover the **exponential distribution**, denote $\mathsf{expo}(\beta)$. The case $\mathsf{gamma}(\nu/2, 1/2)$ corresponds to the chi-square distribution χ^2_ν.

Definition 1.5 (Beta and uniform distribution)**.** The beta distribution $\mathsf{beta}(\alpha_1, \alpha_2)$ is a distribution supported on the unit interval $[0, 1]$ with shape parameters $\alpha_1 > 0$ and $\alpha_2 > 0$. It's density is

$$f(x) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)} x^{\alpha_1 - 1}(1-x)^{1-\alpha_2}, \qquad x \in [0, 1].$$

The case $\mathsf{beta}(1, 1)$, also denoted $\mathsf{unif}(0, 1)$, corresponds to a standard uniform distribution.

The beta distribution is commonly used to model proportions, and can be generalized to the multivariate setting as follows.

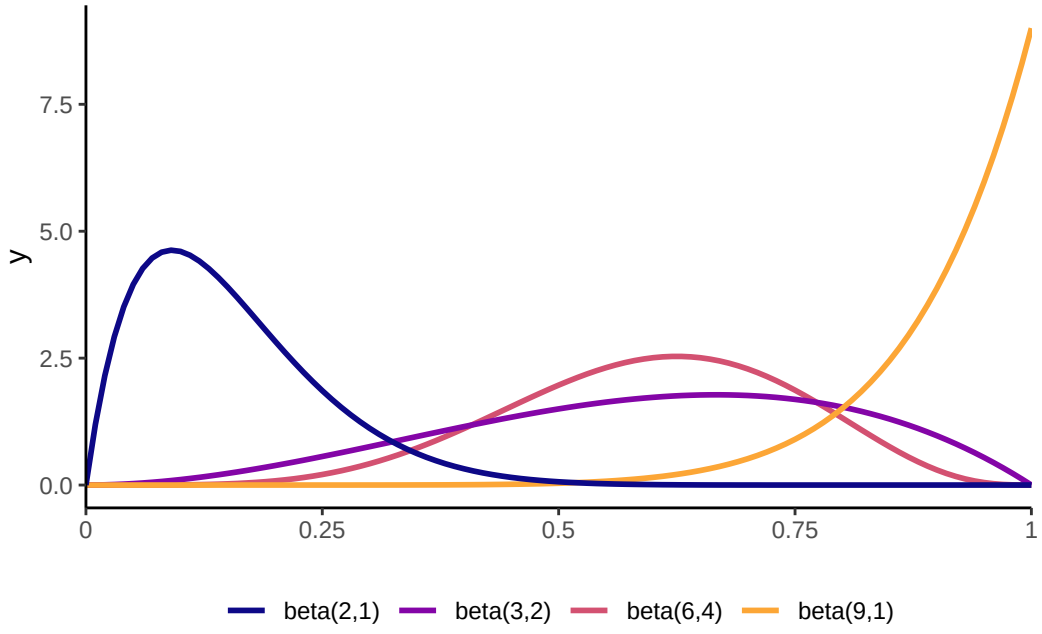

Figure 1.1: Density of beta random variables with different shape parameters

Definition 1.6 (Dirichlet distribution)**.** Let $\boldsymbol{\alpha} \in (0,\infty)^d$ denote shape parameters and cosider a random vector of size d with positive components on the simplex

$$\mathbb{S}_{d-1} : \{0 \leq x_j \leq 1; j = 1,\ldots,d : x_1 + \cdots + x_d = 1\}.$$

The density of a **Dirichlet** random vector, denoted $\boldsymbol{Y} \sim \mathsf{Dirichlet}(\boldsymbol{\alpha})$, is

$$f(\boldsymbol{x}) = \frac{\prod_{j=1}^{d-1} \Gamma(\alpha_j)}{\Gamma(\alpha_1 + \cdots + \alpha_d)} \prod_{j=1}^{d} x_j^{\alpha_j - 1}, \qquad \boldsymbol{x} \in \mathbb{S}_{d-1}$$

Due to the linear dependence, the dth component $x_d = 1 - x_1 - \cdots - x_{d-1}$ is fully determined.

Definition 1.7 (Binomial distribution)**.** The density of the binomial distribution, denoted $Y \sim \mathsf{binom}(n,p)$, is

$$f(x) = \mathsf{Pr}(Y = x) = \binom{m}{x} p^x (1-p)^{m-x}, \quad x = 0,1,\ldots,n.$$

If $n = 1$, we recover the Bernoulli distribution with density $f(x) = p^y(1-p)^{1-y}$. The binomial distribution is closed under convolution, meaning that the number the number of successes Y out of n Bernoulli trials is binomial

Definition 1.8 (Multinomial distribution)**.** If there are more than two outcomes, say d, we can generalize this mass function. Suppose that $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$ denotes the number of realizations of each of the d outcomes based on n trials, so that $0 \leq Y_j \leq n (j = 1, \ldots, d)$ and $Y_1 + \cdots + Y_d = n$. The joint density of the multinomial vector $\boldsymbol{Y} \sim \mathsf{multinom}(\boldsymbol{p})$ with probability vector $\boldsymbol{p} \in \mathbb{S}_{d-1}$ is

$$f(\boldsymbol{x}) = \frac{n!}{\prod_{j=1}^d x_j!} \prod_{j=1}^d p_j^{x_j}, \qquad \boldsymbol{y}/n \in \mathbb{S}_{d-1},$$

where $x! = \Gamma(x+1)$ denotes the factorial function.

Definition 1.9 (Poisson distribution)**.** If the probability of success p of a Bernoulli event is small in the sense that $np \to \lambda$ when the number of trials n increases, then the number of success follows approximately a Poisson distribution with mass function

$$f(x) = \mathsf{Pr}(Y = x) = \frac{\exp(-\lambda)\lambda^y}{\Gamma(y+1)}, \quad x = 0, 1, 2, \ldots$$

where $\Gamma(\cdot)$ denotes the gamma function. The parameter λ of the Poisson distribution is both the expectation and the variance of the distribution, meaning $\mathsf{E}(Y) = \mathsf{Va}(Y) = \lambda$. We denote the distribution as $Y \sim \mathsf{Poisson}(\lambda)$.

The most frequently encountered location-scale distribution is termed normal, but we use the terminology Gaussian in these notes.

Definition 1.10 (Gaussian distribution)**.** Consider a d dimensional vector $\boldsymbol{Y} \sim \mathsf{Gauss}_d(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$ with density

$$f(\boldsymbol{x}) = (2\pi)^{-d/2}|\boldsymbol{Q}|^{1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu})\right\}, \qquad \boldsymbol{x} \in \mathbb{R}^d$$

The mean vector $\boldsymbol{\mu}$ is the vector of expectation of individual observations, whereas $\boldsymbol{Q}^{-1} \equiv \boldsymbol{\Sigma}$ is the $d \times d$ covariance matrix of $\boldsymbol{Y}$ and $\boldsymbol{Q}$, the canonical parameter, is called the precision matrix.

In the univariate case, the density of $\mathsf{Gauss}(\mu, \sigma^2)$ reduces to

$$f(x) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \qquad x \in \mathbb{R}.$$

Proposition 1.1 (Simulation of Gaussian vectors)**.** *The Gaussian distribution is an elliptical distribution and a location-scale family: if* $\boldsymbol{L} = \mathrm{chol}(\boldsymbol{Q})$*, meaning* $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^\top$ *for some lower triangular matrix* $\boldsymbol{L}$*, then*

$$\boldsymbol{L}^\top(\boldsymbol{Y} - \boldsymbol{\mu}) \sim \mathsf{Gauss}_d(\mathbf{0}_d, \mathbf{I}_d).$$

Conversely, we can use the Cholesky root to sample multivariate Gaussian vectors by first drawing d *independent standard Gaussians* $\boldsymbol{Z} = (Z_1, \ldots, Z_d)^\top$*, then computing*

$$\boldsymbol{Y} \leftarrow \boldsymbol{L}^{-1}\boldsymbol{Z} + \boldsymbol{\mu}.$$

Definition 1.11 (Student-t distribution)**.** The name "Student" comes from the pseudonym used by William Gosset in Gosset (1908), who introduced the asymptotic distribution of the t-statistic. The density of the standard Student-t univariate distribution with ν degrees of freedom is

$$f(y;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}}\left(1 + \frac{y^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

The density of the random vector $\boldsymbol{Y} \sim \mathsf{Student}_d(\boldsymbol{\mu}, \boldsymbol{Q}^{-1}, \nu)$, with location vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{Q}^{-1}$ and ν degrees of freedom is

$$f(\boldsymbol{x}) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)|\boldsymbol{Q}|^{1/2}}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{d/2}}\left(1 + \frac{(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{\nu+d}{2}}, \qquad \boldsymbol{x} \in \mathbb{R}^d$$

The Student distribution is a location-scale family and an elliptical distribution. The distribution has polynomial tails, is symmetric around $\boldsymbol{\mu}$ and is unimodal. As $\nu \to \infty$, the Student distribution converges to a normal distribution. It has heavier tails than the normal distribution and only the first $\nu - 1$ moments of the distribution exist. The case $\nu = 1$ is termed Cauchy distribution.

Definition 1.12 (Weibull distribution)**.** The distribution function of a **Weibull** random variable with scale $\lambda > 0$ and shape $\alpha > 0$ is

$$F(x;\lambda,\alpha) = 1 - \exp\left\{-(x/\lambda)^\alpha\right\}, \qquad x \geq 0, \lambda > 0, \alpha > 0,$$

while the corresponding density is

$$f(x;\lambda,\alpha) = \frac{\alpha}{\lambda^\alpha}x^{\alpha-1}\exp\left\{-(x/\lambda)^\alpha\right\}, \qquad x \geq 0, \lambda > 0, \alpha > 0.$$

1.1.2 Marginal and conditional distributions

Definition 1.13 (Marginal distribution)**.** The **marginal distribution** of a subvector $\boldsymbol{X}_{1:k} = (X_1, \ldots, X_k)^\top$, without loss of generality consisting of the k first components of $\boldsymbol{X}$ $(1 \leq k < d)$ is

$$F_{\boldsymbol{X}_{1:k}}(\boldsymbol{x}_{1:k}) = \Pr(\boldsymbol{X}_{1:k} \leq \boldsymbol{x}_{1:k}) = F_{\boldsymbol{X}}(x_1, \ldots, x_k, \infty, \ldots, \infty).$$

and thus the marginal distribution of component j, $F_j(x_j)$, is obtained by evaluating all components but the jth at ∞.

We likewise obtain the marginal density

$$f_{1:k}(\boldsymbol{x}_{1:k}) = \frac{\partial^k F_{1:k}(\boldsymbol{x}_{1:k})}{\partial x_1 \cdots \partial x_k},$$

or through integration from the joint density as

$$f_{1:k}(\boldsymbol{x}_{1:k}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(x_1, \ldots, x_k, z_{k+1}, \ldots, z_d) \mathrm{d}z_{k+1} \cdots \mathrm{d}z_d.$$

Definition 1.14 (Conditional distribution)**.** Let $(\boldsymbol{X}^\top, \boldsymbol{Y}^\top)^\top$ be a d-dimensional random vector with joint density or mass function $f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y})$ and marginal distribution $f_{\boldsymbol{X}}(\boldsymbol{x})$. The conditional distribution function of $\boldsymbol{Y}$ given $\boldsymbol{X} = \boldsymbol{x}$, is

$$f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}; \boldsymbol{x}) = \frac{f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y})}{f_{\boldsymbol{X}}(\boldsymbol{x})}$$

for any value of $\boldsymbol{x}$ in the support of $\boldsymbol{X}$, i.e., the set of values with non-zero density or mass, meaning $f_{\boldsymbol{X}}(\boldsymbol{x}) > 0$; it is undefined otherwise.

Theorem 1.1 (Bayes theorem)**.** *Denote by $p(X) \equiv \Pr(X)$ denotes the marginal density of X, $p(X \mid Y)$ the conditional of X given Y and $p(X, Y)$ the joint density. Bayes' theorem states that*

$$p(X = x \mid Y = y) = \frac{p(Y = y \mid X = x) p(X = x)}{p(Y = y)}$$

In the case of discrete random variable X with support $\mathcal{X}$, the denominator can be evaluated using the law of total probability as

$$\Pr(Y = y) = \sum_{x \in \mathcal{X}} \Pr(Y = y \mid X = x) \Pr(X = x).$$

Example 1.1 (Covid rapid tests)**.** Back in January 2021, the Quebec government was debating whether or not to distribute antigen rapid test, with strong reluctance from authorities given the paucity of available resources and the poor sensitivity.

A Swiss study analyse the efficiency of rapid antigen tests, comparing them to repeated polymerase chain reaction (PCR) test output, taken as benchmark (Jegerlehner et al. 2021). The results are presented in Table 1.1

Table 1.1: Confusion matrix of Covid test results for PCR tests versus rapid antigen tests, from Jegerlehner et al. (2021).

	PCR +	PCR −
rapid +	92	2
rapid −	49	1319
total	141	1321

Estimated seropositivity at the end of January 2021 according to projections of the Institute for Health Metrics and Evaluation (IHME) of 8.18M out of 38M inhabitants (Mathieu et al. 2020), a prevalence of 21.4%. Assuming the latter holds uniformly over the country, what is the probability of having Covid if I get a negative result to a rapid test?

Let R^- (R^+) denote a negative (positive) rapid test result and C^+ (C^-) Covid positivity (negativity). Bayes' formula gives

$$\begin{aligned}\Pr(C^+ \mid R^-) &= \frac{\Pr(R^- \mid C^+)\Pr(C^+)}{\Pr(R^- \mid C^+)\Pr(C^+) + \Pr(R^- \mid C^-)\Pr(C^-)} \\ &= \frac{49/141 \cdot 0.214}{49/141 \cdot 0.214 + 1319/1321 \cdot 0.786}\end{aligned}$$

so there is a small, but non-negligible probability of 8.66% that the rapid test result is misleading. Jegerlehner et al. (2021) indeed found that the sensitivity was 65.3% among symptomatic individuals, but dropped down to 44% for asymptomatic cases. This may have fueled government experts skepticism.

Bayes' rule is central to updating beliefs: given initial beliefs (priors) and information in the form of data, we update our beliefs. We can apply

Example 1.2 (Conditional and marginal for contingency table)**.** Consider a bivariate distribution for (Y_1, Y_2) supported on $\{1,2,3\} \times \{1,2\}$, whose joint probability mass function is given in Table 1.2

Table 1.2: Bivariate mass function with probability of each outcome for (Y_1, Y_2).

	$Y_1 = 1$	$Y_1 = 2$	$Y_1 = 3$	total
$Y_2 = 1$	0.20	0.3	0.10	0.6
$Y_2 = 2$	0.15	0.2	0.05	0.4
total	0.35	0.5	0.15	1.0

The marginal distribution of Y_1 is obtain by looking at the total probability for each column, as

$$\Pr(Y_1 = i) = \Pr(Y_1 = i, Y_2 = 1) + \Pr(Y_1 = i, Y_2 = 2).$$

This gives $\Pr(Y_1 = 1) = 0.35$, $\Pr(Y_1 = 2) = 0.5$ and $\Pr(Y_1 = 3) = 0.15$. Similarly, we find that $\Pr(Y_2 = 1) = 0.6$ and $\Pr(Y_2 = 2) = 0.4$ for the other random variable.

The conditional distribution

$$\Pr(Y_2 = i \mid Y_1 = 2) = \frac{\Pr(Y_1 = 2, Y_2 = i)}{\Pr(Y_1 = 2)},$$

so $\Pr(Y_2 = 1 \mid Y_1 = 2) = 0.3/0.5 = 0.6$ and $\Pr(Y_2 = 2 \mid Y_1 = 2) = 0.4$. We can condition on more complicated events, for example

$$\Pr(Y_2 = i \mid Y_1 \geq 2) = \frac{\Pr(Y_1 = 2, Y_2 = i) + \Pr(Y_1 = 3, Y_2 = i)}{\Pr(Y_1 = 2) + \Pr(Y_1 = 3)}.$$

Example 1.3 (Margins and conditional distributions of multinomial vectors)**.** Consider $\boldsymbol{Y} = (Y_1, Y_2, n - Y_1 - Y_2)$ a trinomial vector giving the number of observations in group $j \in \{1, 2, 3\}$ with n trials and probabilities of each component respectively $(p_1, p_2, 1-p_1-p_2)$. The marginal distribution of Y_2 is obtained by summing over all possible values of Y_1, which ranges from 0 to n, so

$$f(y_2) = \frac{n! p_2^{y_2}}{y_2!} \sum_{y_1=1}^{n} \frac{p_1^{y_1}(1 - p_1 - p_2)^{n-y_1-y_2}}{y_1!(n - y_1 - y_2)!}$$

A useful trick is to complete the expression on the right so that it sum (in the discrete case) or integrate (in the continuous case) to 1. If we multiply and divide by $(1 - p_2)^{n-y_2}/(n - y_2)!$, we get $p_1^* = p_1/(1 - p_2)$ and

$$\begin{aligned} f(y_2) &= \frac{n! p_2^{y_2}}{(1 - p_2)^{n-y_2} y_2!(n - y_2)!} \sum_{y_1=1}^{n} \binom{n - y_2}{y_1} p_1^{\star y_1}(1 - p_1^\star)^{n-y_2} \\ &= \frac{n! p_2^{y_2}}{(1 - p_2)^{n-y_2} y_2!(n - y_2)!} \end{aligned}$$

is binomial with n trials and probability of success p_2. We can generalize this argument to multinomials of arbitrary dimensions.

The conditional density of $Y_2 \mid Y_1 = y_1$ is, up to proportionality,

$$f_{Y_2|Y_1}(y_2; y_1) \propto \frac{p_2^{y_2}(1-p_1-p_2)^{n-y_1-y_2}}{y_2!(n-y_1-y_2)!}$$

If we write $p_2^\star = p_2/(1-p_1)$, we find that $Y_2 \mid Y_1 \sim \mathsf{binom}(n-y_1, p_2^\star)$. Indeed, we can see that

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}) &= f_{Y_2|Y_1}(y_2; y_1) f_{Y_1}(y_1) \\ &= \binom{n-y_1}{y_2} \left(\frac{p_2}{1-p_1}\right)^{y_2} \left(\frac{1-p_1-p_2}{1-p_1}\right)^{n-y_1-y_2} \cdot \binom{n}{y_1} p_1^{y_1}(1-p_1)^{n-y_1}. \end{aligned}$$

Example 1.4. Consider the bivariate density function of the pair (X, Y), where for $\lambda > 0$,

$$f(x,y) = \frac{\lambda y^{1/2}}{(2\pi)^{1/2}} \exp\left\{-y(x^2+\lambda)\right\}, \qquad x \in \mathbb{R}, y > 0.$$

We see that the conditional distribution of $X \mid Y = y \sim \mathsf{Gauss}(0, y^{-1})$. The marginals are

$$\begin{aligned} f(y) &= \int_{-\infty}^{\infty} \frac{\lambda y^{1/2}}{(2\pi)^{1/2}} \exp\left\{-y(x^2+\lambda)\right\} \mathrm{d}x \\ &= \lambda \exp(-\lambda y) \end{aligned}$$

so marginally Y follows an exponential distribution with rate λ. The marginal of X can be obtained by noting that the joint distribution, as a function of y, is proportional to the kernel of a gamma distribution with shape $3/2$ and rate $x^2+\lambda$, with $Y \mid X = x \sim \mathsf{gamma}(3/2, x^2+\lambda)$. If we pull out the normalizing constant, we find

$$\begin{aligned} f(x) &= \int_0^{\infty} \frac{\lambda y^{1/2}}{(2\pi)^{1/2}} \exp\left\{-y(x^2+\lambda)\right\} \mathrm{d}y \\ &= \frac{\lambda \Gamma(3/2)}{(2\pi)^{1/2}(x^2+\lambda)^{3/2}} \int_0^{\infty} f_{Y|X}(y \mid x) \mathrm{d}y \\ &= \frac{\lambda}{2^{3/2}(x^2+\lambda)^{3/2}} \end{aligned}$$

since $\Gamma(a+1) = a\Gamma(a)$ for $a > 0$ and $\Gamma(1/2) = \sqrt{\pi}$. We conclude that marginally $X \sim \mathsf{Student}(0, \lambda, 2)$, a Student distribution with scale λ and two degrees of freedom.

Example 1.5 (Bivariate geometric distribution of Marshall and Olkin)**.** Consider a couple (U_1, U_2) of Bernoulli random variables whose mass function is $\Pr(U_1 = i, U_2 = j = p_{ij}$ for $(i, j) \in \{0, 1\}^2$. The marginal distributions are, by the law of total probability

$$\begin{aligned}\Pr(U_1 = i) &= \Pr(U_1 = i, U_2 = 0) + \Pr(U_1 = i, U_2 = 1) = p_{i0} + p_{i1} = p_{i\bullet}\\ \Pr(U_2 = j) &= \Pr(U_1 = 0, U_2 = j) + \Pr(U_1 = 1, U_2 = j) = p_{0j} + p_{1j} = p_{\bullet j}\end{aligned}$$

We consider a joint geometric distribution (Marshall and Olkin (1985), Section 6) and the pair (Y_1, Y_2) giving the number of zeros for (U_1, U_2) before the variable equals one for the first time. The bivariate mass function is (Nadarajah 2008)

$$\Pr(Y_1 = k, Y_2 = l) = \begin{cases} p_{00}^k p_{01} p_{0\bullet}^{l-k-1} p_{1\bullet} & 0 \leq k < l; \\ p_{00}^k p_{11} & k = l; \\ p_{00}^l p_{10} p_{\bullet 0}^{k-l-1} p_{\bullet 1} & 0 \leq l < k. \end{cases}$$

We can compute the joint survival function $\Pr(Y_1 \geq k, Y_2 \geq l)$ by using properties of the partial sum of geometric series, using the fact $\sum_{i=0}^n p^i = p^n/(1-p)$. Thus, for the case $0 \leq k < l$, we have

$$\begin{aligned}\Pr(Y_1 \geq k, Y_2 \geq l) &= \sum_{i=k}^{\infty}\sum_{j=l}^{\infty} \Pr(Y_1 = i, Y_2 = j)\\ &= \sum_{i=k}^{\infty} p_{00}^i p_{01} p_{0\bullet}^{-i-1} p_{1\bullet} \sum_{j=l}^{\infty} p_{0\bullet}^j\\ &= \sum_{i=k}^{\infty} p_{00}^i p_{01} p_{0\bullet}^{-i-1} p_{1\bullet} \frac{p_{0\bullet}^l}{1-p_{0\bullet}}\\ &= p_{0\bullet}^{l-1} p_{01} \sum_{i=k}^{\infty} \left(\frac{p_{00}}{p_{0\bullet}}\right)^i\\ &= p_{00}^k p_{0\bullet}^{l-k}\end{aligned}$$

since $p_{0\bullet} + p_{1\bullet} = 1$. We can proceed similarly with other subcases to find

$$\Pr(Y_1 \geq k, Y_2 \geq l) = \begin{cases} p_{00}^k p_{0\bullet}^{l-k} & 0 \leq k < l \\ p_{00}^k & 0 \leq k = l \\ p_{00}^l p_{\bullet 0}^{k-l} & 0 \leq l < k \end{cases}$$

and we can obtain the marginal survival function by considering $\Pr(Y_1 \geq 0, Y_2 \geq l)$, etc., which yields $\Pr(Y_2 \geq l) = p_{0\bullet}^l$, whence

$$\begin{aligned}\Pr(Y_2 = l) &= \Pr(Y_2 \geq l) - \Pr(Y_2 \geq l+1)\\ &= p_{0\bullet}^l(1 - p_{0\bullet})\\ &= p_{0\bullet}^l p_{1\bullet}\end{aligned}$$

and so both margins are geometric.

Definition 1.15 (Independence)**.** We say that $\boldsymbol{Y}$ and $\boldsymbol{X}$ are independent if their joint distribution function factorizes as

$$F_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) = F_{\boldsymbol{X}}(\boldsymbol{x})F_{\boldsymbol{Y}}(\boldsymbol{y})$$

for any value of $\boldsymbol{x}$, $\boldsymbol{y}$. It follows from the definition of joint density that, should the latter exists, it also factorizes as

$$f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) = f_{\boldsymbol{X}}(\boldsymbol{x})f_{\boldsymbol{Y}}(\boldsymbol{y}).$$

If two subvectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, then the conditional density $f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y};\boldsymbol{x})$ equals the marginal $f_{\boldsymbol{Y}}(\boldsymbol{y})$.

Proposition 1.2 (Gaussian vectors, independence and conditional independence properties)**.**

A unique property of the multivariate normal distribution is the link between independence and the covariance matrix: components Y_i and Y_j are independent if and only if the (i,j) off-diagonal entry of the covariance matrix $\boldsymbol{Q}^{-1}$ is zero.

If $q_{ij} = 0$, then Y_i and Y_j are conditionally independent given the other components.

Proposition 1.3 (Law of iterated expectation and variance)**.** *Let $\boldsymbol{Z}$ and $\boldsymbol{Y}$ be random vectors. The expected value of $\boldsymbol{Y}$ is*

$$\mathsf{E}_{\boldsymbol{Y}}(\boldsymbol{Y}) = \mathsf{E}_{\boldsymbol{Z}}\left\{\mathsf{E}_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{Y})\right\}.$$

*The **tower** property gives a law of iterated variance*

$$\mathsf{Va}_{\boldsymbol{Y}}(\boldsymbol{Y}) = \mathsf{E}_{\boldsymbol{Z}}\left\{\mathsf{Va}_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{Y})\right\} + \mathsf{Va}_{\boldsymbol{Z}}\left\{\mathsf{E}_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{Y})\right\}.$$

In a hierarchical model, the variance of the unconditional distribution is larger than that of the conditional distribution.

Example 1.6. Let $Y \mid X \sim \mathsf{Gauss}(X, \sigma^2)$ and $X \sim \mathsf{Gauss}(0, \tau^2)$. The unconditional mean and variance of Y are

$$\mathsf{E}(Y) = \mathsf{E}_X\{\mathsf{E}_{Y|X}(Y)\} = \mathsf{E}_X(X) = 0$$

and

$$\begin{aligned}\mathsf{Va}(Y) &= \mathsf{E}_X\{\mathsf{Va}_{Y|X}(Y)\} + \mathsf{Va}_X\{\mathsf{E}_{Y|X}(Y)\} \\ &= \mathsf{E}_X(\sigma^2) + \mathsf{Va}_X(X) \\ &= \sigma^2 + \tau^2\end{aligned}$$

Example 1.7 (Negative binomial as a Poisson mixture)**.**

One restriction of the Poisson model is that the restriction on its moments is often unrealistic. The most frequent problem encountered is that of **overdispersion**, meaning that the variability in the counts is larger than that implied by a Poisson distribution.

One common framework for handling overdispersion is to have $Y \mid \Lambda = \lambda \sim \mathsf{Poisson}(\lambda)$, where the mean of the Poisson distribution is itself a positive random variable with mean μ, if Λ follows a gamma distribution with shape $k\mu$ and rate $k > 0$, $\Lambda \sim \mathsf{gamma}(k\mu, k)$. Since the joint density of Y and Λ can be written

$$\begin{aligned}p(y, \lambda) &= p(y \mid \lambda)p(\lambda) \\ &= \frac{\lambda^y \exp(-\lambda)}{\Gamma(y+1)} \frac{k^{k\mu}\lambda^{k\mu-1}\exp(-k\lambda)}{\Gamma(k\mu)}\end{aligned}$$

so the conditional distribution of $\Lambda \mid Y = y$ can be found by considering only terms that are function of λ, whence

$$f(\lambda \mid Y = y) \overset{\lambda}{\propto} \lambda^{y+k\mu-1}\exp(-(k+1)\lambda)$$

and the conditional distribution is $\Lambda \mid Y = y \sim \mathsf{gamma}(k\mu + y, k+1)$.

We can isolate the marginal density

$$\begin{aligned}p(y) &= \frac{p(y,\lambda)}{p(\lambda \mid y)} \\ &= \frac{\frac{\lambda^y \exp(-\lambda)}{\Gamma(y+1)} \frac{k^{k\mu}\lambda^{k\mu-1}\exp(-k\lambda)}{\Gamma(k\mu)}}{\frac{(k+1)^{k\mu+y}\lambda^{k\mu+y-1}\exp\{-(k+1)\lambda\}}{\Gamma(k\mu+y)}} \\ &= \frac{\Gamma(k\mu+y)}{\Gamma(k\mu)\Gamma(y+1)} k^{k\mu}(k+1)^{-k\mu-y} \\ &= \frac{\Gamma(k\mu+y)}{\Gamma(k\mu)\Gamma(y+1)} \left(1 - \frac{1}{k+1}\right)^{k\mu} \left(\frac{1}{k+1}\right)^y\end{aligned}$$

and this is the density of a negative binomial distribution with probability of success $1/(k+1)$. We can thus view the negative binomial as a Poisson mean mixture.

By the laws of iterated expectation and iterative variance,

$$\begin{aligned}\mathsf{E}(Y) &= \mathsf{E}_\Lambda\{\mathsf{E}(Y \mid \Lambda\} \\ &= \mathsf{E}(\Lambda) = \mu \\ \mathsf{Va}(Y) &= \mathsf{E}_\Lambda\{\mathsf{Va}(Y \mid \Lambda)\} + \mathsf{Va}_\Lambda\{\mathsf{E}(Y \mid \Lambda)\} \\ &= \mathsf{E}(\Lambda) + \mathsf{Va}(\Lambda) \\ &= \mu + \mu/k.\end{aligned}$$

The marginal distribution of Y, unconditionally, has a variance which exceeds its mean, as

$$\mathsf{E}(Y) = \mu, \qquad \mathsf{Va}(Y) = \mu(1 + 1/k).$$

In a negative binomial regression model, the term k is a dispersion parameter, which is fixed for all observations, whereas $\mu = \exp(\boldsymbol{\beta}\mathbf{X})$ is a function of covariates $\mathbf{X}$. As $k \to \infty$, the distribution of Λ degenerates to a constant at μ and we recover the Poisson model.

Proposition 1.4 (Partitioning of covariance matrices)**.** *Let $\boldsymbol{\Sigma}$ be a $d \times d$ positive definite covariance matrix. We define the precision matrix $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$. Suppose the matrices are partitioned into blocks,*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \text{ and } \boldsymbol{\Sigma}^{-1} = \boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{pmatrix}$$

with $\dim(\boldsymbol{\Sigma}_{11}) = k \times k$ and $\dim(\boldsymbol{\Sigma}_{22}) = (d-k) \times (d-k)$. One can show the following relationships:

- $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} = -\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}$
- $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{Q}_{11}^{-1}$
- $\det(\boldsymbol{\Sigma}) = \det(\boldsymbol{\Sigma}_{22})\det(\boldsymbol{\Sigma}_{1|2})$ *where* $\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.

Proposition 1.5 (Conditional distribution of Gaussian vectors)**.** *Let $\boldsymbol{Y} \sim \mathsf{Gauss}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and consider the partition*

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

where $\boldsymbol{Y}_1$ is a $k \times 1$ and $\boldsymbol{Y}_2$ is a $(d-k) \times 1$ vector for some $1 \leq k < d$. Then, we have the conditional distribution

$$\begin{aligned}\boldsymbol{Y}_1 \mid \boldsymbol{Y}_2 = \boldsymbol{y}_2 &\sim \mathsf{Gauss}_k(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{1|2}) \\ &\sim \mathsf{Gauss}_k(\boldsymbol{\mu}_1 - \boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2), \boldsymbol{Q}_{11}^{-1})\end{aligned}$$

and $\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ is the Schur complement of $\boldsymbol{\Sigma}_{22}$.

Proof. It is easier to obtain this result by expressing the density of the Gaussian distribution in terms of the precision matrix $\boldsymbol{Q} = \left(\begin{smallmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{smallmatrix}\right)$ rather than in terms of the covariance matrix $\boldsymbol{\Sigma}$.

Consider the partition $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2)$. The log conditional density $\log f(\boldsymbol{y}_1 \mid \boldsymbol{y}_2)$ as a function of $\boldsymbol{y}_1$ is, up to proportionality,

$$
\begin{aligned}
&-\frac{1}{2}\left(\boldsymbol{y}_1-\boldsymbol{\mu}_1\right)^{\top}\boldsymbol{Q}_{11}\left(\boldsymbol{y}_1-\boldsymbol{\mu}_1\right)-\left(\boldsymbol{y}_1-\boldsymbol{\mu}_1\right)^{\top}\boldsymbol{Q}_{12}\left(\boldsymbol{y}_2-\boldsymbol{\mu}_2\right) \\
&-\frac{1}{2}\boldsymbol{y}_1^{\top}\boldsymbol{Q}_{11}\boldsymbol{y}_1-\boldsymbol{y}_1^{\top}\left\{\boldsymbol{Q}_{11}\boldsymbol{\mu}_1-\boldsymbol{Q}_{12}\left(\boldsymbol{y}_2-\boldsymbol{\mu}_2\right)\right\}
\end{aligned}
$$

upon completing the square in $\boldsymbol{y}_1$. This integrand is proportional to the density of a Gaussian distribution (and hence must be Gaussian) with precision matrix $\boldsymbol{Q}_{11}$, while the mean vector and covariance matrix are

$$
\begin{aligned}
\boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1-\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\left(\boldsymbol{y}_2-\boldsymbol{\mu}_2\right) \\
&= \boldsymbol{\mu}_1+\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\boldsymbol{y}_2-\boldsymbol{\mu}_2\right) \\
\boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.
\end{aligned}
$$

Note that $\boldsymbol{\Sigma}_{1|2} = \boldsymbol{Q}_{11}^{-1}$ corresponds to the Schur complement of $\boldsymbol{\Sigma}_{22}$.

Remark that the above is sufficient (why?) The quadratic form appearing in the exponential term of the density of a Gaussian vector with mean $\boldsymbol{\nu}$ and precision $\boldsymbol{\Psi}$ is

$$
(\boldsymbol{x}-\boldsymbol{\nu})^{\top}\boldsymbol{\Psi}(\boldsymbol{x}-\boldsymbol{\nu}) = \boldsymbol{x}^{\top}\boldsymbol{\Psi}\boldsymbol{x}-\boldsymbol{x}^{\top}\boldsymbol{\Psi}\boldsymbol{\nu}-\boldsymbol{\nu}^{\top}\boldsymbol{\Psi}\boldsymbol{x}+\boldsymbol{\nu}^{\top}\boldsymbol{\Psi}\boldsymbol{\nu}.
$$

uniquely determines the parameters of the Gaussian distribution. The quadratic term in $\boldsymbol{x}$ forms a sandwich around the precision matrix, while the linear term identifies the location vector. Since any (conditional) density function integrates to one, there is a unique normalizing constant and the latter need not be computed.

□

1.2 Likelihood

Definition 1.16 (Likelihood)**.** The **likelihood** $L(\boldsymbol{\theta})$ is a function of the parameter vector $\boldsymbol{\theta}$ that gives the probability (or density) of observing a sample under a postulated distribution, treating the observations as fixed,

$$
L(\boldsymbol{\theta}; \boldsymbol{y}) = f(\boldsymbol{y}; \boldsymbol{\theta}),
$$

where $f(\boldsymbol{y}; \boldsymbol{\theta})$ denotes the joint density or mass function of the n-vector containing the observations.

If the latter are independent, the joint density factorizes as the product of the density of individual observations, and the likelihood becomes

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{n} f_i(y_i; \boldsymbol{\theta}) = f_1(y_1; \boldsymbol{\theta}) \times \cdots \times f_n(y_n; \boldsymbol{\theta}).$$

The corresponding log likelihood function for independent and identically distributions observations is

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \ln f(y_i; \boldsymbol{\theta})$$

Definition 1.17 (Score and information matrix)**.** Let $\ell(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$, be the log likelihood function. The gradient of the log likelihood $U(\boldsymbol{\theta}) = \partial \ell(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ is termed **score** function.

The **observed information matrix** is the hessian of the negative log likelihood

$$j(\boldsymbol{\theta}; \boldsymbol{y}) = -\frac{\partial^2 \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top},$$

evaluated at the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$, so $j(\widehat{\boldsymbol{\theta}})$. Under regularity conditions, the **expected information**, also called **Fisher information** matrix, is

$$i(\boldsymbol{\theta}) = \mathsf{E}\left\{U(\boldsymbol{\theta}; \boldsymbol{Y}) U(\boldsymbol{\theta}; \boldsymbol{Y})^\top\right\} = \mathsf{E}\left\{j(\boldsymbol{\theta}; \boldsymbol{Y})\right\}$$

Both the Fisher (or expected) and the observed information matrices are symmetric and encode the curvature of the log likelihood and provide information about the variability of $\widehat{\boldsymbol{\theta}}$.

The information of an independent and identically distributed sample of size n is n times that of a single observation, so information accumulates at a linear rate.

Example 1.8 (Likelihood for right-censoring)**.** Consider a survival analysis problem for independent time-to-event data subject to (noninformative) random right-censoring. We assume failure times $Y_i (i = 1, \ldots, n)$ are drawn from a common distribution $F(\cdot; \boldsymbol{\theta})$ supported on $(0, \infty)$ and complemented with an independent censoring indicator $C_i \in \{0, 1\}$, with 0 indicating right-censoring and $C_i = 1$ observed failure time. If individual observation i has not experienced the event at the end of the collection period, then the likelihood contribution $\Pr(Y > y) = 1 - F(y; \boldsymbol{\theta})$, where y_i is the maximum time observed for Y_i.

We write the log likelihood in terms of the right-censoring binary indicator as

$$\ell(\boldsymbol{\theta}) = \sum_{i:c_i=0} \log\{1 - F(y_i; \boldsymbol{\theta})\} + \sum_{i:c_i=1} \log f(y_i; \boldsymbol{\theta})$$

Suppose for simplicity that $Y_i \sim \mathsf{expo}(\lambda)$ and let $m = c_1 + \cdots + c_n$ denote the number of observed failure times. Then, the log likelihood and the Fisher information are

$$\begin{aligned}\ell(\lambda) &= \lambda \sum_{i=1}^{n} y_i + \log \lambda m \\ i(\lambda) &= m/\lambda^2\end{aligned}$$

and the right-censored observations for the exponential model do not contribute to the information.

Example 1.9 (Information for the Gaussian distribution)**.** Consider $Y \sim \mathsf{Gauss}(\mu, \tau^{-1})$, parametrized in terms of precision τ. The likelihood contribution for an n sample is, up to proportionality,

$$\ell(\mu, \tau) \propto \frac{n}{2} \log(\tau) - \frac{\tau}{2} \sum_{i=1}^{n} (Y_i^2 - 2\mu Y_i + \mu^2)$$

The observed and Fisher information matrices are is

$$\begin{aligned}j(\mu, \tau) &= \begin{pmatrix} n\tau & -\sum_{i=1}^{n}(Y_i - \mu) \\ -\sum_{i=1}^{n}(Y_i - \mu) & \frac{n}{2\tau^2} \end{pmatrix}, \\ i(\mu, \tau) &= n \begin{pmatrix} \tau & 0 \\ 0 & \frac{1}{2\tau^2} \end{pmatrix}\end{aligned}$$

Since $\mathsf{E}(Y_i) = \mu$, the expected value of the off-diagonal entries of the Fisher information matrix are zero.

Example 1.10 (Likelihood, score and information of the Weibull distribution)**.** The log likelihood for a simple random sample whose realizations are $y_1, \ldots, y_n$ of size n from a $\mathsf{Weibull}(\lambda, \alpha)$ model is

$$\ell(\lambda, \alpha) = n \ln(\alpha) - n\alpha \ln(\lambda) + (\alpha - 1) \sum_{i=1}^{n} \ln y_i - \lambda^{-\alpha} \sum_{i=1}^{n} y_i^{\alpha}.$$

The score, which is the gradient of the log likelihood, is easily obtained by differentiation[1]

$$\begin{aligned} U(\lambda,\alpha) &= \begin{pmatrix} \frac{\partial \ell(\lambda,\alpha)}{\partial \lambda} \\ \frac{\partial \ell(\lambda,\alpha)}{\partial \alpha} \end{pmatrix} \\ &= \begin{pmatrix} -\frac{n\alpha}{\lambda} + \alpha\lambda^{-\alpha-1}\sum_{i=1}^n y_i^\alpha \\ \frac{n}{\alpha} + \sum_{i=1}^n \ln(y_i/\lambda) - \sum_{i=1}^n \left(\frac{y_i}{\lambda}\right)^\alpha \times \ln\left(\frac{y_i}{\lambda}\right). \end{pmatrix} \end{aligned}$$

and the observed information is the 2×2 matrix-valued function

$$j(\lambda,\alpha) = -\begin{pmatrix} \frac{\partial^2 \ell(\lambda,\alpha)}{\partial \lambda^2} & \frac{\partial^2 \ell(\lambda,\alpha)}{\partial \lambda \partial \alpha} \\ \frac{\partial^2 \ell(\lambda,\alpha)}{\partial \alpha \partial \lambda} & \frac{\partial^2 \ell(\lambda,\alpha)}{\partial \alpha^2} \end{pmatrix} = \begin{pmatrix} j_{\lambda,\lambda} & j_{\lambda,\alpha} \\ j_{\lambda,\alpha} & j_{\alpha,\alpha} \end{pmatrix}$$

whose entries are

$$\begin{aligned} j_{\lambda,\lambda} &= \lambda^{-2}\left\{-n\alpha + \alpha(\alpha+1)\sum_{i=1}^n (y_i/\lambda)^\alpha\right\} \\ j_{\lambda,\alpha} &= \lambda^{-1}\sum_{i=1}^n [1 - (y_i/\lambda)^\alpha\{1 + \alpha\ln(y_i/\lambda)\}] \\ j_{\alpha,\alpha} &= n\alpha^{-2} + \sum_{i=1}^n (y_i/\lambda)^\alpha \{\ln(y_i/\lambda)\}^2 \end{aligned}$$

To compute the expected information matrix, we need to compute expectation of $\mathsf{E}\{(Y/\lambda)^\alpha\}$, $\mathsf{E}[(Y/\lambda)^\alpha \ln\{(Y/\lambda)^\alpha\}]$ and $\mathsf{E}[(Y/\lambda)^\alpha \ln^2\{(Y/\lambda)^\alpha\}]$. By definition,

$$\begin{aligned} \mathsf{E}\left\{(Y/\lambda)^\alpha\right\} &= \int_0^\infty (x/\lambda)^\alpha \frac{\alpha}{\lambda^\alpha} x^{\alpha-1} \exp\left\{-(x/\lambda)^\alpha\right\} \mathrm{d}x \\ &= \int_0^\infty s\exp(-s)\mathrm{d}s = 1 \end{aligned}$$

making a change of variable $S = (Y/\lambda)^\alpha \sim \mathsf{Exp}(1)$. The two other integrals are tabulated in Gradshteyn and Ryzhik (2014), and are equal to $1-\gamma$ and $\gamma^2 - 2\gamma + \pi^2/6$, respectively, where $\gamma \approx 0.577$ is the Euler–Mascherroni constant. The expected information matrix of the Weibull distribution has entries

$$\begin{aligned} i_{\lambda,\lambda} &= n\lambda^{-2}\alpha\left\{(\alpha+1) - 1\right\} \\ i_{\lambda,\alpha} &= -n\lambda^{-1}(1-\gamma) \\ i_{\alpha,\alpha} &= n\alpha^{-2}(1 + \gamma^2 - 2\gamma + \pi^2/6) \end{aligned}$$

We can check this result numerically by comparing the expected value of the observed information matrix

[1] Using for example a symbolic calculator.

```
exp_info_weib <- function(scale, shape){
  i11 <- shape*((shape + 1) - 1)/(scale^2)
  i12 <- -(1+digamma(1))/scale
  i22 <- (1+digamma(1)^2+2*digamma(1)+pi^2/6)/(shape^2)
  matrix(c(i11, i12, i12, i22), nrow = 2, ncol = 2)
}
obs_info_weib <- function(y, scale, shape){
  ys <- y/scale # scale family
  o11 <- shape*((shape + 1)*mean(ys^shape)-1)/scale^2
  o12 <- (1-mean(ys^shape*(1+shape*log(ys))))/scale
  o22 <- 1/(shape*shape) + mean(ys^shape*(log(ys))^2)
  matrix(c(o11, o12, o12, o22), nrow = 2, ncol = 2)
}
nll_weib <- function(pars, y){
  -sum(dweibull(x = y, scale = pars[1], shape = pars[2], log = TRUE))}
# Fix parameters
scale <- rexp(n = 1, rate = 0.5)
shape <- rexp(n = 1)
nobs <- 1000L
dat <- rweibull(n = nobs, scale = scale, shape = shape)
# Compare Hessian with numerical differentiation
o_info <- obs_info_weib(dat, scale = scale, shape = shape)
all.equal(
  numDeriv::hessian(nll_weib, x = c(scale, shape), y = dat) / nobs,
  o_info)
# Compute approximation to Fisher information
exp_info_sim <- replicate(n = 1000, expr = {
  obs_info_weib(y = rweibull(n = nobs,
                         shape = shape,
                         scale = scale),
           scale = scale, shape = shape)})
all.equal(apply(exp_info_sim, 1:2, mean),
          exp_info_weib(scale, shape))
```

The joint density function only factorizes for independent data, but an alternative sequential decomposition can be helpful. For example, we can write the joint density $f(y_1, \ldots, y_n)$ using the factorization

$$f(\boldsymbol{y}) = f(y_1) \times f(y_2 \mid y_1) \times \ldots f(y_n \mid y_1, \ldots, y_n)$$

in terms of conditional. Such a decomposition is particularly useful in the context of time series, where data are ordered from time 1 until time n and models typically relate observation y_n to it's past.

Example 1.11 (First-order autoregressive process)**.** Consider a Gaussian $\mathsf{AR}(1)$ model of the form

$$Y_t = \mu + \phi(Y_{t-1} - \mu) + \varepsilon_t,$$

where ϕ is the lag-one correlation, μ the global mean and ε_t is an iid innovation with mean zero and variance σ^2. If $|\phi| < 1$, the process is stationary, and the variance does not increase with t.

The Markov property states that the current realization depends on the past, $Y_t \mid Y_1, \ldots, Y_{t-1}$, only through the most recent value Y_{t-1}. The log likelihood thus becomes

$$\ell(\boldsymbol{\theta}) = \ln f(y_1) + \sum_{i=2}^{n} f(y_i \mid y_{i-1}).$$

If innovations are Gaussian, we have

$$Y_t \mid Y_{t-1} = y_{t-1} \sim \mathsf{Gauss}\{\mu(1-\phi) + \phi y_{t-1}, \sigma^2\}, \qquad t > 1.$$

The $\mathsf{AR}(1)$ stationarity process Y_t, marginally, has mean μ and unconditional variance $\sigma^2/(1-\phi^2)$. The $\mathsf{AR}(1)$ process is first-order Markov since the conditional distribution $p(Y_t \mid Y_{t-1}, \ldots, Y_{t-p})$ equals $p(Y_t \mid Y_{t-1})$. This means the likelihood is

$$\begin{aligned}\ell(\mu, \phi, \sigma^2) = &-\frac{n}{2}\log(2\pi) - n\log\sigma + \frac{1}{2}\log(1-\phi^2) \\ &- \frac{(1-\phi^2)(y_1-\mu)^2}{2\sigma^2} - \sum_{i=2}^{n} \frac{(y_t - \mu(1-\phi) - \phi y_{t-1})^2}{2\sigma^2}\end{aligned}$$

1.3 Monte Carlo methods

Consider a target distribution with finite expected value. The law of large numbers guarantees that, if we can draw observations from our target distribution, then the sample average will converge to the expected value of that distribution, as the sample size becomes larger and larger, provided the expectation is finite.

We can thus compute the probability of any event or the expected value of any (integrable) function by computing sample averages; the cost to pay for this generality is randomness.

Specifically, suppose we are interested in the average $\mathsf{E}\{g(X)\}$ of $X \sim F$ for some function g.

Example 1.12. Consider $X \sim \mathsf{gamma}(\alpha, \beta)$, a gamma distribution with shape α and rate β. We can compute the probability that $X < 1$ easily by Monte Carlo since $\Pr(X < 1) = \mathsf{E}\{\mathrm{I}(X < 1)\}$ and this means we only need to compute the proportion of draws less than one. We can likewise compute the mean $g(x) = x$ or variance.

Suppose we have drawn a Monte Carlo sample of size B. If the function $g(\cdot)$ is square integrable,[2] with variance σ_g^2, then a central limit theorem applies. In large samples and for independent observations, our Monte Carlo average $\widehat{\mu}_g = B^{-1}\sum_{b=1}^{B} g(X_i)$ has variance σ_g^2/B. We can approximate the unknown variance σ_g^2 by it's empirical counterpart.[3]. Note that, while the variance decreases linearly with B, the choice of g impacts the speed of convergence: we can compute

$$\sigma_g^2 = \Pr(X \leq 1)\{1 - \Pr(X \leq 1)\} = 0.0434$$

(left) and $\sigma_g^2 = \alpha/\beta^2 = 1/8$ (middle plot).

Figure 1.2 shows the empirical trace plot of the Monte Carlo average (note the $\sqrt{B}$ x-axis scale!) as a function of the Monte Carlo sample size B along with 95% Wald-based confidence intervals (gray shaded region), $\widehat{\mu}_g \pm 1.96 \times \sigma_g/\sqrt{B}$. We can see that the 'likely region' for the average shrinks with B.

What happens if our function is not integrable? The right-hand plot of Figure 1.2 shows empirical averages of $g(x) = x^{-1}$, which is not integrable if $\alpha < 1$. We can compute the empirical average, but the result won't converge to any meaningful quantity regardless of the sample size. The large jumps are testimonial of this.

We have already used Monte Carlo methods to compute posterior quantities of interest in conjugate models. Outside of models with conjugate priors, the lack of closed-form expression for the posterior precludes inference. Indeed, calculating the posterior probability of an event, or posterior moments, requires integration of the normalized posterior density and thus knowledge of the marginal likelihood. It is seldom possible to sample independent and identically distributed (iid) samples from the target, especially if the model is high dimensional: rejection sampling and the ratio of uniform method are examples of Monte Carlo methods which can be used to generate iid draws.

Proposition 1.6 (Rejection sampling)**.** *Rejection sampling (also termed accept-reject algorithm) samples from a random vector with density $p(\cdot)$ by drawing candidates from a*

[2]Meaning $\mathsf{E}\{g^2(X)\} < \infty$, so the variance of $g(X)$ exists.
[3]By contrasts, if data are identically distributed but not independent, more care is needed.

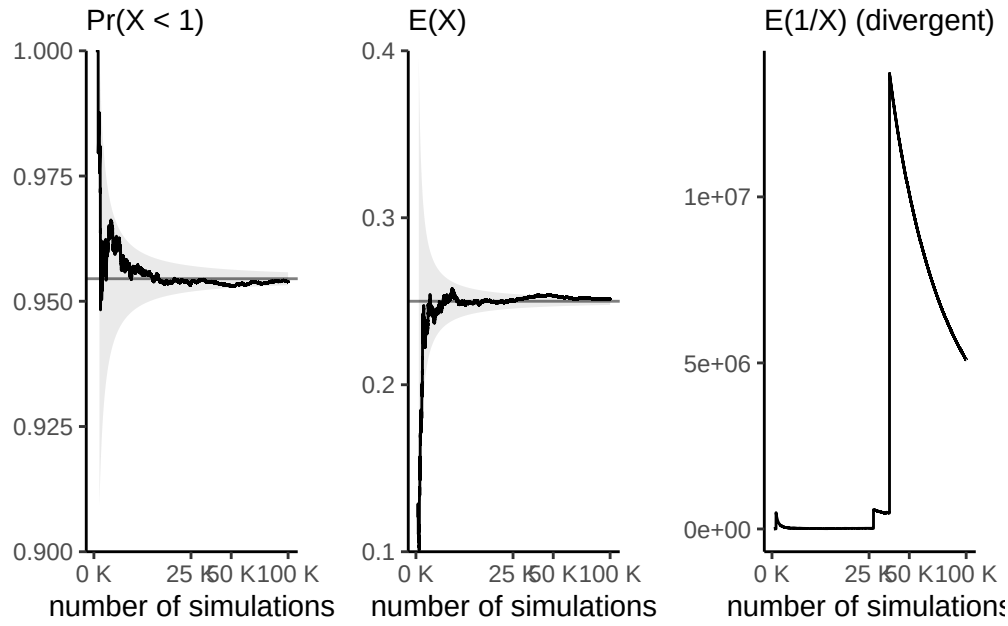

Figure 1.2: Running mean trace plots for $g(x) = \mathrm{I}(x < 1)$ (left), $g(x) = x$ (middle) and $g(x) = 1/x$ (right) for a Gamma distribution with shape 0.5 and rate 2, as a function of the Monte Carlo sample size.

proposal with density $q(\cdot)$ with nested support, $\mathrm{supp}(p) \subseteq \mathrm{supp}(q)$. The density $q(\cdot)$ must be such that $p(\boldsymbol{\theta}) \leq Cq(\boldsymbol{\theta})$ for $C \geq 1$ for all values of $\boldsymbol{\theta}$ in the support of $p(\cdot)$. A proof can be found in Devroye (1986, Theorem 3.1)

1. *Generate $\boldsymbol{\theta}^\star$ from the proposal with density q and $U \sim \mathsf{U}(0,1)$*
2. *Compute the ratio $R \leftarrow p(\boldsymbol{\theta}^\star)/q(\boldsymbol{\theta}^\star)$.*
3. *If $R \geq CU$, return $\boldsymbol{\theta}$, else go back to step 1.*

Rejection sampling requires the proposal q to have a support at least as large as that of p and resemble closely the density. It should be chosen so that the upper bound C is as sharp as possible and close to 1. The dominating density q must have heavier tails than the density of interest. The expected number of simulations needed to accept one proposal is C. Finally, for the method to be useful, we need to be able to simulate easily and cheaply from the proposal. The optimal value of C is $C = \sup_{\boldsymbol{\theta}} p(\boldsymbol{\theta})/q(\boldsymbol{\theta})$. This quantity may be obtained by numerical optimization, by finding the mode of the ratio of the log densities if the maximum is not known analytically.

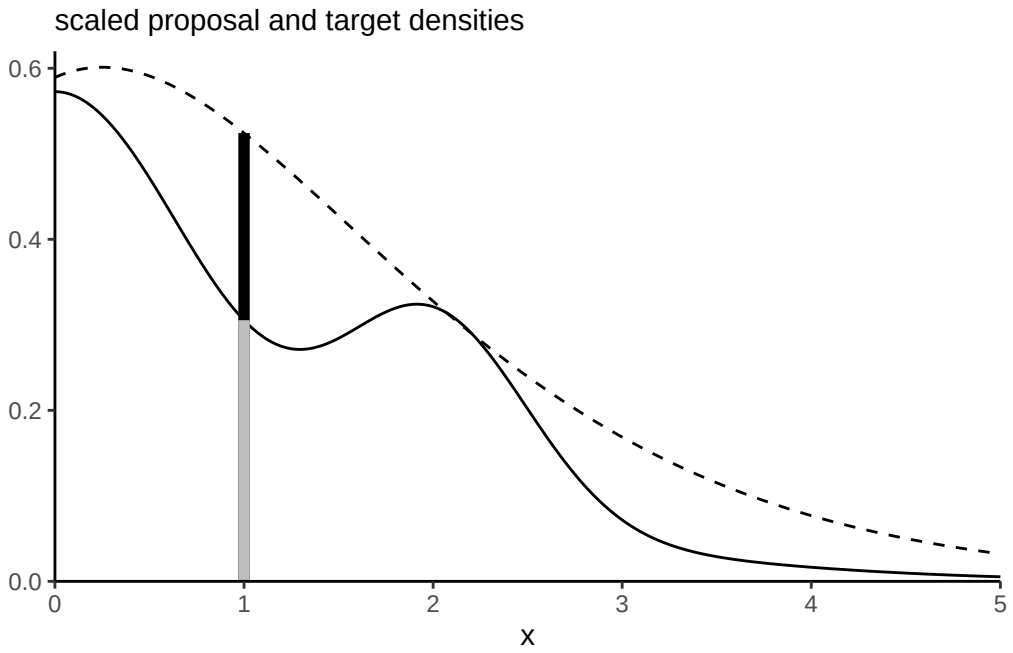

Figure 1.3: Target density (full) and scaled proposal density (dashed): the vertical segment at $x = 1$ shows the percentage of acceptance for a uniform slice under the scaled proposal, giving an acceptance ratio of 0.58.

Example 1.13 (Truncated Gaussian distribution)**.** Consider the problem of sampling from a Gaussian distribution $Y \sim \mathsf{Gauss}(\mu, \sigma^2)$ truncated in the interval $[a, b]$, which has density

$$f(x; \mu, \sigma, a, b) = \frac{1}{\sigma} \frac{\phi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\{(b-\mu)/\sigma\} - \Phi\{(a-\mu)/\sigma\}}.$$

where $\phi(\cdot), \Phi(\cdot)$ are respectively the density and distribution function of the standard Gaussian distribution.

Since the Gaussian is a location-scale family, we can reduce the problem to sampling X from a standard Gaussian truncated on $\alpha = (a-\mu)/\sigma$ and $\beta = (b-\mu)/\sigma$ and back transform the result as $Y = \mu + \sigma X$.

A crude accept-reject sampling algorithm would consider sampling from the same untruncated distribution with density $g(X) = \sigma^{-1}\phi\{(x-\mu)/\sigma\}$, and the acceptance ratio is $C^{-1} = \{\Phi(\beta) - \Phi(\alpha)\}$. We thus simply simulate points from the Gaussian and accept any that falls within the bounds.

```
# Standard Gaussian truncated on [0,1]
candidate <- rnorm(1e5)
trunc_samp <- candidate[candidate >= 0 & candidate <= 1]
# Acceptance rate
length(trunc_samp)/1e5
```

```
[1] 0.34242
```

```
# Theoretical acceptance rate
pnorm(1)-pnorm(0)
```

```
[1] 0.3413447
```

We can of course do better: if we consider a random variable with distribution function F, but truncated over the interval $[a, b]$, then the resulting distribution function is

$$\frac{F(x) - F(a)}{F(b) - F(a)}, \qquad a \leq x \leq b,$$

and we can invert this expression to get the quantile function of the truncated variable in terms of the distribution function F and the quantile function F^{-1} of the original untruncated variable.

For the Gaussian, this gives

$$X \sim \Phi^{-1}\left[\Phi(a) + \{\Phi(b) - \Phi(a)\}U\right]$$

for $U \sim \mathsf{U}(0, 1)$. Although the quantile and distribution functions of the Gaussian, `pnorm` and `qnorm` in **R**, are very accurate, this method will fail for rare event simulation because it will return $\Phi(x) = 0$ for $x \leq -39$ and $\Phi(x) = 1$ for $x \geq 8.3$, implying that $a \leq 8.3$ for this approach to work (Botev and L'Écuyer 2017).

Consider the problem of simulating events in the right tail for a standard Gaussian where $a > 0$; Marsaglia's method (Devroye 1986, 381), can be used for that purpose. Write the density of the Gaussian as $f(x) = \exp(-x^2/2)/c_1$, where $c_1 = \int_a^\infty \exp(-z^2/2)\mathrm{d}z$, and note that

$$c_1 f(x) \leq \frac{x}{a}\exp\left(-\frac{x^2}{2}\right) = a^{-1}\exp\left(-\frac{a^2}{2}\right) g(x), \qquad x \geq a;$$

where $g(x)$ is the density of a Rayleigh variable shifted by a, which has distribution function $G(x) = 1-\exp\{(a^2-x^2)/2\}$ for $x \geq a$. We can simulate such a random variate X through the inversion method. The constant $C = \exp(-a^2/2)(c_1 a)^{-1}$ approaches 1 quickly as $a \to \infty$.

The accept-reject thus proceeds with

1. Generate a shifted Rayleigh above a, $X \leftarrow \{a^2 - 2\log(U)\}^{1/2}$ for $U \sim \mathsf{U}(0,1)$
2. Accept X if $XV \leq a$, where $V \sim \mathsf{U}(0,1)$.

Should we wish to obtain samples on $[a,b]$, we could instead propose from a Rayleigh truncated above at b (Botev and L'Écuyer 2017).

```
a <- 8.3
niter <- 1000L
X <- sqrt(a^2 + 2*rexp(niter))
samp <- X[runif(niter)*X <= a]
```

For a given candidate density g which has a heavier tail than the target, we can resort to numerical methods to compute the mode of the ratio f/g and obtain the bound C; see Albert (2009), Section 5.8 for an insightful example.

2 Bayesics

The Bayesian paradigm is an inferential framework that is used widespread in data science. Numerical challenges that prevented it's widespread adoption until the 90's, when the Markov chain Monte Carlo revolution allowed models estimation.

Bayesian inference, which builds on likelihood-based inference, offers a natural framework for prediction and for uncertainty quantification. The interpretation is more natural than that of classical (i.e., frequentist) paradigm, and it is more easy to generalized models to complex settings, notably through hierarchical constructions. The main source of controversy is the role of the prior distribution, which allows one to incorporate subject-matter expertise but leads to different inferences being drawn by different practitioners; this subjectivity is not to the taste of many and has been the subject of many controversies.

The Bayesian paradigm includes multiples notions that are not covered in undergraduate introductory courses. The purpose of this chapter is to introduce these concepts and put them in perspective; the reader is assumed to be familiar with basics of likelihood-based inference. We begin with a discussion of the notion of probability, then define priors, posterior distributions, marginal likelihood and posterior predictive distributions. We focus on the interpretation of posterior distributions and explain how to summarize the posterior, leading leading to definitions of high posterior density region, credible intervals, posterior mode for cases where we either have a (correlated) sample from the posterior, or else have access to the whole distribution. Several notions, including sequentiality, prior elicitation and estimation of the marginal likelihood, are mentioned in passing. A brief discussion of Bayesian hypothesis testing (and alternatives) is presented.

2.1 Probability and frequency

In classical (frequentist) parametric statistic, we treat observations $\boldsymbol{Y}$ as realizations of a distribution whose parameters $\boldsymbol{\theta}$ are unknown. All of the information about parameters is encoded by the likelihood function.

The interpretation of probability in the classical statistic is in terms of long run frequency, which is why we term this approach frequentist statistic. Think of a fair die: when we state that values $\{1,\dots,6\}$ are equiprobable, we mean that repeatedly tossing the die

should result, in large sample, in each outcome being realized roughly $1/6$ of the time (the symmetry of the object also implies that each facet should be equally likely to lie face up). This interpretation also carries over to confidence intervals: a $(1-\alpha)$ confidence interval either contains the true parameter value or it doesn't, so the probability level $(1-\alpha)$ is only the long-run proportion of intervals created by the procedure that should contain the true fixed value, not the probability that a single interval contains the true value. This is counter-intuitive to most.

In practice, the true value of the parameter $\boldsymbol{\theta}$ vector is unknown to the practitioner, thus uncertain: Bayesians would argue that we should treat the latter as a random quantity rather than a fixed constant. Since different people may have different knowledge about these potential values, the prior knowledge is a form of **subjective probability**. For example, if you play cards, one person may have recorded the previous cards that were played, whereas other may not. They thus assign different probability of certain cards being played. In Bayesian inference, we consider $\boldsymbol{\theta}$ as random variables to reflect our lack of knowledge about potential values taken. Italian scientist Bruno de Finetti, who is famous for the claim "Probability does not exist' ', stated in the preface of Finetti (1974):

> Probabilistic reasoning — always to be understood as subjective — merely stems from our being uncertain about something. It makes no difference whether the uncertainty relates to an unforseeable future, or to an unnoticed past, or to a past doubtfully reported or forgotten: it may even relate to something more or less knowable (by means of a computation, a logical deduction, etc.) but for which we are not willing or able tho make the effort; and so on [...] The only relevant thing is uncertainty — the extent of our knowledge and ignorance. The actual fact of whether or not the events considered are in some sense *determined*, or known by other people, and so on, is of no consequence.

On page 3, de Finetti continues (Finetti 1974)

> only subjective probabilities exist — i.e., the degree of belief in the occurrence of an event attributed by a given person at a given instant and with a given set of information.

2.2 Posterior distribution

We consider a parametric model with parameters $\boldsymbol{\theta}$ defined on $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$. In Bayesian learning, we adjoin to the likelihood $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) \equiv p(\boldsymbol{y} \mid \boldsymbol{\theta})$ a **prior** function $p(\boldsymbol{\theta})$ that reflects the prior knowledge about potential values taken by the p-dimensional parameter vector, before

observing the data $\boldsymbol{y}$. The prior makes $\boldsymbol{\theta}$ random and the distribution of the parameter reflects our uncertainty about the true value of the model parameters.

In a Bayesian analysis, observations are random variables but inference is performed conditional on the observed sample values. By Bayes' theorem, our target is therefore the posterior density $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, defined as

$$\underbrace{p(\boldsymbol{\theta} \mid \boldsymbol{y})}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{y} \mid \boldsymbol{\theta})}^{\text{likelihood}} \times \overbrace{p(\boldsymbol{\theta})}^{\text{prior}}}{\underbrace{\int p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}_{\text{marginal likelihood } p(\boldsymbol{y})}}. \tag{2.1}$$

The posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ is proportional, as a function of θ, to the product of the likelihood and the prior function.

For the posterior to be **proper**, we need the product of the prior and the likelihood on the right hand side to be integrable as a function of $\boldsymbol{\theta}$ over the parameter domain $\boldsymbol{\Theta}$. The integral in the denominator, termed marginal likelihood or prior predictive distribution and denoted $p(\boldsymbol{y}) = \mathsf{E}_{\boldsymbol{\theta}}\{p(\boldsymbol{y} \mid \boldsymbol{\theta})\}$. It represents the distribution of the data before data collection, the respective weights being governed by the prior probability of different parameters values. The denominator of Equation 2.1 is a normalizing constant, making the posterior density integrate to unity. The marginal likelihood plays a central role in Bayesian testing.

If $\boldsymbol{\theta}$ is low dimensional, numerical integration such as quadrature methods can be used to compute the marginal likelihood.

To fix ideas, we consider next a simple one-parameter model where the marginal likelihood can be computed explicitly.

Example 2.1 (Binomial model with beta prior)**.** Consider a binomial likelihood with probability of success $\theta \in [0, 1]$ and n trials, $Y \sim \mathsf{binom}(n, \theta)$. If we take a beta prior, $\theta \sim \mathsf{beta}(\alpha, \beta)$ and observe y successes, the posterior is

$$\begin{aligned} p(\theta \mid y = y) &\propto \binom{n}{y} \theta^y (1-\theta)^{n-y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\overset{\theta}{\propto} \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \end{aligned}$$

and is

$$\int_0^1 \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \mathrm{d}\theta = \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)},$$

a Beta function. Since we need only to keep track of the terms that are function of the parameter θ, we could recognize directly that the posterior distribution is $\mathsf{beta}(y+\alpha, n-y+\beta)$ and deduce the normalizing constant from there.

If $Y \sim \mathsf{binom}(n,\theta)$, the expected number of success is $n\theta$ and the expected number of failures $n(1-\theta)$ and so the likelihood contribution, relative to the prior, will dominate as the sample size n grows.

Another way to see this is to track moments (expectation, variance, etc.) The Beta distribution, whose density is $f(x;\alpha,\beta) \propto x^{\alpha-1}(1-x)^{\beta-1}$, has expectation $\alpha/(\alpha+\beta)$ and variance $\alpha\beta/\{(\alpha+\beta)^2(\alpha+\beta+1)\}$. The posterior mean is

$$\mathsf{E}(\theta \mid y) = w\frac{y}{n} + (1-w)\frac{\alpha}{\alpha+\beta}, \qquad w = \frac{n}{n+\alpha+\beta},$$

a weighted average of the maximum likelihood estimator and the prior mean. We can think of the parameter α (respectively β) as representing the fixed prior number of success (resp. failures). The variance term is $\mathrm{O}(n^{-1})$ and, as the sample size increases, the likelihood weight w dominates.

Figure 2.1 shows three different posterior distributions with different beta priors: the first prior, which favors values closer to 1/2, leads to a more peaked posterior density, contrary to the second which is symmetric, but concentrated toward more extreme values near endpoints of the support. The rightmost panel is truncated: as such, the posterior is zero for any value of θ beyond 1/2 and so the posterior mode may be close to the endpoint of the prior. The influence of such a prior will not necessarily vanish as sample size and should be avoided, unless there are compelling reasons for restricting the domain.

Remark (Proportionality). Any term appearing in the likelihood times prior function that does not depend on parameters can be omitted since they will be absorbed by the normalizing constant. This makes it useful to compute normalizing constants or likelihood ratios.

Remark. An alternative parametrization for the beta distribution sets $\alpha = \mu\kappa$, $\beta = (1-\mu)\kappa$ for $\mu \in (0,1)$ and $\kappa > 0$, so that the model is parametrized directly in terms of mean μ, with κ capturing the dispersion.

Remark. A density integrates to 1 over the range of possible outcomes, but there is no guarantee that the likelihood function, as a function of $\boldsymbol{\theta}$, integrates to one over the parameter domain $\boldsymbol{\Theta}$.

For example, the binomial likelihood with n trials and y successes satisfies

$$\int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}\mathrm{d}\theta = \frac{1}{n+1}.$$

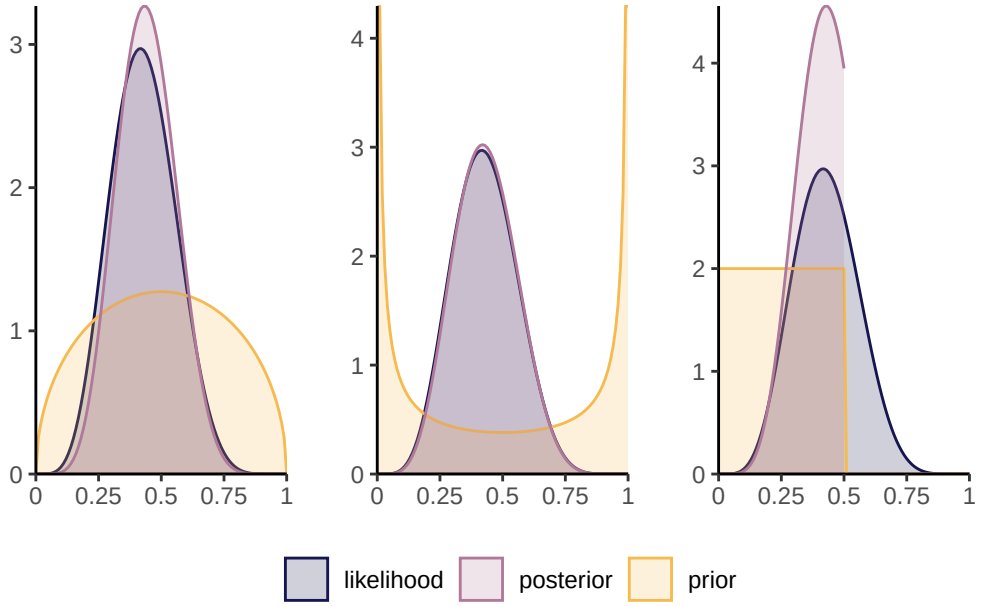

Figure 2.1: Scaled binomial likelihood for six successes out of 14 trials, with $\texttt{beta}(3/2, 3/2)$ prior (left), $\texttt{beta}(1/4, 1/4)$ (middle) and truncated uniform on $[0, 1/2]$ (right), with the corresponding posterior distributions.

Moreover, the binomial distribution is discrete with support $0, \ldots, n$, whereas the likelihood is continuous as a function of the probability of success, as evidenced by Figure 2.2

Proposition 2.1 (Sequentiality and Bayesian updating)**.** *The likelihood is invariant to the order of the observations if they are independent. Thus, if we consider two blocks of observations* $\boldsymbol{y}_1$ *and* $\boldsymbol{y}_2$

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \boldsymbol{y}_2) = p(\boldsymbol{\theta} \mid \boldsymbol{y}_1)p(\boldsymbol{\theta} \mid \boldsymbol{y}_2),$$

so it makes no difference if we treat data all at once or in blocks. More generally, for data exhibiting spatial or serial dependence, it makes sense to consider rather the conditional (sequential) decomposition

$$f(\boldsymbol{y}; \boldsymbol{\theta}) = f(\boldsymbol{y}_1; \boldsymbol{\theta})f(\boldsymbol{y}_2; \boldsymbol{\theta}, \boldsymbol{y}_1) \cdots f(\boldsymbol{y}_n; \boldsymbol{\theta}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n-1})$$

where $f(\boldsymbol{y}_k; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{k-1})$ *denotes the conditional density function given observations* $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{k-1}$.

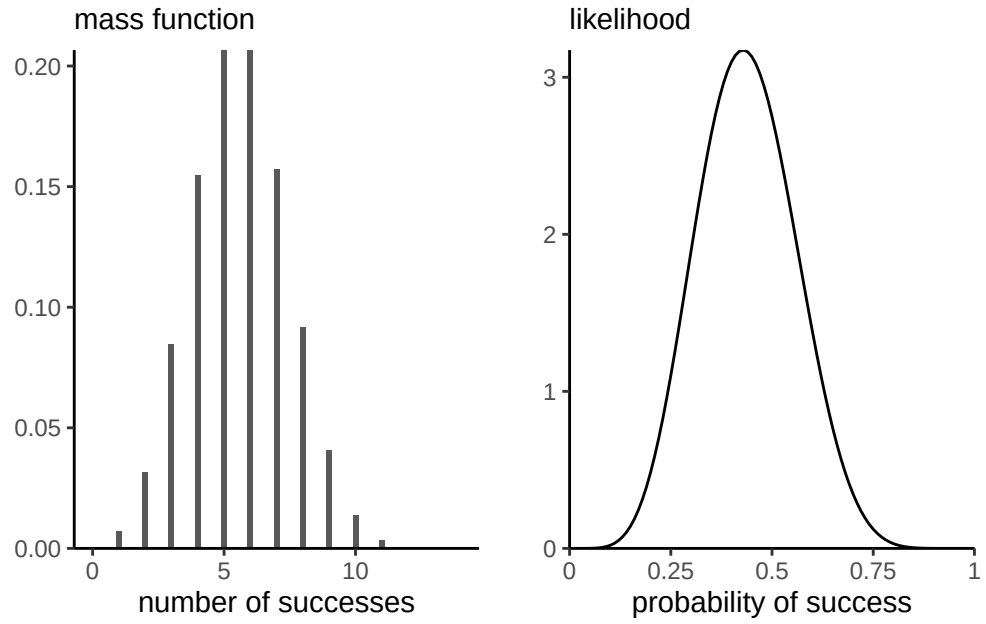

Figure 2.2: Binomial mass function (left) and scaled likelihood function (right).

By Bayes' rule, we can consider updating *the posterior by adding terms to the likelihood, noting that*

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \boldsymbol{y}_2) \propto p(\boldsymbol{y}_2 \mid \boldsymbol{y}_1, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}_1)$$

which amounts to treating the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y}_1)$ *as a prior. If data are exchangeable, the order in which observations are collected and the order of the belief updating is irrelevant to the full posterior. Figure 2.3 shows how the posterior becomes gradually closer to the scaled likelihood as we increase the sample size, and the posterior mode moves towards the true value of the parameter (here 0.3).*

Example 2.2. While we can calculate analytically the value of the normalizing constant for the beta-binomial model, we could also for arbitrary priors use numerical integration or Monte Carlo methods in the event the parameter vector $\boldsymbol{\theta}$ is low-dimensional.

While estimation of the normalizing constant is possible in simple models, the following highlights some challenges that are worth keeping in mind. In a model for discrete data (that is, assigning probability mass to a countable set of outcomes), the terms in the likelihood are probabilities and thus the likelihood becomes smaller as we gather more observations

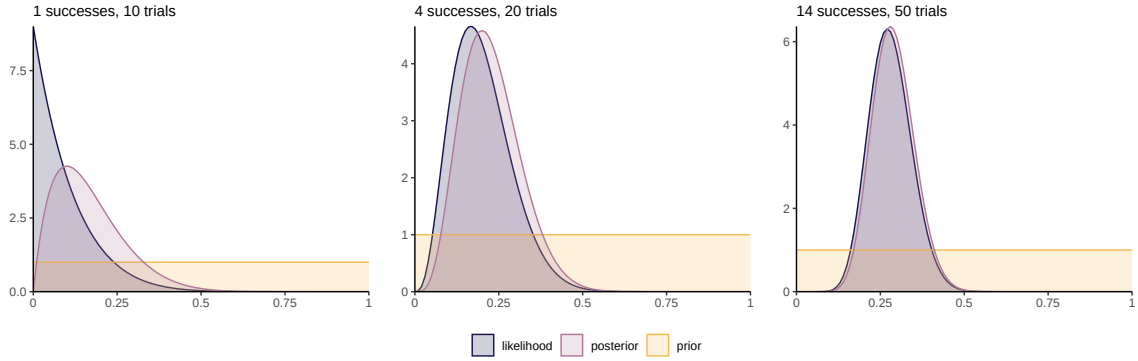

Figure 2.3: Beta posterior and binomial likelihood with a uniform prior for increasing number of observations (from left to right) out of a total of 100 trials.

(since we multiply terms between zero or one). The marginal likelihood term becomes smaller and smaller, so it's reciprocal is big and this can lead to arithmetic underflow.

```
y <- 6L # number of successes
n <- 14L # number of trials
alpha <- beta <- 1.5 # prior parameters
unnormalized_posterior <- function(theta){
  theta^(y+alpha-1) * (1-theta)^(n-y + beta - 1)
}
integrate(f = unnormalized_posterior,
          lower = 0,
          upper = 1)
```

```
1.066906e-05 with absolute error < 1e-12
```

```
# Compare with known constant
beta(y + alpha, n - y + beta)
```

```
[1] 1.066906e-05
```

```
# Monte Carlo integration
mean(unnormalized_posterior(runif(1e5)))
```

```
[1] 1.064067e-05
```

When $\boldsymbol{\theta}$ is high-dimensional, the marginal likelihood is intractable. This is one of the main challenges of Bayesian statistics and the popularity and applicability has grown drastically with the development and popularity of numerical algorithms, following the publication of Geman and Geman (1984) and Gelfand and Smith (1990). Markov chain Monte Carlo methods circumvent the calculation of the denominator by drawing approximate samples from the posterior.

2.3 Posterior predictive distribution

Prediction in the Bayesian paradigm is obtained by considering the *posterior predictive distribution*,

$$p(y_{\text{new}} \mid \boldsymbol{y}) = \int_{\Theta} p(y_{\text{new}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}$$

Given draws from the posterior distribution, say $\boldsymbol{\theta}_b$ $(b = 1, \ldots, B)$, we sample from each a new realization from the distribution appearing in the likelihood $p(y_{\text{new}} \mid \boldsymbol{\theta}_b)$. This is different from the frequentist setting, which fixes the value of the parameter to some estimate $\widehat{\boldsymbol{\theta}}$; by contrast, the posterior predictive, here a beta-binomial distribution $\mathsf{BetaBin}(n, \alpha + y, n - y + \beta)$, carries over the uncertainty so will typically be wider and overdispersed relative to the corresponding binomial model. This can be easily seen from the left-panel of Figure 2.4, which contrasts the binomial mass function evaluated at the maximum likelihood estimator $\widehat{\theta} = 6/14$ with the posterior predictive.

```
npost <- 1e4L
# Sample draws from the posterior distribution
post_samp <- rbeta(n = npost, y + alpha, n - y + beta)
# For each draw, sample new observation
post_pred <- rbinom(n = npost, size = n, prob = post_samp)
```

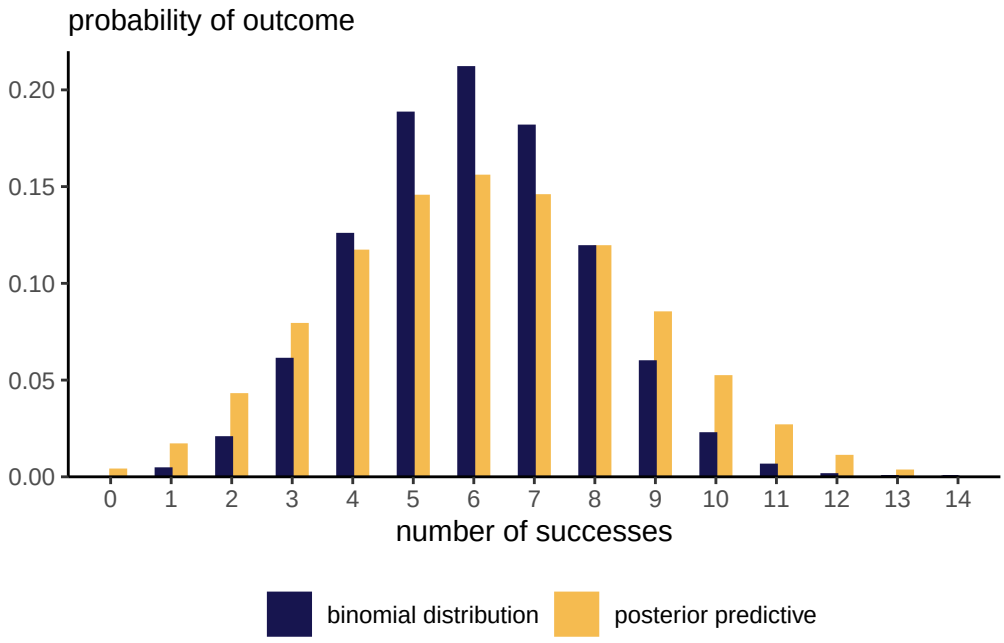

Figure 2.4: Beta-binomial posterior predictive distribution with corresponding binomial mass function evaluated at the maximum likelihood estimator.

Example 2.3 (Posterior predictive distribution of univariate Gaussian with known mean)**.** Consider an n sample of independent and identically distributed Gaussian, $Y_i \sim \mathsf{Gauss}(0, \tau^{-1})$ $(i = 1, \ldots, n)$, where we assign a gamma prior on the precision $\tau \sim \mathsf{gamma}(\alpha, \beta)$. The posterior is

$$p(\tau \mid \boldsymbol{y}) \overset{\tau}{\propto} \prod_{i=1}^{n} \tau^{n/2} \exp\left(-\tau \frac{\sum_{i=1}^{n} y_i^2}{2}\right) \times \tau^{\alpha-1} \exp(-\beta\tau)$$

and rearranging the terms to collect powers of τ, etc. we find that the posterior for τ must also be gamma, with shape parameter $\alpha^* = \alpha + n/2$ and rate $\beta^* = \beta + \sum_{i=1}^{n} y_i^2/2$.

The posterior predictive is

$$\begin{aligned}
p(y_{\text{new}} \mid \boldsymbol{y}) &= \int_0^\infty \frac{\tau^{1/2}}{(2\pi)^{1/2}} \exp(-\tau y_{\text{new}}^2/2) \frac{\beta^{*\alpha^*}}{\Gamma(\alpha^*)} \tau^{\alpha^*-1} \exp(-\beta^*\tau) \mathrm{d}\tau \\
&= (2\pi)^{-1/2} \frac{\beta^{*\alpha^*}}{\Gamma(\alpha^*)} \int_0^\infty \tau^{\alpha^*-1/2} \exp\left\{-\tau(y_{\text{new}}^2/2 + \beta^*)\right\} \mathrm{d}\tau \\
&= (2\pi)^{-1/2} \frac{\beta^{*\alpha^*}}{\Gamma(\alpha^*)} \frac{\Gamma(\alpha^* + 1/2)}{(y_{\text{new}}^2/2 + \beta^*)^{\alpha^*+1/2}} \\
&= \frac{\Gamma\left(\frac{2\alpha^*+1}{2}\right)}{\sqrt{2\pi}\Gamma\left(\frac{2\alpha^*}{2}\right)\beta^{*1/2}} \left(1 + \frac{y_{\text{new}}^2}{2\beta^*}\right)^{-\alpha^*-1/2} \\
&= \frac{\Gamma\left(\frac{2\alpha^*+1}{2}\right)}{\sqrt{\pi}\sqrt{2\alpha^*}\Gamma\left(\frac{2\alpha^*}{2}\right)(\beta^*/\alpha^*)^{1/2}} \left(1 + \frac{1}{2\alpha^*} \frac{y_{\text{new}}^2}{(\beta^*/\alpha^*)}\right)^{-\alpha^*-1/2}
\end{aligned}$$

which entails that Y_{new} is a scaled Student-t distribution with scale $(\beta^*/\alpha^*)^{1/2}$ and $2\alpha + n$ degrees of freedom. This example also exemplifies the additional variability relative to the distribution generating the data: indeed, the Student-t distribution is more heavy-tailed than the Gaussian, but since the degrees of freedom increase linearly with n, the distribution converges to a Gaussian as $n \to \infty$, reflecting the added information as we collect more and more data points and the variance gets better estimated through $\sum_{i=1}^n y_i^2/n$.

2.4 Summarizing posterior distributions

The output of the Bayesian learning problem will be either of:

1. a fully characterized distribution
2. a numerical approximation to the posterior distribution (pointwise)
3. an exact or approximate sample drawn from the posterior distribution

In the first case, we will be able to directly evaluate quantities of interest if there are closed-form expressions for the latter, or else we could draw samples from the distribution and evaluate them via Monte-Carlo. In case of numerical approximations, we will need to resort to numerical integration or otherwise to get our answers.

Often, we will also be interested in the marginal posterior distribution of each component θ_j in turn ($j = 1, \ldots, J$). To get these, we carry out additional integration steps,

$$p(\theta_j \mid \boldsymbol{y}) = \int p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}_{-j}.$$

With a posterior sample, this is trivial: it suffices to keep the column corresponding to θ_j and discard the others.

Most of the field of Bayesian statistics revolves around the creation of algorithms that either circumvent the calculation of the normalizing constant (notably using Monte Carlo and Markov chain Monte Carlo methods) or else provide accurate numerical approximation of the posterior pointwise, including for marginalizing out all but one parameters (integrated nested Laplace approximations, variational inference, etc.) The target of inference is the whole posterior distribution, a potentially high-dimensional object which may be difficult to summarize or visualize. We can thus report only characteristics of the the latter.

The choice of point summary to keep has it's root in decision theory.

Definition 2.1 (Loss function)**.** A loss function $c(\boldsymbol{\theta}, \boldsymbol{\upsilon})$ is a mapping from $\boldsymbol{\Theta} \to \mathbb{R}^k$ that assigns a weight to each value of $\boldsymbol{\theta}$, corresponding to the regret or loss arising from choosing this value. The corresponding point estimator $\widehat{\boldsymbol{\upsilon}}$ is the minimizer of the expected loss,

$$\begin{aligned}\widehat{\boldsymbol{\upsilon}} &= \operatorname*{argmin}_{\boldsymbol{\upsilon}} \mathsf{E}_{\boldsymbol{\Theta}|\boldsymbol{Y}}\{c(\boldsymbol{\theta}, \boldsymbol{\upsilon})\} \\ &= \operatorname*{argmin}_{\boldsymbol{\upsilon}} \int_{\boldsymbol{\Theta}} c(\boldsymbol{\theta}, \boldsymbol{\upsilon}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}\end{aligned}$$

For example, in a univariate setting, the quadratic loss $c(\theta, \upsilon) = (\theta - \upsilon)^2$ returns the posterior mean, the absolute loss $c(\theta, \upsilon) = |\theta - \upsilon|$ returns the posterior median and the 0-1 loss $c(\theta, \upsilon) = \mathrm{I}(\upsilon \neq \theta)$ returns the posterior mode. All of these point estimators are central tendency measures, but some may be more adequate depending on the setting as they can correspond to potentially different values, as shown in the left-panel of Figure 2.5. The choice is application specific: for multimodal distributions, the mode is likely a better choice.

If we know how to evaluate the distribution numerically, we can optimize to find the mode or else return the value for the pointwise evaluation on a grid at which the density achieves it's maximum. The mean and median would have to be evaluated by numerical integration if there is no closed-form expression for the latter.

If we have rather a sample from the posterior with associated posterior density values, then we can obtain the mode as the parameter combination with the highest posterior, the median from the value at rank $\lfloor n/2 \rfloor$ and the mean through the sample mean of posterior draws.

The loss function is often a functional (meaning a one-dimensional summary) from the posterior. The following example shows how it reduces a three-dimensional problem into a single risk measure.

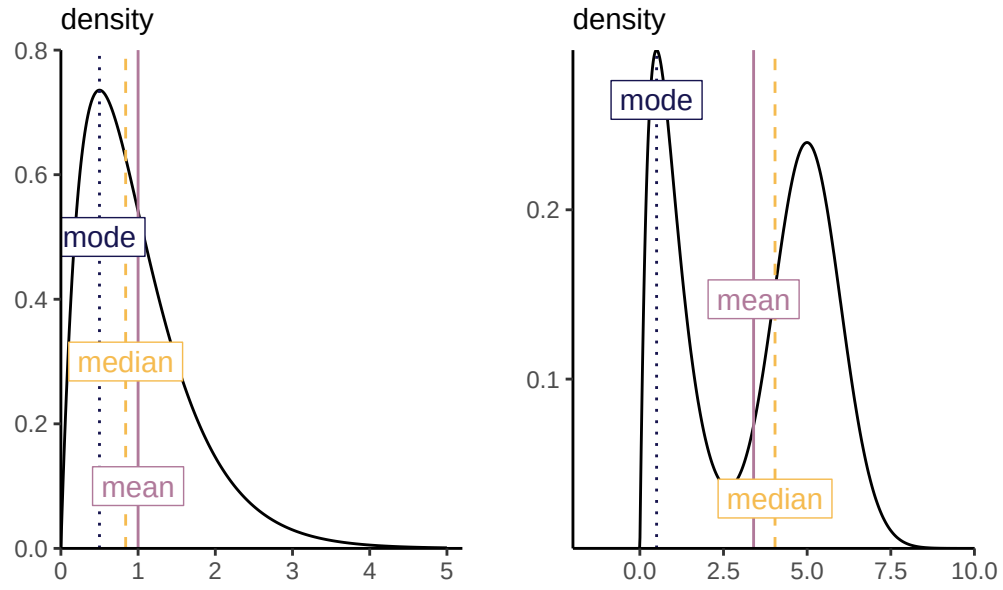

Figure 2.5: Point estimators from a right-skewed distribution (left) and from a multimodal distribution (right).

Example 2.4 (Danish insurance losses)**.** In extreme value, we are often interested in assessing the risk of events that are rare enough that they lie beyond the range of observed data. To provide a scientific extrapolation, it often is justified to fit a generalized Pareto distribution to exceedances of $Z = Y - u$, for some user-specified threshold u which is often taken as a large quantile of the distribution of Y. The generalized Pareto distribution function is

$$F(z;\tau,\xi) = 1 - \begin{cases} (1+\xi/\tau z)_+^{-1/\xi}, & \xi \neq 0 \\ \exp(-z/\tau), & \xi = 0. \end{cases}$$

The shape ξ governs how heavy-tailed the distribution is, while τ is a scale parameter.

Insurance companies provide coverage in exchange for premiums, but need to safeguard themselves against very high claims by buying reinsurance products. These risks are often communicated through the value-at-risk (VaR), a high quantile exceeded with probability p. We model Danish fire insurance claim amounts for inflation-adjusted data collected from January 1980 until December 1990 that are in excess of a million Danish kroner, found in the `evir` package and analyzed in Example 7.23 of McNeil, Frey, and Embrechts (2005). These claims are denoted Y and there are 2167 observations.

We fit a generalized Pareto distribution to exceedances above 10 millions krones, keeping 109 observations or roughly the largest 5% of the original sample. Preliminary analysis shows that we can treat data as roughly independent and identically distributed and goodness-of-fit diagnostics (not shown) for the generalized Pareto suggest that the fit is adequate for all but the three largest observations, which are (somewhat severely) underestimated by the model.

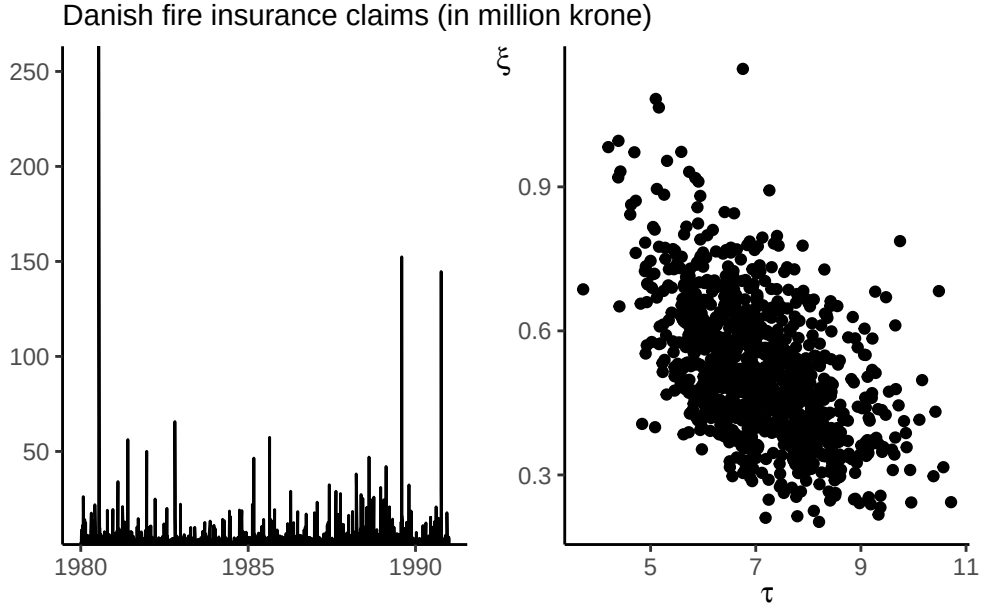

Figure 2.6: Time series of Danish fire claims exceeding a million krone (left) and posterior samples from the scale τ and shape ξ of the generalized Pareto model fitted to exceedances above 10 million krone (right).

The generalized Pareto model only describes the n_u exceedances above $u = 10$, so we need to incorporate in the likelihood a binomial contribution for the probability ζ_u of exceeding the threshold u. Provided that the priors for (τ, ξ) are independent of those for ζ_u, the posterior also factorizes as a product, so ζ_u and (τ, ξ) are a posteriori independent.

Suppose for now that we set a $\texttt{beta}(0.5, 0.5)$ prior for ζ_u and a non-informative prior for the generalized Pareto parameters.

Proposition 2.2 (Ratio of uniform method)**.** *The ratio-of-uniform method (Kinderman and Monahan 1977; Wakefield, Gelfand, and Smith 1991) is a variant of accept-reject used to draw samples from a unnormalized density $f(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d$. For some $r \geq 0$, consider*

the set

$$\mathcal{C}_r = \left\{(u_0, \ldots, u_d) : 0 < u_0 \leq [f(u_1/u_0^r, \ldots, u_d/u_0^r)]^{\frac{1}{rd+1}}\right\}.$$

If we can generate $u_0, \ldots, u_d$ uniformly over $\mathcal{C}_R$, then the draws $(u_1/u_0^r, \ldots, u_d/u_0^r)$ are from the normalized density f. Rejection sampling is used to obtain uniform draws over $\mathcal{C}_r$ under some conditions on the density and marginal moments. See the `rust` *package vignette for technical details and examples. Like with other accept-reject algorithms, the acceptance rate of the proposal goes down with the dimension of the problem.*

Example 2.5. We use the ratio-of-uniform algorithm for the data from Example 2.4 to generate draws from the posterior. We illustrate below the `rust` package with a user-specified prior and posterior. We fit a generalized Pareto distribution $Y \sim \mathsf{GP}(\sigma, \xi)$ to exceedances above 10 millions krones to the `danish` fire insurance data, using a truncated maximal data information prior $p(\sigma, \xi) \propto \sigma^{-1}\exp(-\xi+1)\mathrm{I}(\xi > -1)$.

```
data(danish, package = "evir")
# Extract threshold exceedances
exc <- danish[danish > 10] - 10
# Create a function for the log prior
logmdiprior <- function(par, ...){
  if(isTRUE(any(par[1] <= 0, par[2] < -1))){
    return(-Inf)
  }
  -log(par[1]) - par[2]
}
# Same for log likelihood, assuming independent data
loglik_gp <- function(par, data = exc, ...){
  if(isTRUE(any(par[1] <= 0, par[2] < -1))){
    return(-Inf)
  }
  sum(mev::dgp(x = data, scale = par[1], shape = par[2], log = TRUE))
}
logpost <- function(par, ...){
  logmdiprior(par) + loglik_gp(par)
}
# Sampler using ratio-of-uniform method
ru_output <- rust::ru(
  logf = logpost,  # log posterior function
  n = 10000, # number of posterior draws
```

```
  d = 2, # dimension of the parameter vector
  init = mev::fit.gpd(danish, thresh = 10)$par,
  lower = c(0, -1))
## Acceptance rate
# ru_output$pa
## Posterior samples
postsamp <- ru_output$sim_vals
```

Even without modification, the acceptance rate is 52%, which is quite efficient in the context. The generalized Pareto approximation suggests a very heavy tail: values of $\xi \geq 1$ correspond to distributions with infinite first moment, and those with $\xi \geq 1/2$ to infinite variance.

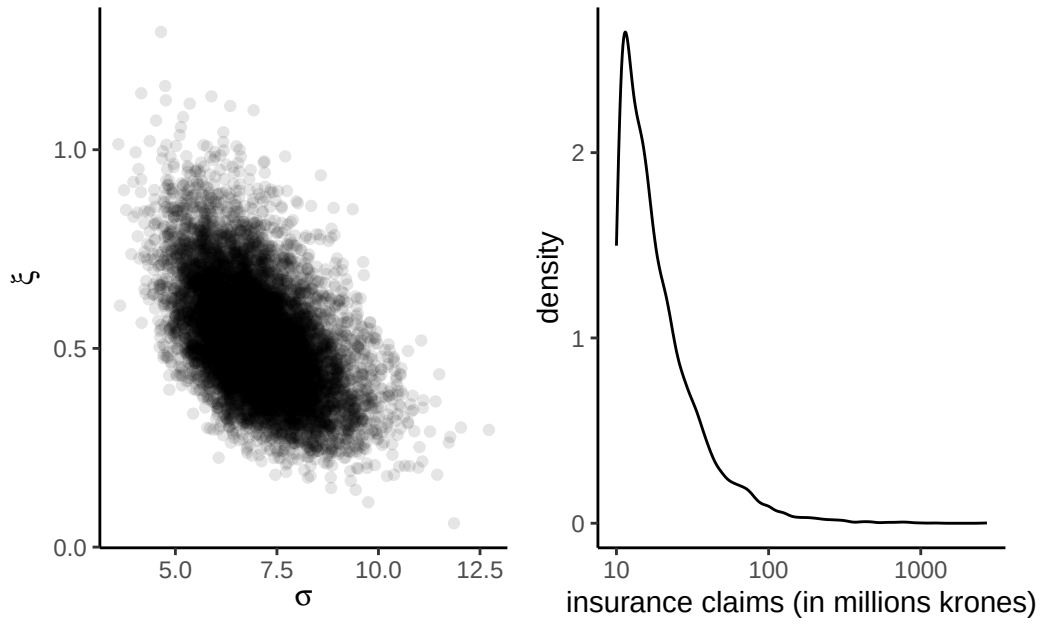

Figure 2.7: Scatterplot of posterior samples from the generalized Pareto model applied to Danish fire insurance losses above 10 millions krones, with maximal data information prior (left) and posterior predictive density on log scale (right).

The `post_samp` matrix contains exact samples from the posterior distribution of (τ, ξ, ζ_u), obtained using a Monte Carlo algorithm. Our aim is to evaluate the posterior distribution for the value-at-risk, the α quantile of Y for high values of α and see what point estimator

one would obtain depending on our choice of loss function. For any $\alpha > 1 - \zeta_u$, the q_α is

$$\begin{aligned} 1-\alpha &= \Pr(Y > q_\alpha \mid Y > u)\Pr(Y > u) \\ &= \left(1 + \xi\frac{q_\alpha - u}{\tau}\right)_+^{-1/\xi}\zeta_u \end{aligned}$$

and solving for q_α gives

$$q_\alpha = u + \frac{\tau}{\xi}\left\{\left(\frac{\zeta_u}{1-\alpha}\right)^\xi - 1\right\}.$$

To obtain the posterior distribution of the α quantile, q_α, it suffices to plug in each posterior sample and evaluate the function: the uncertainty is carried over from the simulated values of the parameters to those of the quantile q_α. The left panel of Figure 2.8 shows the posterior density estimate of the $\mathsf{VaR}(0.99)$ along with the maximum a posteriori (mode) of the latter.

Suppose that we prefer to under-estimate the value-at-risk rather than overestimate: this could be captured by the custom loss function

$$c(q, q_0) = \begin{cases} 0.5(0.99q - q_0), & q > q_0 \\ 0.75(q_0 - 1.01q), & q < q_0. \end{cases}$$

For a given value of the value-at-risk q_0 evaluated on a grid, we thus compute

$$r(q_0) = \int_{\boldsymbol{\Theta}} c(q(\boldsymbol{\theta}), q_0)p(\boldsymbol{\theta} \mid \boldsymbol{y})\mathrm{d}\boldsymbol{\theta}$$

and we seek to minimize the risk, $\widehat{q} = \mathrm{argmin}_{q_0 \in \mathbb{R}_+} r(q_0)$. The value returned that minimizes the loss, shown in Figure 2.8, is to the left of the posterior mean for q_α.

```
# Compute value at risk from generalized Pareto distribution quantile fn
VaR_post <- with(post_samp,   # data frame of posterior draws
            revdbayes::qgp(   # with columns 'probexc', 'scale', 'shape'
  p = 0.01/probexc,
  loc = 10,
  scale = scale,
  shape = shape,
  lower.tail = FALSE))
# Loss function
loss <- function(qhat, q){
    mean(ifelse(q > qhat,
```

```
           0.5*(0.99*q-qhat),
           0.75*(qhat-1.01*q)))
}
# Create a grid of values over which to estimate the loss for VaR
nvals <- 101L
VaR_grid <- seq(
  from = quantile(VaR_post, 0.01),
  to = quantile(VaR_post, 0.99),
  length.out = nvals)
# Create a container to store results
risk <- numeric(length = nvals)
for(i in seq_len(nvals)){
  # Compute integral (Monte Carlo average over draws)
 risk[i] <- loss(q = VaR_post, qhat = VaR_grid[i])
}
```

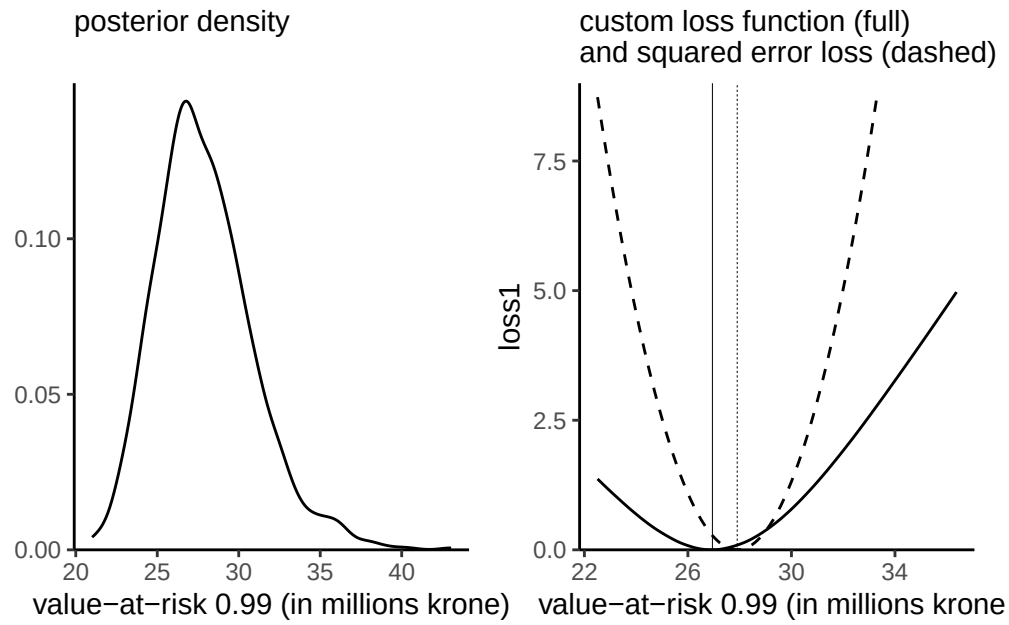

Figure 2.8: Posterior density (left) and losses functions for the 0.99 value-at-risk for the Danish fire insurance data. The vertical lines denote point estimates of the quantiles that minimize the loss functions.

To communicate uncertainty, we may resort to credible regions and intervals.

Definition 2.2. A $(1-\alpha)$ **credible region** (or credible interval in the univariate setting) is a set $\mathcal{S}_\alpha$ such that, with probability level α,

$$\Pr(\boldsymbol{\theta} \in \mathcal{S}_\alpha \mid \boldsymbol{Y} = \boldsymbol{y}) = 1-\alpha$$

These intervals are not unique, as are confidence sets. In the univariate setting, the central or equitailed interval are the most popular, and easily obtained by considering the $\alpha/2, 1-\alpha/2$ quantiles. These are easily obtained from samples by simply taking empirical quantiles. An alternative, highest posterior density credible sets, which may be a set of disjoint intervals obtained by considering the parts of the posterior with the highest density, may be more informative. The top panel Figure 2.9 shows the distinction for a bimodal mixture distribution, and a even more striking difference for 50% credible intervals for a symmetric beta distribution whose mass lie near the endpoints of the distribution, leading to no overlap between the two intervals.

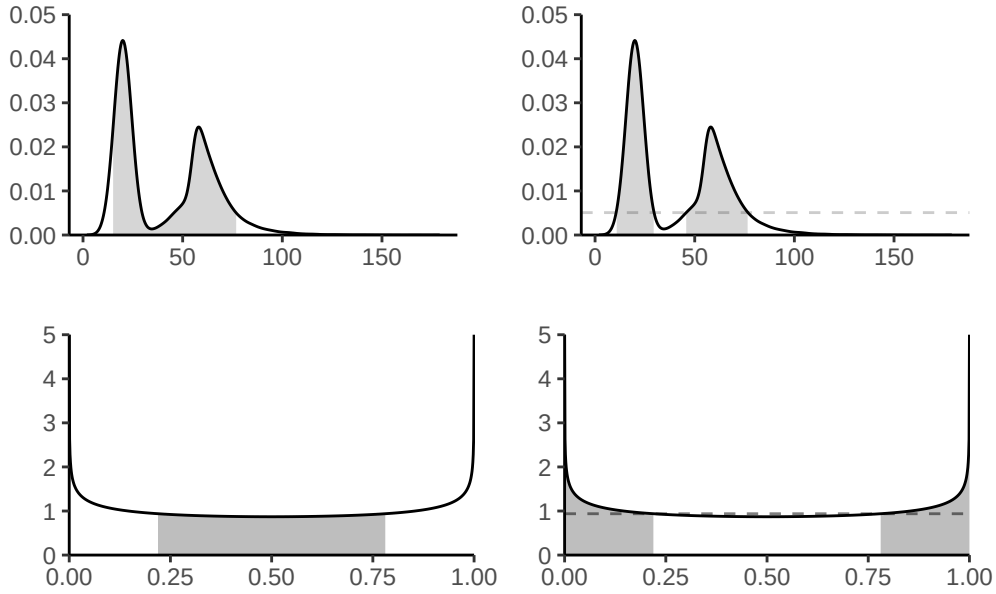

Figure 2.9: Density plots with 89% (top) and 50% (bottom) equitailed or central credible (left) and highest posterior density (right) regions for two data sets, highlighted in grey. The horizontal lign gives the posterior density value determining the cutoff for the HDP region.

3 Priors

The posterior distribution combines two ingredients: the likelihood and the prior. If the former is a standard ingredient of any likelihood-based inference, prior specification requires some care. The purpose of this chapter is to consider different standard way of constructing prior functions, and to specify the parameters of the latter: we term these hyperparameters.

The posterior is a compromise prior and likelihood: the more informative the prior, the more the posterior resembles it, but in large samples, the effect of the prior is often negligible if there is enough information in the likelihood about all parameters. We can assess the robustness of the prior specification through a sensitivity analysis by comparing the outcomes of the posterior for different priors or different values of the hyperparameters.

We can use moment matching to get sensible values, or tune via trial-and-error using the prior predictive draws to assess the implausibility of the prior outcomes. One challenge is that even if we have some prior information (e.g., we can obtain sensible prior values for the mean, quantiles or variance of the parameter of interest), these summary statisticss will not typically be enough to fully characterize the prior: many different functions or distributions could encode the same information. This means that different analysts get different inferences. Generally, we will choose the prior for convenience. Priors are controversial because they could be tuned aposteriori to give any answer an analyst might want.

3.1 Prior simulation

Expert elicitation is difficult and it is hard to grasp what the impacts of the hyperparameters are. One way to see if the priors are reasonable is to sample values from them and generate new observations, resulting in prior predictive draws.

The prior predictive is $\int_{\boldsymbol{\Theta}} p(y \mid \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$: we can simulate outcomes from it by first drawing parameter values from the prior, then sampling new observations from the distribution in the likelihood and keeping only the latter. If there are sensible bounds for the range of the response, we could discard values that do not abide with these.

Working with standardized inputs $x_i \mapsto (x_i - \overline{x})/\mathrm{sd}(\boldsymbol{x})$ is useful. For example, in a simple linear regression (with a sole numerical explanatory), the slope is the correlation between standardized explanatory X and standardized response Y and the intercept should be mean zero.

Example 3.1. Consider the daily number of Bixi bike sharing users for 2017–2019 at the Edouard Montpetit station next to HEC: we can consider a simple linear regression with log counts as a function of temperature,[1]

$$\log(\texttt{nusers}) \sim \mathsf{Gauss}_+\{\beta_0 + \beta_1(\texttt{temp} - 20), \sigma^2\}.$$

The β_1 slope measures units in degree Celsius per log number of person.

The hyperparameters depend of course on the units of the analysis, unless one standardizes response variable and explanatories: it is easier to standardize the temperature so that we consider deviations from, say 20°C, which is not far from the observed mean in the sample. After some tuning, the independent priors $\beta_0 \sim \mathsf{Gauss}(\overline{y}, 0.5^2)$, $\beta_1 \sim \mathsf{Gauss}(0, 0.05^2)$ and $\sigma \sim \mathsf{Exp}(3)$ seem to yield plausible outcomes and relationships.[2]

We can draw regression lines from the prior, as in the left panel of Figure 3.1: while some of the negative relationships appear unlikely after seeing the data, the curves all seem to pass somewhere in the cloud of point. By contrast, a silly prior is one that would result in all observations being above or below the regression line, or yield values that are much too large near the endpoints of the explanatory variable. Indeed, given the number of bikes for rental is limited (a docking station has only 20 bikes), it is also sensible to ensure that simulations do not return overly large numbers. The maximum number of daily users in the sample is 68, so priors that return simulations with more than 200 (rougly 5.3 on the log scale) are not that plausible. The prior predictive draws can help establish this and the right panel of Figure 3.1 shows that, expect for the lack of correlation between temperature and number of users, the simulated values from the prior predictive are plausible even if overdispersed.

3.2 Conjugate priors

In very simple models, there may exists prior densities that result in a posterior distribution of the same family. We can thus directly extract characteristics of the posterior. Conjugate

[1] If counts are Poisson, then the log transform is variance stabilizing.

[2] One can object to the prior parameters depending on the data, but an alternative would be to model centered data $y - \overline{y}$, in which case the prior for the intercept parameter β_0 would be zero.

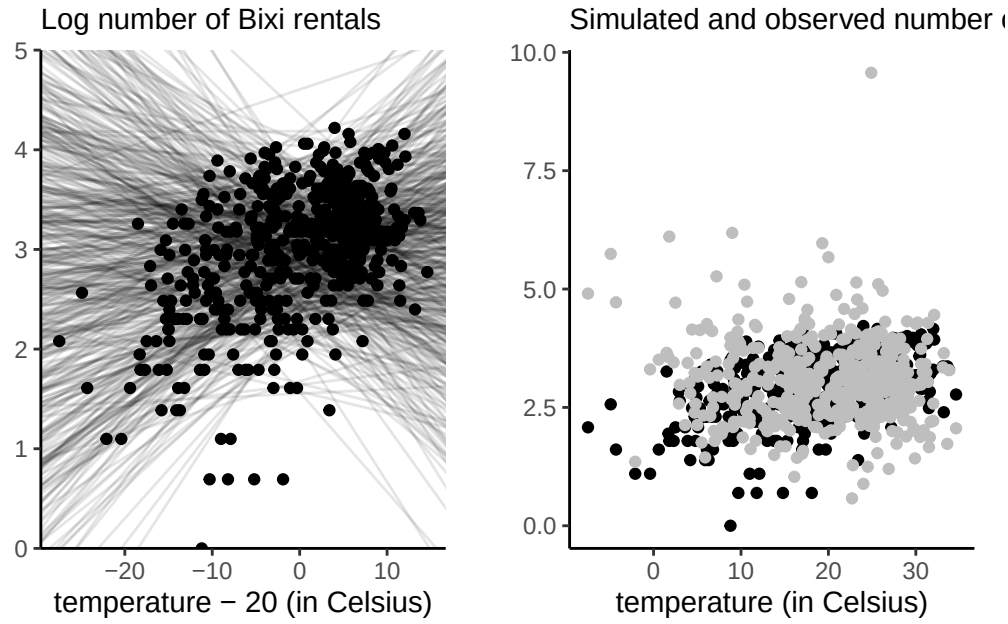

Figure 3.1: Prior draws of the linear regressions with observed data superimposed (left), and draws of observations from the prior predictive distribution (in gray) against observed data (right).

priors are chosen for computational convenience and because interpretation is convenient, as the parameters of the posterior will often be some weighted average of prior and likelihood component.

Definition 3.1. A prior density $p(\boldsymbol{\theta})$ is conjugate for likelihood $L(\boldsymbol{\theta}; \boldsymbol{y})$ if the product $L(\boldsymbol{\theta}; \boldsymbol{y})p(\boldsymbol{\theta})$, after renormalization, is of the same parametric family as the prior.

Exponential families (including the binomial, Poisson, exponential, Gaussian distributions) admit conjugate priors[3]

[3]A distribution belongs to an exponential family with parameter vector $\boldsymbol{\theta} \in \mathbb{R}^D$ if it can be written as

$$f(y; \boldsymbol{\theta}) = \exp\left\{\sum_{k=1}^{K} Q_k(\boldsymbol{\theta}) t_k(y) + D(\boldsymbol{\theta})\right\}$$

and in particular, the support does not depend on unknown parameters. If we have an independent and

Example 3.2 (Conjugate prior for the binomial model)**.** The binomial log density with y successes out of n trials is proportional to

$$y \log(p) + (n - y) \log(1 - p) = y \log\left(\frac{p}{1-p}\right) + n \log(1 - p)$$

with canonical parameter $\operatorname{logit}(p)$.[4] The binomial distribution is thus an exponential family.

Since the density of the binomial is of the form $p^y(1-p)^{n-y}$, the beta distribution $\mathsf{Beta}(\alpha, \beta)$ with density

$$f(x) \propto x^{\alpha-1}(1-x)^{\beta-1}$$

is the conjugate prior.

The beta distribution is also the conjugate prior for the negative binomial, geometric and Bernoulli distributions, since their likelihoods are all proportional to that of the beta. The fact that different sampling schemes that result in proportional likelihood functions give the same inference is called likelihood principle.

Example 3.3 (Conjugate prior for the Poisson model)**.** The Poisson distribution with mean μ has log density proportional to $f(y; \mu) \propto y \log(\mu) - \mu$, so is an exponential family with natural parameter $\log(\mu)$. The gamma density,

$$f(x) \propto \beta^\alpha / \Gamma(\alpha) x^{\alpha-1} \exp(-\beta x)$$

with shape α and rate β is the conjugate prior for the Poisson. For an n-sample of independent observations $\mathsf{Poisson}(\mu)$ observations with $\mu \sim \mathsf{Gamma}(\alpha, \beta)$, the posterior is $\mathsf{Gamma}(\sum_{i=1}^n y_i + \alpha, \beta + n)$.

Knowing the analytic expression for the posterior can be useful for calculations of the marginal likelihood, as Example 1.7 demonstrated.

---

identically distributed sample of observations $y_1, \ldots, y_n$, the log likelihood is thus of the form

$$\ell(\boldsymbol{\theta}) = \sum_{k=1}^{K} \phi_k(\boldsymbol{\theta}) \sum_{i=1}^{n} t_k(y_i) + nD(\boldsymbol{\theta}),$$

where the collection $\sum_{i=1}^n t_k(y_i)$ $(k = 1, \ldots, K)$ are sufficient statistics and $\phi_k(\boldsymbol{\theta})$ are the canonical parameters. The number of sufficient statistics are the same regardless of the sample size. Exponential families play a prominent role in generalized linear models, in which the natural parameters are modeled as linear function of explanatory variables. A log prior density with parameters $\eta, \nu_1, \ldots, \nu_K$ that is proportional to

$$\log p(\boldsymbol{\theta}) \propto \eta D(\boldsymbol{\theta}) + \sum_{k=1}^{K} Q_k(\boldsymbol{\theta}) \nu_k$$

is conjugate.

[4] The canonical link function for Bernoulli gives rise to logistic regression model.

Example 3.4 (Posterior rates for A/B tests using conjugate Poisson model)**.** Upworthy.com, a US media publisher, revolutionized headlines online advertisement by running systematic A/B tests to compare the different wording of headlines, placement and image and what catches attention the most. The Upworthy Research Archive (Matias et al. 2021) contains results for 22743 experiments, with a click through rate of 1.58% on average and a standard deviation of 1.23%. The `clickability_test_id` gives the unique identifier of the experiment, `clicks` the number of conversion out of `impressions`. See Section 8.5 of Alexander (2023) for more details about A/B testing and background information.

Consider an A/B test from November 23st, 2014, that compared four different headlines for a story on Sesame Street workshop with interviews of children whose parents were in jail and visiting them in prisons. The headlines tested were:

1. Some Don't Like It When He Sees His Mom. But To Him? Pure Joy. Why Keep Her From Him?
2. They're Not In Danger. They're Right. See True Compassion From The Children Of The Incarcerated.
3. Kids Have No Place In Jail ... But In This Case, They *Totally* Deserve It.
4. Going To Jail *Should* Be The Worst Part Of Their Life. It's So Not. Not At All.

At first glance, the first and third headlines seem likely to lead to a curiosity gap. The wording of the second is more explicit (and searchable), whereas the first is worded as a question.

We model the conversion rate λ_i for each headline separately using a Poisson distribution and compare the posterior distributions for all four choices. Using a conjugate prior and selecting the parameters by moment matching yields approximately $\alpha = 1.65$ and $\beta = 0.01$ for the hyperparameters, setting $\alpha\beta = 0.0158$ and $\alpha\beta^2 = 0.0123^2$ and solving for the two unknown parameters.

Table 3.1: Number of views, clicks for different headlines for the Upworthy data.

headline	impressions	clicks
H1	3060	49
H2	2982	20
H3	3112	31
H4	3083	9

We can visualize the posterior distributions. In this context, the large sample size lead to the dominance of the likelihood contribution $p(Y_i \mid \lambda_i) \sim \mathsf{Poisson}(n_i\lambda_i)$ relative to the prior. We can see there is virtually no overlap between different rates for headers H1 (preferred)

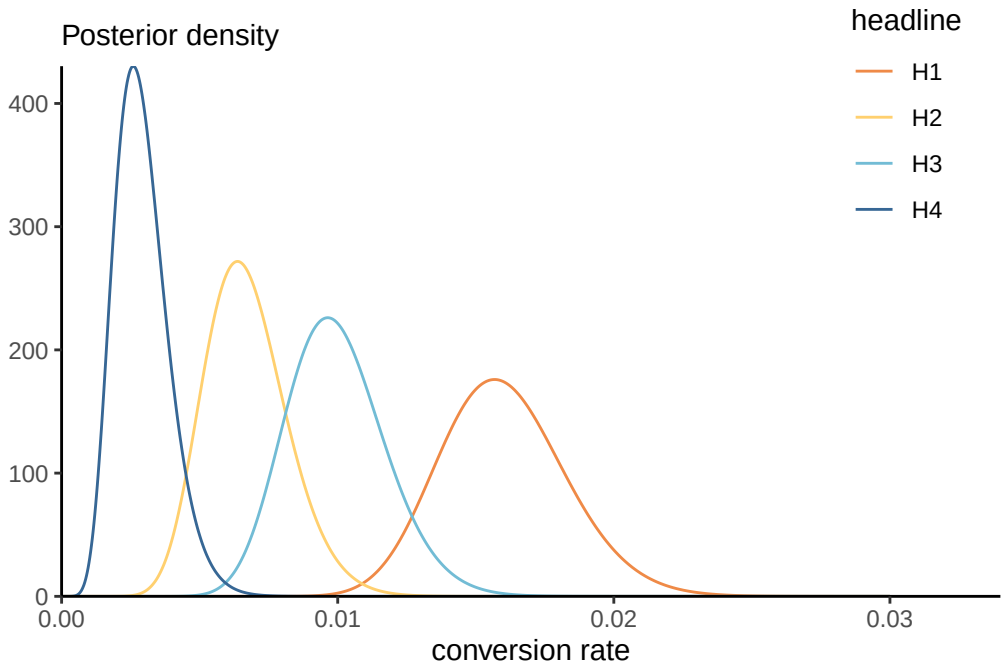

Figure 3.2: Gamma posterior for conversion rate of the different Upworthy Sesame Street headline.

relative to H4 (least favorable). The probability that the conversion rate for Headline 3 is higher than Headline 1 can be approximated by simulating samples from both posteriors and computing the proportion of times one is larger: we get 1.7% for `H3` relative to `H1`, indicating a clear preference for the first headline `H1`.

Example 3.5 (Should you phrase your headline as a question?)**.** We can also consider aggregate records for Upworthy, as Alexander (2023) did. The `upworthy_question` database contains a balanced sample of all headlines where at least one of the choices featured a question, with at least one alternative statement. Whether a headline contains a question or not is determined by querying for the question mark. We consider aggregated counts for all such headlines, with the `question` factor encoding whether there was a question, `yes` or `no`. For simplicity, we treat the number of views as fixed, but keep in mind that A/B tests are often sequential experiments with a stopping rule.[5]

We model first the rates using a Poisson regression; the corresponding frequentist analysis would include an offset to account for differences in views. If λ_j $(j = 1, 2)$ are the average

[5]The stopping rule means that data stops being collected once there is enough evidence to determine if an option is more suitable, or if a predetermined number of views has been reached.

rate for each factor level (yes and no), then $\mathsf{E}(Y_{ij}/n_{ij}) = \lambda_j$. In the frequentist setting, we can fit a simple Poisson generalized linear regression model with an offset term and a binary variable.

```
data(upworthy_question, package = "hecbayes")
poismod <- glm(
  clicks ~ offset(log(impressions)) + question,
  family = poisson(link = "log"),
  data = upworthy_question)
coef(poismod)
```

```
(Intercept)  questionno
-4.51264669  0.07069677
```

The coefficients represent the difference in log rate (multiplicative effect) relative to the baseline rate, with an increase of 6.3 percent when the headline does not contain a question. A likelihood ratio test can be performed by comparing the deviance of the null model (intercept-only), indicating strong evidence that including question leads to significatively different rates. This is rather unsurprising given the enormous sample sizes.

Consider instead a Bayesian analysis with conjugate prior: we model separately the rates of each group (question or not). Suppose we think apriori that the click-rate is on average 1%, with a standard deviation of 2%, with no difference between questions or not. For a $\mathsf{Gamma}(\alpha, \beta)$ prior, this would translate, using moment matching, into a rate of $\beta = 0.04 = \mathsf{Var}_0(\lambda_j)/\mathsf{E}_0(\lambda_j)$ and a shape of $\alpha = 2.5$ $(j = 1, 2)$. If λ_j is the average rate for each factor level (yes and no), then $\mathsf{E}(Y_{ij}/n_{ij}) = \lambda_j$ so the log likelihood is proportional, as a function of λ_1 and λ_2, to

$$\ell(\boldsymbol{\lambda}; \boldsymbol{y}, \boldsymbol{n}) \overset{\boldsymbol{\lambda}}{\propto} \sum_{i=1}^{n}\sum_{j=1}^{2} y_{ij} \log \lambda_j - \lambda_j n_{ij}$$

and we can recognize that the posterior for λ_i is gamma with shape $\alpha + \sum_{i=1}^{n} y_{ij}$ and rate $\beta + \sum_{i=1}^{n} n_{ij}$. For inference, we thus only need to select hyperparameters and calculate the total number of clicks and impressions per group. We can then consider the posterior difference $\lambda_1 - \lambda_2$ or, to mimic the Poisson multiplicative model, of the ratio λ_1/λ_2. The former suggests very small differences, but one must keep in mind that rates are also small. The ratio, shown in the right-hand panel of Figure 3.3, gives a more easily interpretable portrait that is in line with the frequentist analysis.

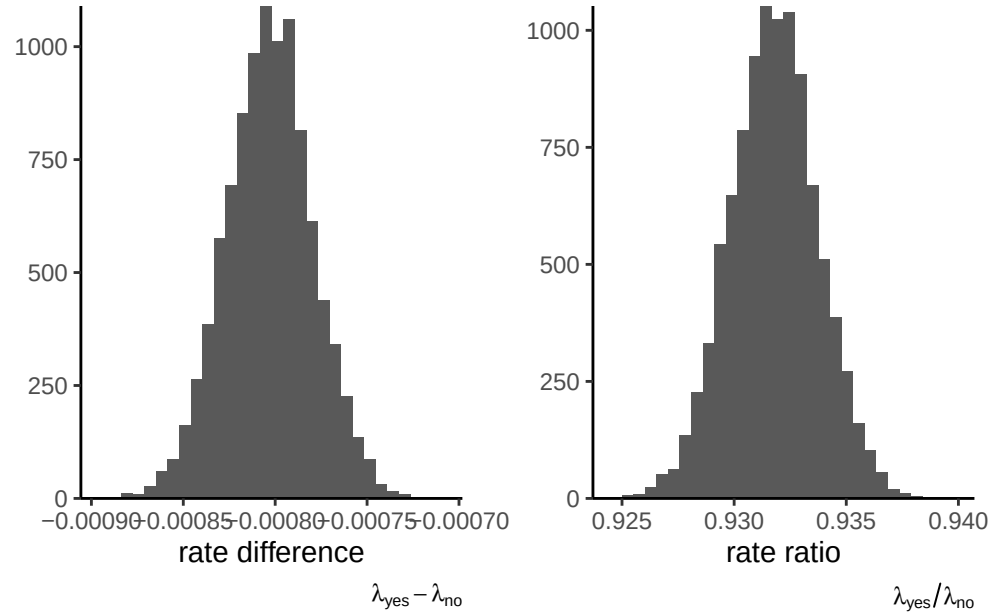

Figure 3.3: Histograms of posterior summaries for differences (left) and rates (right) based on 1000 simulations from the independent gamma posteriors.

To get an approximation to the posterior mean of the ratio λ_1/λ_2, it suffices to draw independent observations from their respective posterior, compute the ratio and take the sample mean of those draws. We can see that the sampling distribution of the ratio is nearly symmetrical, so we can expect Wald intervals to perform well should one be interested in building confidence intervals. This is however hardly surprising given the sample size at play.

Example 3.6 (Conjugate prior for Gaussian mean with known variance)**.** Consider an n simple random sample of independent and identically distributed Gaussian variables with mean μ and standard deviation σ, denoted $Y_i \sim \mathsf{Gauss}(\mu, \sigma^2)$. We pick a Gaussian prior for the location parameter, $\mu \sim \mathsf{Gauss}(\nu, \tau^2)$ where we assume μ, τ are fixed hyperparameter values. For now, we consider only inference for $p(\mu \mid \sigma)$: discarding any term that is not a function of μ, the conditional posterior is

$$\begin{aligned} p(\mu \mid \sigma) &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}\exp\left\{-\frac{1}{2\tau^2}(\mu-\nu)^2\right\} \\ &\propto \exp\left\{\left(\frac{\sum_{i=1}^n y_i}{\sigma^2}+\frac{\nu}{\tau^2}\right)\mu-\left(\frac{n}{2\sigma^2}+\frac{1}{2\tau^2}\right)\mu^2\right\}. \end{aligned}$$

The log of the posterior density conditional on σ is quadratic in μ, it must be a Gaussian distribution truncated over the positive half line. This can be seen by completing the square in μ, or by comparing this expression to the density of $\mathsf{Gauss}(\mu, \sigma^2)$,

$$f(x;\mu,\sigma) \overset{\mu}{\propto} \exp\left(-\frac{1}{2\sigma^2}\mu^2 + \frac{x}{\sigma^2}\mu\right)$$

we can deduce by matching mean and variance that the conditional posterior $p(\mu \mid \sigma)$ is Gaussian with reciprocal variance (precision) $n/\sigma^2+1/\tau^2$ and mean $(n\overline{y}\tau^2+\nu\sigma^2)/(n\tau^2+\sigma^2)$. The precision is an average of that of the prior and data, but assigns more weight to the latter, which increases linearly with the sample size n. Likewise, the posterior mean is a weighted average of prior and sample mean, with weights proportional to the relative precision.

The exponential family is quite large; Fink (1997) *A Compendium of Conjugate Priors* gives multiple examples of conjugate priors and work out parameter values.

In general, unless the sample size is small and we want to add expert opinion, we may wish to pick an *uninformative prior*, i.e., one that does not impact much the outcome. For conjugate models, one can often show that the relative weight of prior parameters (relative to the random sample likelihood contribution) becomes negligible by investigating their relative weights.

3.3 Uninformative priors

Definition 3.2 (Proper prior)**.** We call a prior function *proper* if it's integral is finite over the parameter space; such prior function automatically leads to a valid posterior.

The best example of prior priors arise from probability density function. We can still employ this rule for improper priors: for example, taking $\alpha, \beta \to 0$ in the beta prior leads to a prior proportional to $x^{-1}(1-x)^{-1}$, the integral of which diverges on the unit interval $[0,1]$. However, as long as the number of success and the number of failures is larger than 1, meaning $k \geq 1, n-k \geq 1$, the posterior distribution would be proper, i.e., integrable. To find the posterior, normalizing constants are also superfluous.

Many uninformative priors are flat, or proportional to a uniform on some subset of the real line and therefore improper. It may be superficially tempting to set a uniform prior on a large range to ensure posterior property, but the major problem is that a flat prior may be informative in a different parametrization, as the following example suggests.

Gelman et al. (2013) uses the following taxonomy for various levels of prior information: uninformative priors are generally flat or uniform priors with $p(\beta) \propto 1$, vague priors are

typically nearly flat even if proper, e.g., $\beta \sim \mathsf{Gauss}(0, 100)$, weakly informative priors provide little constraints $\beta \sim \mathsf{Gauss}(0, 10)$, and informative prior are typically application-specific, but constrain the ranges. Uninformative and vague priors are generally not recommended unless they are known to give valid posterior inference and the amount of information from the likelihood is high.

Example 3.7 (Transformation of flat prior for scales)**.** Consider the parameter $\log(\tau) \in \mathbb{R}$ and the prior $p(\log \tau) \propto 1$. If we reparametrize the model in terms of τ, the new prior (including the Jacobian of the transformation) is τ^{-1}

Some priors are standard and widely used. In location scale families with location ν and scale τ, the density is such that

$$f(x; \nu, \tau) = \frac{1}{\tau} f\left(\frac{x-\nu}{\tau}\right), \qquad \nu \in \mathbb{R}, \tau > 0.$$

We thus wish to have a prior so that $p(\tau) = c^{-1}p(\tau/c)$ for any scaling $c > 0$, whence it follows that $p(\tau) \propto \tau^{-1}$, which is uniform on the log scale.

The priors $p(\nu) \propto 1$ and $p(\tau) \propto \tau^{-1}$ are both improper but lead to location and scale invariance, hence that the result is the same regardless of the units of measurement.

One criticism of the Bayesian approach is the arbitrariness of prior functions. However, the role of the prior is often negligible in large samples (consider for example the posterior of exponential families with conjugate priors). Moreover, the likelihood is also chosen for convenience, and arguably has a bigger influence on the conclusion. Data fitted using a linear regression model seldom follow Gaussian distributions conditionally, in the same way that the linearity is a convenience (and first order approximation).

Definition 3.3 (Jeffrey's prior)**.** In single parameter models, taking a prior function for θ proportional to the square root of the determinant of the information matrix, $p(\theta) \propto |\imath(\theta)|^{1/2}$ yields a prior that is invariant to reparametrization, so that inferences conducted in different parametrizations are equivalent.[6]

To see this, consider a bijective transformation $\theta \mapsto \vartheta$. Under the reparametrized model and suitable regularity conditions[7], the chain rule implies that

$$\begin{aligned} i(\vartheta) &= -\mathsf{E}\left(\frac{\partial^2 \ell(\vartheta)}{\partial^2 \vartheta}\right) \\ &= -\mathsf{E}\left(\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right)\left(\frac{\mathrm{d}\theta}{\mathrm{d}\vartheta}\right)^2 + \mathsf{E}\left(\frac{\partial \ell(\theta)}{\partial \theta}\right)\frac{\mathrm{d}^2\theta}{\mathrm{d}\vartheta^2} \end{aligned}$$

[6]The Fisher information is linear in the sample size for independent and identically distributed data so we can derive the result for $n = 1$ without loss of generality.

[7]Using Bartlett's identity; Fisher consistency can be established using the dominated convergence theorem.

Since the score has mean zero, $\mathsf{E}\left\{\partial\ell(\theta)/\partial\theta\right\} = 0$ and the rightmost term vanishes. We can thus relate the Fisher information in both parametrizations, with

$$\imath^{1/2}(\vartheta) = \imath^{1/2}(\theta)\left|\frac{\mathrm{d}\theta}{\mathrm{d}\vartheta}\right|,$$

implying invariance.

In multiparameter models, the system isn't invariant to reparametrization if we consider the determinant of the Fisher information.

Example 3.8 (Jeffrey's prior for the binomial distribution)**.** Consider the binomial distribution $\mathsf{Bin}(1,\theta)$ with density $f(y;\theta) \propto \theta^y(1-\theta)^{1-y}\mathsf{I}_{\theta\in[0,1]}$. The negative of the second derivative of the log likelihood with respect to p is

$$\jmath(\theta) = -\partial^2\ell(\theta;y)/\partial\theta^2 = y/\theta^2 + (1-y)/(1-\theta)^2$$

and since $\mathsf{E}(Y) = \theta$, the Fisher information is

$$\imath(\vartheta) = \mathsf{E}\{\jmath(\theta)\} = 1/\theta + 1/(1-\theta) = 1/\{\theta(1-\theta)\}$$

Jeffrey's prior is thus $p(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, a conjugate Beta prior $\mathsf{Beta}(0.5, 0.5)$.

Exercise 3.1 (Jeffrey's prior for the normal distribution)**.** Check that for the Gaussian distribution $\mathsf{Gauss}(\mu,\sigma^2)$, the Jeffrey's prior obtained by treating each parameter as fixed in turn, are $p(\mu) \propto 1$ and $p(\sigma) \propto 1/\sigma$, which also correspond to the default uninformative priors for location-scale families.

Example 3.9 (Jeffrey's prior for the Poisson distribution)**.** The Poisson distribution with $\ell(\lambda) \propto -\lambda + y\log\lambda$, with second derivative $-\partial^2\ell(\lambda)/\partial\lambda^2 = y/\lambda^2$. Since the mean of the Poisson distribution is λ, the Fisher information is $\imath(\lambda) = \lambda^{-1}$ and Jeffrey's prior is $\lambda^{-1/2}$.

3.4 Informative priors

One strength of the Bayesian approach is the capability of incorporating expert and domain-based knowledge through priors. Often, these will take the form of moment constraints, so one common way to derive a prior is to perform moment matching to related ellicited quantities with moments of the prior distribution. It may be easier to set priors on a different scale than those of the observations, as Example 3.10 demonstrates.

Example 3.10 (Gamma quantile difference priors for extreme value distributions)**.** The generalized extreme value distribution arises as the limiting distribution for the maximum of m independent observations from some common distribution F. The $\mathsf{GEV}(\mu, \sigma, \xi)$ distribution is a location-scale with distribution function

$$F(x) = \exp\left[-\{1+\xi(x-\mu)/\sigma\}_+^{-1/\xi}\right]$$

where $x_+ = \max\{0, x\}$.

Inverting the distribution function yields the quantile function

$$Q(p)\mu + \sigma\frac{(-\log p)^{-\xi}-1}{\xi}$$

In environmental data, we often model annual maximum. Engineering designs are often specified in terms of the k-year return levels, defined as the quantile of the annual maximum exceeded with probability $1/k$ in any given year. Using a GEV for annual maximum, Coles and Tawn (1996) proposed modelling annual daily rainfall and specifying a prior on the quantile scale $q_1 < q_2 < q_3$ for tail probabilities $p_1 > p_2 > p_3$. To deal with the ordering constraints, gamma priors are imposed on the differences

- $q_1 - o \sim \mathsf{Gamma}(\alpha_1, \beta_1)$,
- $q_2 - q_1 \sim \mathsf{Gamma}(\alpha_2, \beta_2)$ and
- $q_3 - q_2 \sim \mathsf{Gamma}(\alpha_3, \beta_3)$,

where o is the lower bound of the support. The prior is thus of the form

$$p(\boldsymbol{q}) \propto q_1^{\alpha_1-1}\exp(-\beta_1 q_1)\prod_{i=2}^{3}(q_i-q_{i-1})^{\alpha_i-1}\exp\{\beta_i(q_i-q_{i-1})\}.$$

where $0 \leq q_1 \leq q_2 \leq q_3$. The fact that these quantities refer to moments or risk estimates which practitioners often must compute as part of regulatory requirements makes it easier to specify sensible values for hyperparameters.

Example 3.11 (Extreme rainfall in Abisko, Sweden)**.** As illustrating example, consider maximum daily cumulated rainfall in Abisko, Sweden. The time series spans from 1913 until December 2014; we compute the 102 yearly maximum, which range from 11mm to 62mm, and fit a generalized extreme value distribution to these.

For the priors, suppose an expert elicits quantiles of the 10, 50 and 100 years return levels; say 30mm, 45mm and 70mm, respectively, for the median and likewise 40mm, 70mm and 120mm for the 90% percentile of the return levels. We can compute the differences and calculate the parameters of the gamma distribution through moment-matching: this gives

roughly a shape of $\alpha_1 = 18.27$ and $\beta_1 = 0.6$, etc. Figure 3.4 shows the transfer from the prior predictive to the posterior distribution. The prior is much more dispersed and concentrated on the tail, which translates in a less peaked posterior than using a weakly informative prior (dotted line): the mode of the latter is slightly to the left and with lower density in the tail.

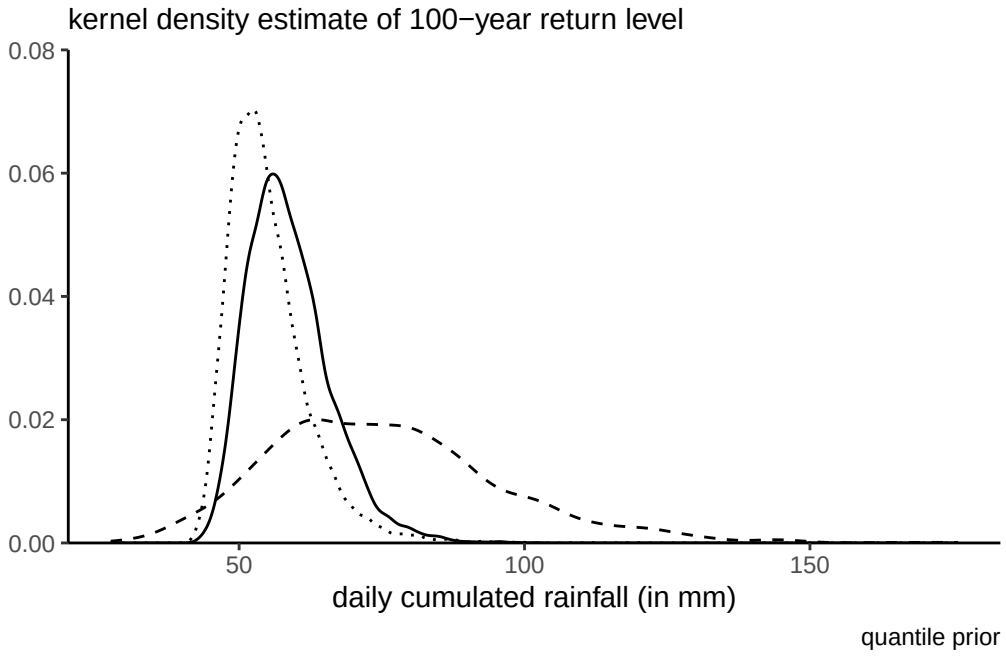

Figure 3.4: Kernel density estimates of draws from the posterior distribution of 100 year return levels with a Coles–Tawn quantile prior (full line) and from the corresponding prior predictive (dashed). The dotted line gives the posterior distribution for a maximum domain information prior on the shape with improper priors on location and scale.

What would you do if we you had prior information from different sources? One way to combine these is through a mixture: given M different prior distributions $p_m(\boldsymbol{\theta})$, we can assign each a positive weight w_m to form a mixture of experts prior through the linear combination

$$p(\boldsymbol{\theta}) \propto \sum_{m=1}^{M} w_m p_m(\boldsymbol{\theta})$$

3.4.1 Penalized complexity priors

Oftentimes, there will be a natural family of prior density to impose on some model component, $p(\boldsymbol{\theta} \mid \zeta)$, with hyperparameter ζ. The flexibility of the underlying construction leads itself to overfitting. Penalized complexity priors (Simpson et al. 2017) aim to palliate this by penalizing models far away from a simple baseline model, which correspond to a fixed value ζ_0. The prior will favour the simpler parsimonious model the more prior mass one places on ζ_0, which is in line with Occam's razor principle.

To construct a penalized-complexity prior, we compute the Kullback–Leibler divergence between the model $p_\zeta \equiv p(\boldsymbol{\theta} \mid \zeta)$ relative to the baseline with ζ_0, $p_0 \equiv p(\boldsymbol{\theta} \mid \zeta_0)$; the Kullback–Leibler divergence is

$$\mathsf{KL}(p_\zeta \| \, p_0) = \int p_\zeta \log\left(\frac{p_\zeta}{p_0}\right) \mathrm{d}\boldsymbol{\theta}.$$

The distance between the prior densities is then set to $d(\zeta) = \{2\mathsf{KL}(p_\zeta \mid\mid p_0)\}^{1/2}$. which is zero at the model with ζ_0. The PC prior then constructs an exponential prior on the distance scale, which after back-transformation gives $p(\zeta \mid \lambda) = \lambda \exp(-\lambda d(\zeta)) \, |\partial d(\zeta)/\partial \zeta|$. To choose λ, the authors recommend elicitation of a pair (U, α) such that $\Pr(\lambda > U) = \alpha$.

Example 3.12 (Penalized complexity prior for random effects models)**.** Simpson et al. (2017) give the example of a Gaussian prior for random effects $\boldsymbol{\alpha}$, of the form $\boldsymbol{\alpha} \mid \zeta \sim \mathsf{Gauss}_J(\mathbf{0}_J, \zeta^2 \mathbf{I}_J)$ where $\zeta_0 = 0$ corresponds to the absence of random subject-variability. The penalized complexity prior for the scale ζ is then an exponential with rate λ,[8] with density $p(\zeta \mid \lambda) = \lambda \exp(-\lambda\zeta)$. Using the recommendation for setting λ, we get that $\lambda = -\ln(\alpha/U)$ and this can be directly interpreted in terms of standard deviation of ζ; simulation from the prior predictive may also be used for calibration.

Example 3.13 (Penalized complexity prior for autoregressive model of order 1)**.** Sørbye and Rue (2017) derive penalized complexity prior for the Gaussian stationary AR(1) model with autoregressive parameter $\phi \in (-1, 1)$, where $Y_t \mid Y_{t-1}, \phi, \sigma^2 \sim \mathsf{Gauss}(\phi Y_{t-1}, \sigma^2)$. There are two based models that could be of interest: one with $\phi = 0$, corresponding to a memoryless model with no autocorrelation, and a static mean $\phi = 1$ for no change in time; note that the latter is not stationary. For the former, the penalized complexity prior is

$$p(\phi \mid \lambda) = \frac{\lambda}{2} \exp\left[-\lambda \left\{-\ln(1-\phi^2)\right\}^{1/2}\right] \frac{|\phi|}{(1-\phi^2)\left\{-\ln(1-\phi^2)\right\}^{1/2}}.$$

One can set λ by considering plausible values by relating the parameter to the variance of the one-step ahead forecast error.

[8] Possibly truncated above if the support of ζ has a finite upper bound.

Gaussian components are widespread: not only for linear regression models, but more generally for the specification of random effects that capture group-specific effects, residuals spatial or temporal variability. In the Bayesian paradigm, there is no difference between fixed effects $\boldsymbol{\beta}$ and the random effect parameters: both are random quantities that get assigned priors, but we will treat these priors differently.

The reason why we would like to use a penalized complexity prior for a random effect, say $\alpha_j \sim \mathsf{Gauss}(0, \zeta^2)$, is because we don't know a prior if there is variability between groups. The inverse gamma prior for ζ, $\zeta \sim \mathsf{InvGamma}(\epsilon, \epsilon)$ does not have a mode at zero unless it is improper with $\epsilon \to 0$. Generally, we want our prior for the variance to have significant probability density at the null $\zeta = 0$. The penalized complexity prior is not the only sensible choice. Posterior inference is unfortunately sensitive to the value of ϵ in hierarchical models when the random effect variance is close to zero, and more so when there are few levels for the groups since the relative weight of the prior relative to that of the likelihood contribution is then large.

Example 3.14 (Student-t prior for variance components)**.** Gelman (2006) recommends a Student-t distribution truncated below at 0, with low degrees of freedom. The rationale for this choice comes from the simple two level model with n_j independent in each group $j = 1, \ldots, J$: for observation i in group j,

$$\begin{aligned} Y_{ij} &\sim \mathsf{Gauss}(\mu + \alpha_j, \sigma^2), \\ \alpha_j &\sim \mathsf{Gauss}(0, \tau_\alpha^2), \end{aligned}$$

The conditionally conjugate prior $p(\tau \mid \boldsymbol{\alpha}, \mu, \sigma)$ is inverse gamma. Standard inference with this parametrization is however complicated, because there is strong dependence between parameters.

To reduce this dependence, one can add a parameter, taking $\alpha_j = \xi\eta_j$ and $\tau_\alpha = |\xi|\tau_\eta$; the model is now overparametrized. Suppose $\eta_j \sim \mathsf{Gauss}(0, \tau_\eta^2)$ and consider the likelihood conditional on μ, η_j: we have that $(y_{ij} - \mu)/\eta_j \sim \mathsf{Gauss}(\xi, \sigma^2/\eta_j)$ so conditionally conjugate priors for ξ and τ_η are respectively Gaussian and inverse gamma. This translates into a prior distribution for τ_α which is that of the absolute value of a noncentral Student-t with location, scale and degrees of freedom ν. If we set the location to zero, the prior puts high mass at the origin, but is heavy tailed with polynomial decay. We recommend to set degrees of freedom so that the variance is heavy-tailed, e.g., $\nu = 3$. While this prior is not conjugate, it compares favorably to the $\mathsf{inv.gamma}(\epsilon, \epsilon)$.

Example 3.15 (Poisson random effect models)**.** We consider data from an experimental study conducted at Tech3Lab on road safety. In Brodeur et al. (2021), 31 participants were

asked to drive in a virtual environment; the number of road violation was measured for different type of distractions (phone notification, phone on speaker, texting and smartwatch). The data are balanced, with each participant exposed to each task exactly once.

We model the data using a Poisson mixed model to measure the number of violations, `nviolation`, with a fixed effect for `task`, which captures the type of distraction, and a random effect for participant `id`. The hierarchical model fitted for individual i $(i = 1, \ldots, 34)$ and distraction type j $(j = 1, \ldots, 4)$ is

$$\begin{aligned} Y_{ij} &\sim \mathsf{Poisson}\{\mu = \exp(\beta_j + \alpha_i)\}, \\ \beta_j &\sim \mathsf{Gauss}(0, 100), \\ \alpha_i &\sim \mathsf{Gauss}(0, \kappa^2), \\ \kappa &\sim \mathsf{Student}_+(3). \end{aligned}$$

so observations are conditionally independent given hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

In frequentist statistics, there is a distinction made in mixed-effect models between parameters that are treated as constants, termed fixed effects and corresponding in this example to $\boldsymbol{\beta}$, and random effects, equivalent to $\boldsymbol{\alpha}$. There is no such distinction in the Bayesian paradigm, except perhaps for the choice of prior.

We can look at some of posterior distribution of the 31 random effects (here the first five individuals) and the fixed effect parameters $\boldsymbol{\beta}$, plus the variance of the random effect κ: there is strong evidence that the latter is non-zero, suggesting strong heterogeneity between individuals. The distraction which results in the largest number of violation is texting, while the other conditions all seem equally distracting on average (note that there is no control group with no distraction to compare with, so it is hard to draw conclusions).

3.5 Sensitivity analysis

Do priors matter? The answer to that question depends strongly on the model, and how much information the data provides about hyperparameters. While this question is easily answered in conjugate models (the relative weight of hyperparameters relative to data can be derived from the posterior parameters), it is not so simple in hierarchical models, where the interplay between prior distributions is often more intricate. To see the impact, one often has to rely on doing several analyses with different values fr the prior and see the sensitivity of the conclusions to these changes, for example by considering a vague prior or modifying the parameters values (say halving or doubling). If the changes are immaterial, then this provides reassurance that our analyses are robust.

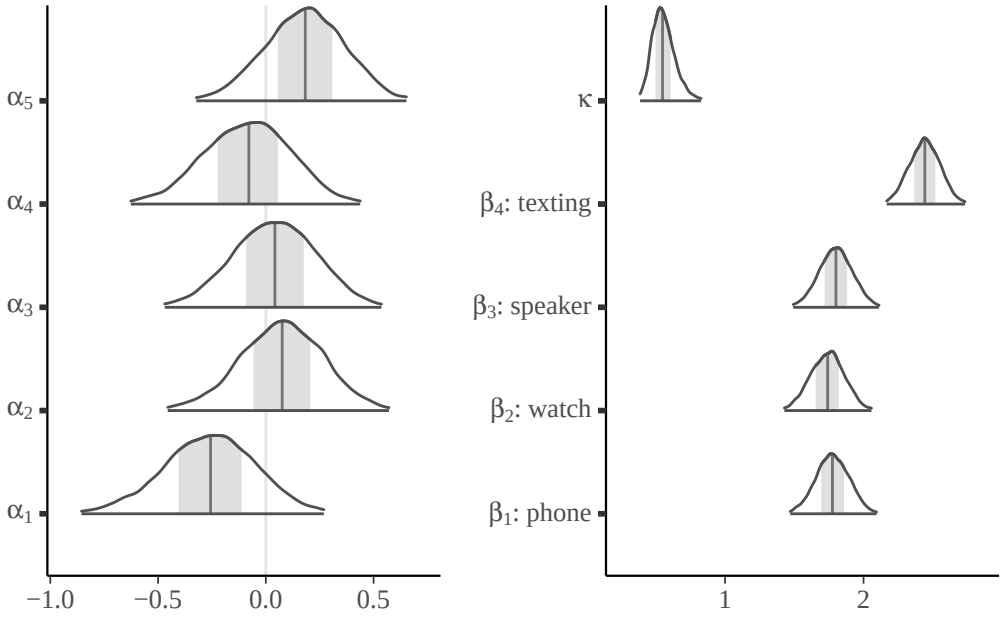

Figure 3.5: Posterior density plots with 50% credible intervals and median value for the random effects of the first five individuals (left) and the fixed effects and random effect variance (right).

Example 3.16. To check the sensitivity of the conclusion, we revisit the modelling of the `smartwatch` experiment data using a Poisson regression and compare four priors: a uniform prior truncated to $[0, 10]$, an inverse gamma $\mathsf{InvGamma}(0.01, 0.01)$ prior, a penalized complexity prior such that the 0.95 percentile of the scale is 5, corresponding to $\mathsf{Exp}(0.6)$. Since each distraction type appears 31 times, there is plenty of information to reliably estimate the dispersion κ of the random effects α: the different density plots in Figure 3.6 are virtually indistinguishable from one another. This is perhaps unsurprising given the large number of replicates, and the significant variability between groups.

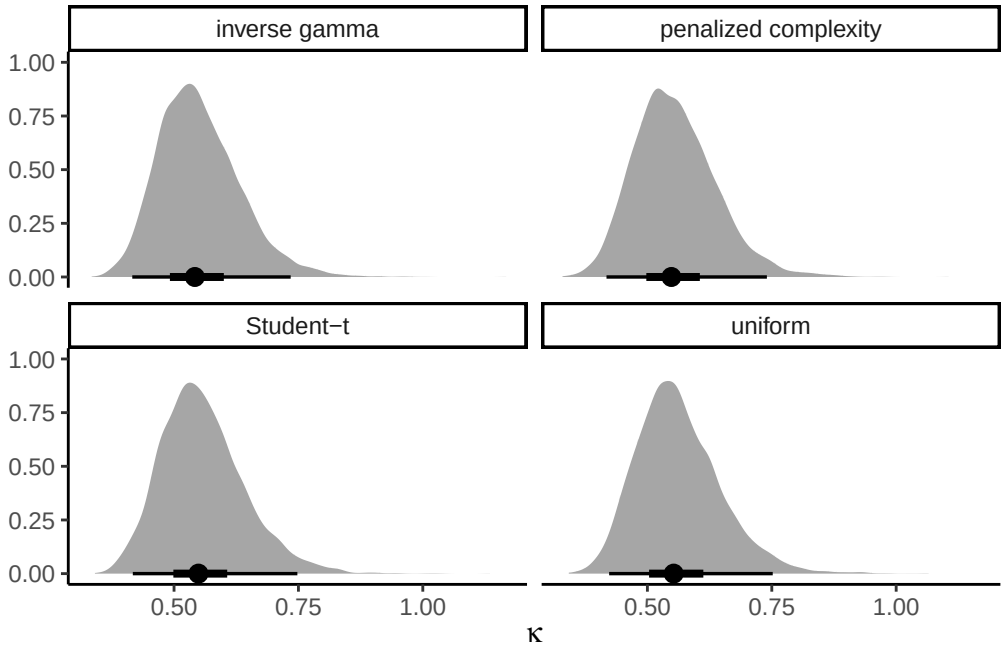

Figure 3.6: Posterior density of the scale of the random effects with uniform, inverse gamma, penalized complexity and folded Student-t with three degrees of freedom. The circle denotes the median and the bars the 50% and 95% percentile credible intervals.

4 Markov chain Monte Carlo methods

There are two major approaches to handling the problem of the unknown normalizing constant: deterministic and stochastic approximations. The former includes Laplace and nested Laplace approximations, variational methods and expectation propagation. This chapter covers the latter, stochastic approximations, and focuses on implementation of basic Markov chain Monte Carlo algorithms. The simulation algorithms circumvent the need to calculate the normalizing constant of the posterior entirely. We present several examples of implementations, several tricks for tuning and diagnostics of convergence.

Ordinary Monte Carlo methods suffer from the curse of dimensionality, with few algorithms are generic enough to be useful in complex high-dimensional problems. Instead, we will construct a Markov chain with a given invariant distribution corresponding to the posterior. Markov chain Monte Carlo methods generate correlated draws that will target the posterior under suitable conditions.[1]

4.1 Markov chains

Before going forward with algorithms for sampling, we introduce some terminology that should be familiar to people with a background in time series analysis.

Definition 4.1 (Stationarity and Markov property)**.** A stochastic (i.e., random) process is (weakly) stationary if the distribution of $\{X_1, \ldots, X_t\}$ is the same as that of $\{X_{n+1}, \ldots X_{t+n}\}$ for any value of n and given t.

It is Markov if it satisfies the Markov property: given the current state of the chain, the future only depends on the current state and not on the past.

Autoregressive processes are not the only ones we can consider, although their simplicity lends itself to analytic calculations.

[1] While we won't focus on the fine prints of the contract, there are conditions for validity and these matter!

Proposition 4.1 (Effective sample size)**.** *Intuitively, a sample of correlated observations carries less information than an independent sample of draws. If we want to compute sample averages* $\overline{Y}_T = (Y_1 + \cdots + Y_T)/T$*, the variance will be*

$$\mathsf{Va}\left(\overline{Y}_T\right) = \frac{1}{T}\sum_{t=1}^{T}\mathsf{Va}(Y_t) + \frac{2}{T}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\mathsf{Co}(Y_t, Y_s).$$

In the independent case, the covariance is zero so we get the sum of variances. If the process is stationary, the covariances at lag k *are the same regardless of the time index and the variance is some constant, say* σ^2*; this allows us to simplify calculations,*

$$\mathsf{Va}(\overline{Y}_T) = \sigma^2\left\{1 + \frac{2}{T}\sum_{t=1}^{T-1}(T-t)\mathsf{Cor}(Y_{T-k}, Y_T)\right\}.$$

Denote the lag-k *autocorrelation* $\mathsf{Cor}(Y_t, Y_{t+k})$ *by* γ_k*. Under technical conditions*[2]*, a central limit theorem applies and we get an asymptotic variance for the mean of*

$$\lim_{T\to\infty} T\mathsf{Va}\left(\overline{Y}_T\right) = \sigma^2\left\{1 + 2\sum_{t=1}^{\infty}\gamma_t\right\}.$$

This statement holds only if we start with draws from the stationary distribution, otherwise bets are off.

We need the ***effective sample size*** *of our Monte Carlo averages based on a Markov chain of length* B *to be sufficient for the estimates to be meaningful. The effective sample size is, loosely speaking, the equivalent number of observations if the marginal posterior draws where independent and more formally*

$$\mathsf{ESS} = \frac{B}{\{1 + 2\sum_{t=1}^{\infty}\gamma_t\}} \tag{4.1}$$

where γ_t *is the lag* t *correlation. The relative effective sample size is simply the fraction of the effective sample size over the Monte Carlo number of replications: small values of* ESS/B *indicate pathological or inefficient samplers. If the ratio is larger than one, it indicates the sample is superefficient (as it generates negatively correlated draws).*

In practice, we replace the unknown autocorrelations by sample estimates and truncate the series in Equation 4.1 at the point where they become negligible — typically when the consecutive sum of two consecutive becomes negative; see Section 1.4 of the Stan manual or Section 1.10.2 of Geyer (2011) for details.

[2] Geometric ergodicity and existence of moments, among other things.

Example 4.1. The lag-k correlation of the stationary autoregressive process of order 1 is ϕ^k, so summing the series gives an asymptotic variance of $\sigma^2(1+\phi)/(1-\phi)$. We can constrast that to the variance of the stationary distribution for an independent sample, which is $\sigma^2/(1-\phi^2)$. The price to pay for having correlated samples is inefficiency: the higher the autocorrelation, the larger the variability of our mean estimators.

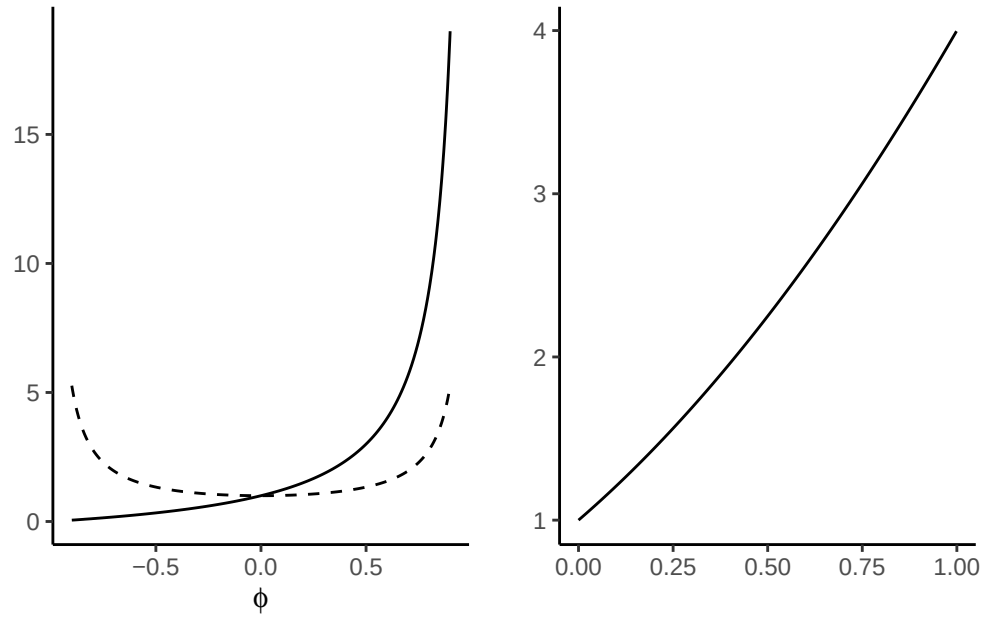

Figure 4.1: Scaled asymptotic variance of the sample mean for a stationary autoregressive first-order process with unit variance (full line) and a corresponding sample of independent observations with the same marginal variance (dashed line). The right panel gives the ratio of variances for positive correlation coefficients.

We can see from Figure 4.1 that, when the autocorrelation is positive (as will be the cause in all applications of interest), we will suffer from variance inflation. To get the same uncertainty estimates for the mean with an $\mathsf{AR}(1)$ process with $\phi \approx 0.75$ than with an iid sample, we would need nine times as many observations: this is the prize to pay.

4.1.1 Uncertainty estimation with Markov chains

With a simple random sample containing independent and identically distributed observations, the standard error of the sample mean is $\sigma/\sqrt{n}$ and we can use the empirical standard deviation $\widehat{\sigma}$ to estimate the first term. For Markov chains, the correlation prevents us from

using this approach. The output of the `coda` package are based on fitting a high order autoregressive process to the Markov chain and using the formula of the unconditional variance of the AR(p) to obtain the central limit theorem variance. An alternative method recommended by Geyer (2011) and implemented in his **R** package `mcmc`, is to segment the time series into batch, compute the means of each non-overlapping segment and use this standard deviation with suitable rescaling to get the central limit variance for the posterior mean. Figure 4.2 illustrate the method of batch means.

1. Break the chain of length B (after burn in) in K blocks of size $\approx K/B$.
2. Compute the sample mean of each segment. These values form a Markov chain and should be approximately uncorrelated.
3. Compute the standard deviation of the segments mean. Rescale by $K^{-1/2}$ to get standard error of the global mean.

Why does the approach work? If the chain samples from the stationary distribution, all samples have the same mean. If we partition the sample into long enough, the sample mean of each blocks should be roughly independent (otherwise we could remove an overlapping portion). We can then compute the empirical standard deviation of the estimators. We can then compute the overall mean and use a scaling argument to relate the variability of the global estimator with the variability of the means of the smaller blocks.

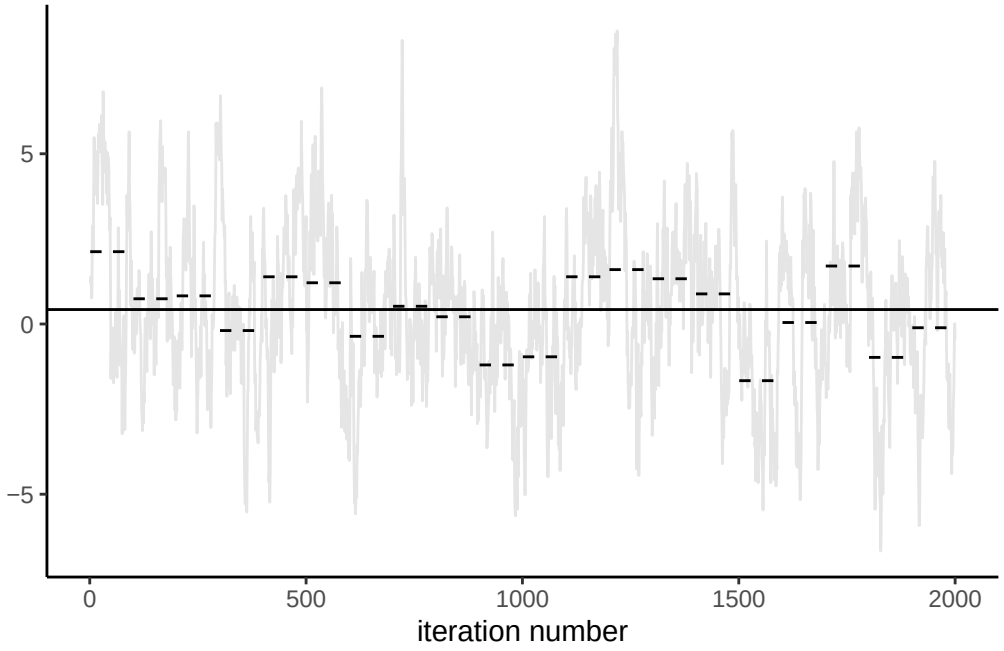

Figure 4.2: Calculation of the standard error of the posterior mean using the batch method.

When can we use output from a Markov chain in place of independent Monte Carlo draws? The assumptions laid out in the ergodic theorem are that the chain is irreducible and acyclic, ensuring that the chain has a unique stationary distribution. The ergodic theorem is a result about convergence of averages.

A discrete-time stochastic process is a random sequences whose elements are part of some set (finite or countable), termed state space $\mathcal{S}$. We can encode the probability of moving from one state to the next via a transition matrix, whose rows contain the probabilities of moving from one state to the next and thus sum to one. We can run a Markov chain by sampling an initial state X_0 at random from $\mathcal{S}$ and then consider the transitions from the conditional distribution, sampling $p(X_t \mid X_{t-1})$.

Consider a Markov chain on integers $\{1, 2, 3\}$. Because of the Markov property, the history of the chain does not matter: we only need to read the value $i = X_{t-1}$ of the state and pick the ith row of the transition matrix $\mathbf{P}$ to know the probability of the different moves from the current state.

Irreducible means that the chain can move from anywhere to anywhere, so it doesn't get stuck in part of the space forever. A transition matrix such as P_1 below describes a reducible Markov chain, because once you get into state 2 or 3, you won't escape. With reducible chains, the stationary distribution need not be unique, and so the target would depend on the starting values.

Cyclical chains loop around and visit periodically a state: P_2 is an instance of transition matrix describing a chain that cycles from 1 to 3, 3 to 2 and 2 to 1 every three iteration. An acyclic chain is needed for convergence of marginals.

$$P_1 = \begin{pmatrix} 0.5 & 0.3 & 0.2 \\ 0 & 0.4 & 0.6 \\ 0 & 0.5 & 0.5 \end{pmatrix}, \qquad P_2 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

If a chain is irreducible and aperiodic, it has a unique stationary distribution and the limiting distribution of the Markov chain will converge there. For example, we consider a transition P_3 on $1, \ldots, 5$ defined as

$$P_3 = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 0 & 0 \\ 0 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 0 \\ 0 & 0 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ 0 & 0 & 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

The stationary distribution is the value of the row vector $\boldsymbol{p}$, such that $\boldsymbol{p} = \boldsymbol{p}\mathbf{P}$ for transition matrix $\mathbf{P}$: we get $\boldsymbol{p}_1 = (0, 5/11, 6/11)$ for P_1, $(1/3, 1/3, 1/3)$ for P_2 and $(1, 2, 2, 2, 1)/8$ for P_3.

Figure 4.3 shows the path of the walk and the empirical proportion of the time spent in each state, as time progress. Since the Markov chain has a unique stationary distribution, we expect these to converge to it.

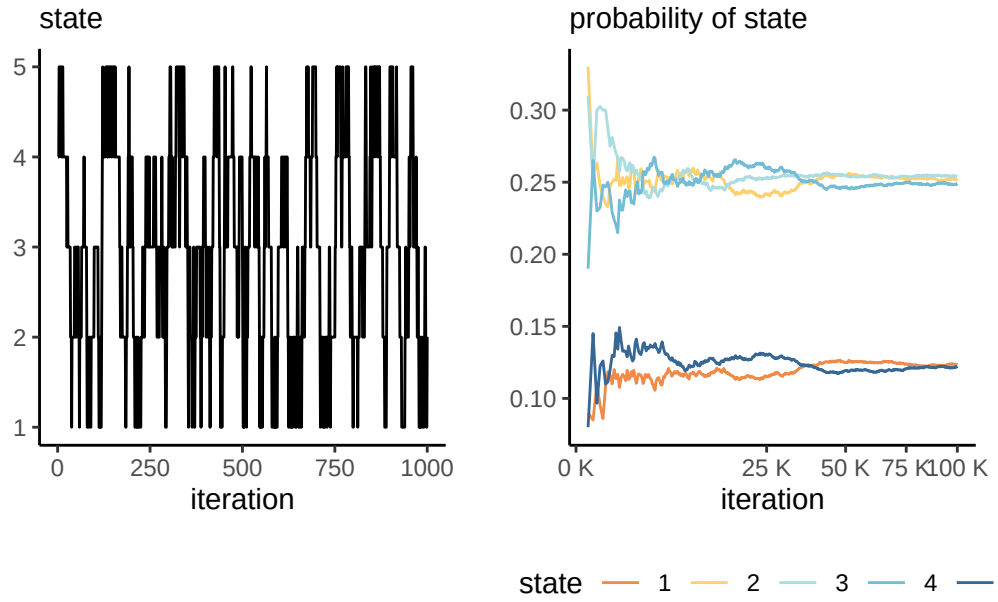

Figure 4.3: Discrete Markov chain on integers from 1 to 5, with transition matrix P_3, with traceplot of 1000 first iterations (left) and running mean plots of sample proportion of each state visited per 100 iterations (right).

4.2 Markov chain Monte Carlo algorithms

The Markov chain Monte Carlo revolution in the 1990s made Bayesian inference mainstream by allowing inference for models when only approximations were permitted, and coincided with a time at which computers became more widely available. The idea is to draw correlated samples from a posterior via Markov chains, constructed to have the posterior as invariant stationary distribution.

4.2.1 Metropolis–Hastings algorithm

Named after Metropolis et al. (1953), Hastings (1970), its relevance took a long time to gain traction in the statistical community. The idea of the Metropolis–Hastings algorithm is to construct a Markov chain targeting a distribution $p(\cdot)$.

Proposition 4.2 (Metropolis–Hastings algorithm)**.** *We consider from a density function $p(\boldsymbol{\theta})$, known up to a normalizing factor not depending on $\boldsymbol{\theta}$. We use a (conditional) proposal density $q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*)$ which has non-zero probability over the support of $p(\cdot)$, as transition kernel to generate proposals.*

The Metropolis–Hastings build a Markov chain starting from an initial value $\boldsymbol{\theta}_0$:

1. *draw a proposal value $\boldsymbol{\theta}_t^\star \sim q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{t-1})$.*
2. *Compute the acceptance ratio*

$$R = \frac{p(\boldsymbol{\theta}_t^\star)}{p(\boldsymbol{\theta}_{t-1})} \frac{q(\boldsymbol{\theta}_{t-1} \mid \boldsymbol{\theta}_t^\star)}{q(\boldsymbol{\theta}_t^\star \mid \boldsymbol{\theta}_{t-1})} \tag{4.2}$$

3. *With probability $\min\{R, 1\}$, accept the proposal and set $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_t^\star$, otherwise set the value to the previous state, $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1}$.*

The Metropolis–Hastings algorithm generates samples from the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ if the Markov chain it defines is reversible: we say it satisfies the *detailed balance condition* when the density of $\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t$, say $f(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t)$. Detailed balance means

$$f(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) p(\boldsymbol{\theta}_t \mid \boldsymbol{y}) = f(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t+1}) p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{y}).$$

This guarantees that, if $\boldsymbol{\theta}_t$ is drawn from the posterior, then the left hand side is the joint density of $(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1})$ and the marginal distribution obtained by integrating over $\boldsymbol{\theta}_t$,

$$\begin{aligned}\int f(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) p(\boldsymbol{\theta}_t \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}_t &= \int f(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t+1}) p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}_t \\ &= p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{y})\end{aligned}$$

and any draw from the posterior will generate a new realization from the posterior. We also ensure that, provided the starting value as non-zero probability under the posterior, the chain will converge to the stationarity distribution (albeit perhaps slowly).

Remark (Interpretation of the algorithm). If $R > 1$, the proposal has higher density and we always accept the move. If the ratio is less than one, the proposal is in a lower probability region, we accept the move with probability R and set $\boldsymbol{\theta}_t = \boldsymbol{\theta}_t^\star$; if we reject, the Markov chain stays at the current value, which induces autocorrelation. Since the acceptance

probability depends only on the density through ratios, we can work with unnormalized density functions and this is what allows us, if our proposal density is the (marginal) posterior of the parameter, to obtain approximate posterior samples without having to compute the marginal likelihood.

Remark (Blank run). To check that the algorithm is well-defined, we can remove the log likelihood component and run the algorithm: if it is correct, the resulting draws should be drawn from the prior provided the latter is proper (Green 2001, 55).

Remark (Symmetric proposals). Suppose we generate a candidate sample $\boldsymbol{\theta}_t^\star$ from a symmetric distribution $q(\cdot \mid \cdot)$ centered at $\boldsymbol{\theta}_{t-1}$, such as the random walk $\boldsymbol{\theta}_t^\star = \boldsymbol{\theta}_{t-1} + Z$ where Z has a symmetric distribution. Then, the proposal density ratio cancels so need not be computed in the Metropolis ratio of Equation 4.2.

Remark (Calculations). In practice, we compute the log of the acceptance ratio, $\ln R$, to avoid numerical overflow. If our target is log posterior density, we have

$$\ln\left\{\frac{p(\boldsymbol{\theta}_t^\star)}{p(\boldsymbol{\theta}_{t-1})}\right\} = \ell(\boldsymbol{\theta}_t^\star) + \ln p(\boldsymbol{\theta}_t^\star) - \ell(\boldsymbol{\theta}_{t-1}) - \ln p(\boldsymbol{\theta}_{t-1})$$

and we proceed likewise for the log of the ratio of transition kernels. We then compare the value of $\ln R$ (if less than zero) to $\log(U)$, where $U \sim \mathsf{U}(0,1)$. We accept the move if $\ln(R) > \log(U)$ and keep the previous value otherwise.

Example 4.2. Consider again the Upworthy data from Example 3.5. We model the Poisson rates λ_i $(i = 1, 2)$, this time with the usual Poisson regression parametrization in terms of log rate for the baseline yes, $\log(\lambda_2) = \beta$, and log odds rates $\kappa = \log(\lambda_1) - \log(\lambda_2)$. Our model is

$$\begin{aligned}
Y_i &\sim \mathsf{Po}(n_i\lambda_i), \qquad (i = 1, 2)\\
\lambda_1 &= \exp(\beta + \kappa)\\
\lambda_2 &= \exp(\beta)\\
\beta &\sim \mathsf{Gauss}(\log 0.01, 1.5)\\
\kappa &\sim \mathsf{Gauss}(0, 1)
\end{aligned}$$

There are two parameters in the model, which can be updated in turn or jointly.

```r
data(upworthy_question, package = "hecbayes")
# Compute sufficient statistics
data <- upworthy_question |>
  dplyr::group_by(question) |>
  dplyr::summarize(ntot = sum(impressions),
                   y = sum(clicks))
# Code log posterior as sum of log likelihood and log prior
loglik <- function(par, counts = data$y, offset = data$ntot, ...){
  lambda <- exp(c(par[1] + log(offset[1]), par[1] + par[2] + log(offset[2])))
 sum(dpois(x = counts, lambda = lambda, log = TRUE))
}
logprior <- function(par, ...){
  dnorm(x = par[1], mean = log(0.01), sd = 1.5, log = TRUE) +
    dnorm(x = par[2], log = TRUE)
}
logpost <- function(par, ...){
  loglik(par, ...) + logprior(par, ...)
}
# Compute maximum a posteriori (MAP)
map <- optim(
  par = c(-4, 0.07),
  fn = logpost,
  control = list(fnscale = -1),
  offset = data$ntot,
  counts = data$y,
  hessian = TRUE)
# Use MAP as starting value
cur <- map$par
# Compute logpost_cur - we can keep track of this to reduce calculations
logpost_cur <- logpost(cur)
# Proposal covariance
cov_map <- -2*solve(map$hessian)
chol <- chol(cov_map)

set.seed(80601)
niter <- 1e4L
chain <- matrix(0, nrow = niter, ncol = 2L)
colnames(chain) <- c("beta","kappa")
naccept <- 0L
```

```r
for(i in seq_len(niter)){
  # Multivariate normal proposal - symmetric random walk
  prop <- chol %*% rnorm(n = 2) + cur
  logpost_prop <- logpost(prop)
  # Compute acceptance ratio (no q because the ratio is 1)
  logR <- logpost_prop - logpost_cur
  if(logR > -rexp(1)){
    cur <- prop
    logpost_cur <- logpost_prop
    naccept <- naccept + 1L
  }
  chain[i,] <- cur
}
# Posterior summaries
summary(coda::as.mcmc(chain))
```

```
Iterations = 1:10000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean       SD  Naive SE Time-series SE
beta  -4.51268 0.001697 1.697e-05      6.176e-05
kappa  0.07075 0.002033 2.033e-05      9.741e-05

2. Quantiles for each variable:

          2.5%      25%      50%      75%    97.5%
beta  -4.51591 -4.51385 -4.51273 -4.51154 -4.50929
kappa  0.06673  0.06933  0.07077  0.07212  0.07463
```

```r
# Computing standard errors using batch means
sqrt(diag(mcmc::olbm(chain, batch.length = niter/40)))
```

```
[1] 5.717097e-05 8.220816e-05
```

The acceptance rate of the algorithm is 35.1% and the posterior means are $\beta = -4.51$ and $\kappa = 0.07$.

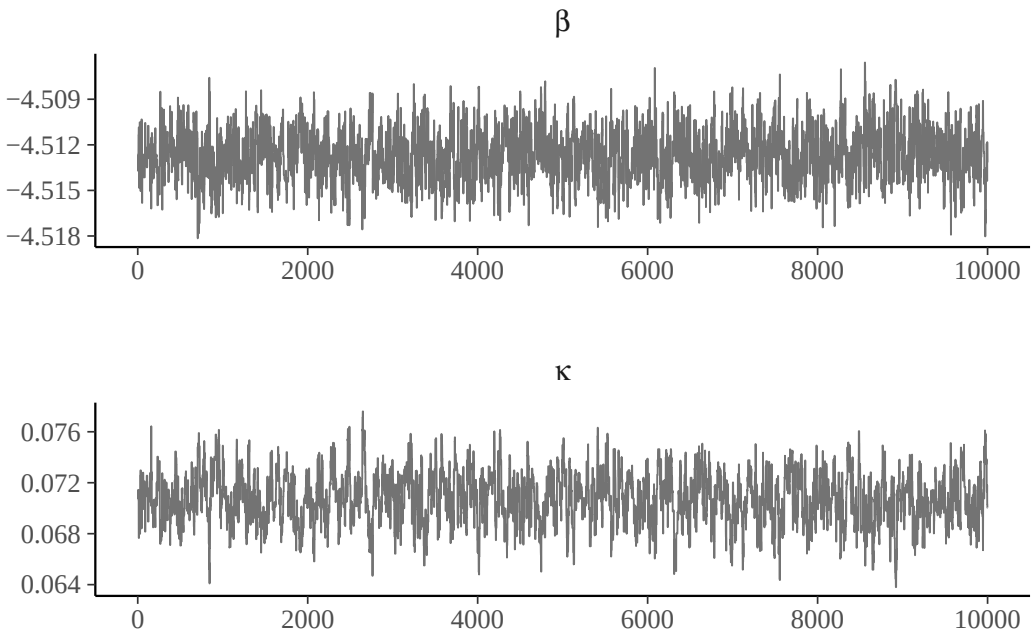

Figure 4.4: Traceplots of Markov chain of log rate and log odds rate for the Metropolis–Hastings sampler applied to the Upworthy question data.

Figure 4.5 shows the posterior samples, which are very nearly bivariate Gaussian. The parametrization in terms of log odds ratio induces strong negative dependence, so if we were to sample κ, then β, we would have much larger inefficiency and slower exploration. Instead, the code used a bivariate Gaussian random walk proposal whose covariance matrix was taken as a multiple of the inverse of the negative hessian (equivalently, to the observed information matrix of the log posterior), evaluated at of the maximum a posteriori. This Gaussian approximation is called **Laplace approximation**: it is advisable to reparametrize the model so that the distribution is nearly symmetric, so that the approximation is good. In this example, because of the large sample, the Gaussian approximation implied by Bernstein–von Mises' theorem is excellent.

The quality of the mixing of the chain (autocorrelation), depends on the proposal variance, which can obtain by trial and error. Trace plots Figure 4.4 show the values of the chain as

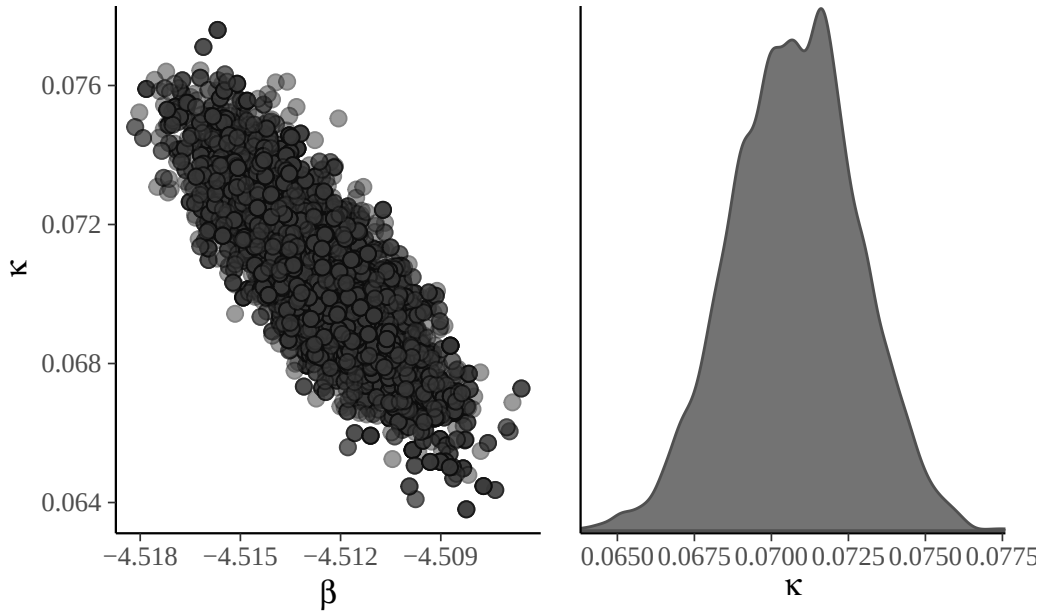

Figure 4.5: Scatterplot of posterior draws (left) and marginal density plot of log odds rate (right).

a function of iteration number. If our algorithm works well, we expect the proposals to center around the posterior mode and resemble a fat hairy caterpillar. If the variance is too small, the acceptance rate will increase but most steps will be small. If the variance of the proposal is too high, the acceptance rate will decrease (as many proposal moves will have much lower posterior), so the chain will get stuck for long periods of time. This is Goldilock's principle, as illustrated in Figure 4.6.

One way to calibrate is to track the acceptance rate of the proposals: for the three chains in Figure 4.6, these are 0.932, 0.33, 0.12. In one-dimensional toy problems with Gaussian distributions, an acceptance rate of 0.44 is optimal, and this ratio decreases to 0.234 when $D \geq 2$ (Roberts and Rosenthal 2001; Sherlock 2013). This need not generalize to other settings and depends on the context. Optimal rate for alternative algorithms, such as Metropolis-adjusted Langevin algorithm, are typically higher.

We can tune the variance of the global proposal (Andrieu and Thoms 2008) to improve the mixing of the chains at approximate stationarity. This is done by increasing (decreasing) the variance if the historical acceptance rate is too high (respectively low) during the burn in period, and reinitializing after any change with an acceptance target of 0.44. We stop adapting to ensure convergence to the posterior after a suitable number of initial iterations.

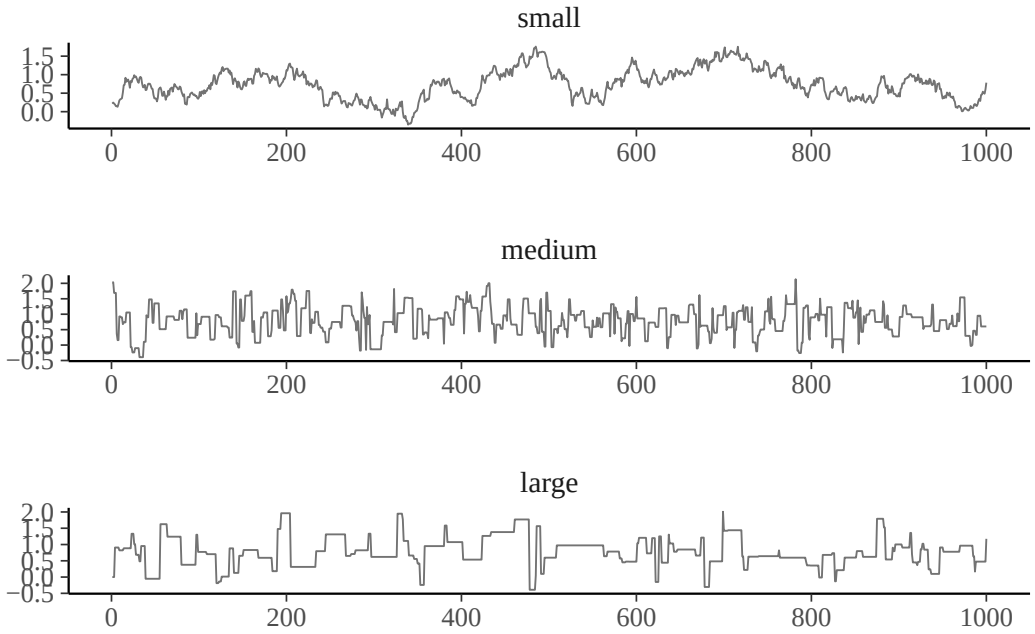

Figure 4.6: Example of traceplot with proposal variance that is too small (top), adequate (middle) and too large (bottom).

Adaptive MCMC methods use an initial warm up period to find good proposals: we can consider a block of length L, compute the acceptance rate, multiply the variance by a scaling factor and run the chain a little longer. We only keep samples obtained after the adaptation phase.

We can also plot the autocorrelation of the entries of the chain as a function of lags, a display known as correlogram in the time series literature but colloquially referred to as autocorrelation function (acf). The higher the autocorrelation, the more variance inflation one has and the longer the number of steps before two draws are treated as independent. Figure 4.7 shows the effect of the proposal variance on the correlation for the three chains. Practitioners designing very inefficient Markov chain Monte Carlo algorithms often thin their series: that is, they keep only every k iteration. This is not recommended practice unless storage is an issue and usually points towards inefficient sampling algorithms.

Remark (Independence Metropolis–Hastings). If the proposal density $q(\cdot)$ does not depend on the current state $\boldsymbol{\theta}_{t-1}$, the algorithm is termed *independence*. To maximize acceptance, we could design a candidate distribution whose mode is at the maximum a posteriori value. To efficiently explore the state space, we need to place enough density in all regions, for example by taking a heavy-tailed distributions, so that we explore the full support. Such

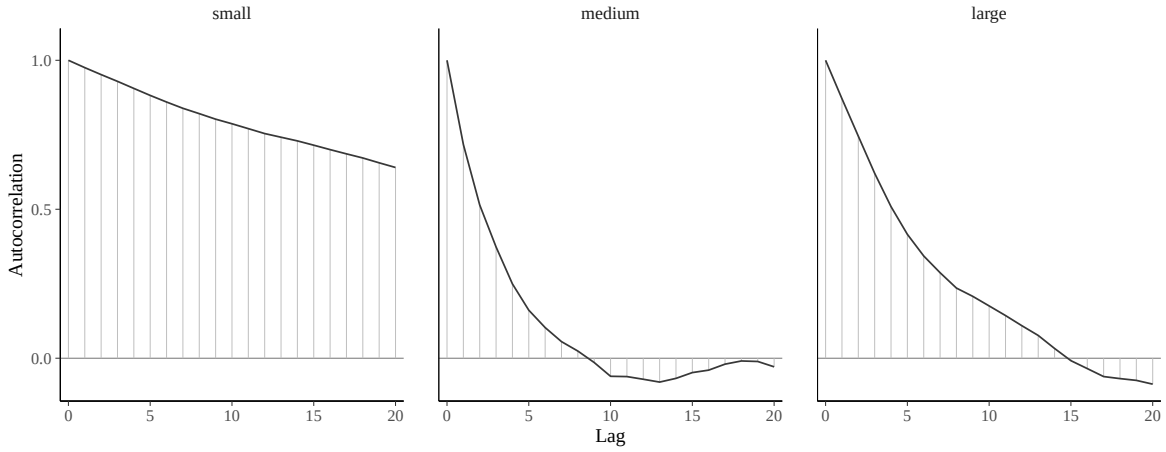

Figure 4.7: Correlogram for the three Markov chains.

proposals can be however inefficient and fail when the distribution of interest is multimodal. The independence Metropolis–Hastings algorithm then resembles accept-reject. If the ratio $p(\boldsymbol{\theta})/q(\boldsymbol{\theta})$ is bounded above by $C \geq 1$, then we can make comparisons with rejection sampling. Lemma 7.9 of Robert and Casella (2004) shows that the probability of acceptance of a move for the Markov chain is at least $1/C$, which is larger than the accept-reject.

In models with multiple parameter, we can use Metropolis–Hastings algorithm to update every parameter in turn, fixing the value of the others, rather than update them in block. The reason behind this pragmatic choice is that, as for ordinary Monte Carlo sampling, the acceptance rate goes down sharply with the dimension of the vector. Updating parameters one at a time can lead to higher acceptance rates, but slower exploration as a result of the correlation between parameters.

If we can factorize the log posterior, then some updates may not depend on all parameters: in a hierarchical model, hyperpriors parameter only appear through priors, etc. This can reduce computational costs.

Proposition 4.3 (Parameter transformation)**.** *If a parameter is bounded in the interval* (a,b)*, where* $-\infty \leq a < b \leq \infty$*, we can consider a bijective transformation* $\vartheta \equiv t(\theta) : (a,b) \to \mathbb{R}$ *with differentiable inverse. The log density of the transformed variable, assuming it exists, is*

$$f_\vartheta(\vartheta) = f_\theta\{t^{-1}(\vartheta)\}\left|\frac{\mathrm{d}}{\mathrm{d}\vartheta}t^{-1}(\vartheta)\right|$$

For example, we can use of the following transformations for finite a, b *in the software:*

- *if* $\theta \in (a,\infty)$ *(lower bound only), then* $\vartheta = \log(\theta - a)$ *and* $f_\vartheta(\vartheta) = f_\theta\{\exp(\vartheta)+a\}\cdot\exp(\vartheta)$

- *if* $\theta \in (-\infty, b)$ *(upper bound only), then* $\vartheta = \log(b-\theta)$ *and* $f_\vartheta(\vartheta) = f_\theta\{b-\exp(\vartheta)\}\cdot\exp(\vartheta)$
- *if* $\theta \in (a, b)$ *(both lower and upper bound), then* $\vartheta = \text{logit}\{(\theta-a)/(b-a)\}$ *and*

$$f_\vartheta(\vartheta) = f_\theta\{a + (b-a)\text{expit}(\vartheta)\}(b-a) \times \text{expit}(\vartheta)\{1-\text{expit}(\vartheta)\}$$

To guarantee that our proposals fall in the support of θ*, we can thus run a symmetric random walk proposal on the* transformed scale *by drawing* $\vartheta_t^\star \sim \vartheta_{t-1} + \tau Z$ *where* $Z \sim \mathsf{Gauss}(0,1)$*. Due to the transformation, the kernel ratio now contains the Jacobian.*

Proposition 4.4 (Truncated proposals)**.** *As an alternative, if we are dealing with parameters that are restricted in* $[a, b]$*, we can simulate using a random walk but with truncated Gaussian steps, taking* $\theta_t^\star \sim \mathsf{trunc.Gauss}(\vartheta_{t-1}, \tau^2, a, b)$*. The benefits of using the truncated proposal becomes more apparent when we move to more advanced proposals whose mean and variance depends on the gradient and or the hessian of the underlying unnormalized log posterior, as the mean can be lower than* a *or larger than* b*: this would garantee zero acceptance with regular Gaussian random walk. The* `TruncatedNormal` *package can be used to efficiently evaluate such instances using results from Botev and L'Écuyer (2017) even when the truncation bounds are far from the mode. the normalizing constant of the truncated Gaussian in the denominator of the density is a function of the location and scale parameters: if these depend on the current value of* $\boldsymbol{\theta}_{t-1}$*, as is the case for a random walk, we need to keep these terms as part of the Metropolis ratio. The mean and standard deviation of the truncated Gaussian are not equal to the parameters* μ *(which corresponds to the mode, provided* $a < \mu < b$*) and* σ*.*

Proposition 4.5 (Efficient proposals)**.** *Rather than simply build a random walk, we can exploit the geometry of the posterior using the gradient, via Metropolis-ajusted Langevin algorithm (MALA), or using local quadratic approximations of the target.*

Let $p(\theta)$ *denote the conditional (unnormalized) log posterior for a scalar parameter* $\theta \in (a, b)$*. We considering a Taylor series expansion of* $p(\cdot)$ *around the current parameter value* θ_{t-1}*,*

$$p(\theta) \approx p(\theta_{t-1}) + p'(\theta_{t-1})(\theta-\theta_{t-1}) + \frac{1}{2}p''(\theta_{t-1})(\theta-\theta_{t-1})^2$$

plus remainder, which suggests a Gaussian approximation with mean $\mu_{t-1} = \theta_{t-1} - f'(\theta_{t-1})/f''(\theta_{t-1})$ *and precision* $\tau^{-2} = -f''(\theta_{t-1})$*. We can use truncated Gaussian distribution on* (a, b) *with mean* μ *and standard deviation* τ*, denoted* $\mathsf{trunc.Gauss}(\mu, \tau, a, b)$ *with corresponding density function* $q(\cdot; \mu, \tau, a, b)$*. The Metropolis acceptance ratio for a proposal* $\theta_t^\star \sim \mathsf{trunc.Gauss}(\mu_{t-1}, \tau_{t-1}, a, b)$ *is*

$$\alpha = \frac{p(\theta_t^\star)}{p(\theta_{t-1})}\frac{q(\theta_{t-1} \mid \mu_t^\star, \tau_t^\star, a, b)}{q(\theta_t^\star \mid \mu_{t-1}, \tau_{t-1}, a, b)}$$

and we set $\theta^{(t+1)} = \theta_t^\star$ with probability $\min\{1, r\}$ and $\theta^{(t+1)} = \theta_{t-1}$ otherwise. To evaluate the ratio of truncated Gaussian densities $q(\cdot; \mu, \tau, a, b)$, we need to compute the Taylor approximation from the current parameter value, but also the reverse move from the proposal $\theta_t^\star$. Another option is to modify the move dictated by the rescaled gradient by taking instead

$$\mu_{t-1} = \theta_{t-1} - \eta f'(\theta_{t-1})/f''(\theta_{t-1}).$$

The proposal includes an additional learning rate parameter, $\eta \leq 1$, whose role is to prevent oscillations of the quadratic approximation, as in a Newton–Raphson algorithm. Relative to a random walk Metropolis–Hastings, the proposal automatically adjusts to the local geometry of the target, which guarantees a higher acceptance rate and lower autocorrelation for the Markov chain despite the higher evaluation costs. The proposal requires that both $f''(\theta_{t-1})$ and $f''(\theta_t^\star)$ be negative since the variance is $-1/f''(\theta)$: this shouldn't be problematic in the vicinity of the mode. Otherwise, one could use a global scaling derived from the hessian at the mode.

The simpler Metropolis-adjusted Langevin algorithm is equivalent to using a Gaussian random walk where the proposal has mean $\boldsymbol{\theta}_{t-1} + \mathbf{A}\eta\nabla \log p(\boldsymbol{\theta}_{t-1}; \boldsymbol{y})$ and variance $\tau^2\mathbf{A}$, for some mass matrix $\mathbf{A}$ and learning rate $\eta < 1$. Taking $\mathbf{A}$ as the identity matrix, which assumes the parameters are isotropic (same variance, uncorrelated) is the default choice although seldom far from optimal.

For MALA to work well, we need both to start near stationarity, to ensure that the gradient is relatively small and to prevent oscillations. One can dampen the size of the step initially if needed to avoid overshooting. The proposal variance, the other tuning parameter, is critical to the success of the algorithm. The usual target for the variance is one that gives an acceptance rate of roughly 0.574. These more efficient methods require additional calculations of the gradient and Hessian, either numerically or analytically. Depending on the situation and the computational costs of such calculations, the additional overhead may not be worth it.

Example 4.3. We revisit the Upworthy data, this time modelling each individual headline as a separate observation. We view n =`nimpression` as the sample size of a binomial distribution and `nclick` as the number of successes. Since the number of trials is large, the sample average `nclick/nimpression`, denoted y in the sequel, is approximately Gaussian. We assume that each story has a similar population rate and capture the heterogeneity inherent to each news story by treating each mean as a sample. The variance of the sample average or click rate is proportional to n^{-1}, where n is the number of impressions. To allow for underdispersion or overdispersion, we thus consider a Gaussian likelihood $Y_i \sim \mathsf{Gauss}(\mu, \sigma^2/n_i)$. We perform Bayesian inference for μ, σ after assigning a truncated Gaussian prior for $\mu \sim \mathsf{trunc.Gauss}(0.01, 0.1^2)$ over $[0, 1]$ and an penalized complexity prior for $\sigma \sim \mathsf{Exp}(0.7)$.

```r
data(upworthy_question, package = "hecbayes")
# Select data for a single question
qdata <- upworthy_question |>
  dplyr::filter(question == "yes") |>
  dplyr::mutate(y = clicks/impressions,
                no = impressions)
# Create functions with the same signature (...) for the algorithm
logpost <- function(par, data, ...){
  mu <- par[1]; sigma <- par[2]
  no <- data$no
  y <- data$y
  if(isTRUE(any(sigma <= 0, mu < 0, mu > 1))){
    return(-Inf)
  }
  dnorm(x = mu, mean = 0.01, sd = 0.1, log = TRUE) +
  dexp(sigma, rate = 0.7, log = TRUE) +
  sum(dnorm(x = y, mean = mu, sd = sigma/sqrt(no), log = TRUE))
}

logpost_grad <- function(par, data, ...){
   no <- data$no
  y <- data$y
  mu <- par[1]; sigma <- par[2]
  c(sum(no*(y-mu))/sigma^2 -(mu - 0.01)/0.01,
    -length(y)/sigma + sum(no*(y-mu)^2)/sigma^3 -0.7
  )
}

# Starting values - MAP
map <- optim(
  par = c(mean(qdata$y), 0.5),
  fn = function(x){-logpost(x, data = qdata)},
  gr = function(x){-logpost_grad(x, data = qdata)},
  hessian = TRUE,
  method = "BFGS")
# Set initial parameter values
curr <- map$par
# Check convergence
logpost_grad(curr, data = qdata)
```

```
[1] 7.650733e-03 5.575424e-05
```

```
# Compute a mass matrix
Amat <- solve(map$hessian)
# Cholesky root - for random number generation
cholA <- chol(Amat)



# Create containers for MCMC
B <- 1e4L # number of iterations
warmup <- 1e3L # adaptation period
npar <- 2L # number of parameters
prop_sd <- rep(1, npar) #updating both parameters jointly
chains <- matrix(nrow = B, ncol = npar)
damping <- 0.8 # learning rate
acceptance <- attempts <- 0
colnames(chains) <- names(curr) <- c("mu","sigma")
prop_var <- diag(prop_sd) %*% Amat %*% diag(prop_sd)
for(i in seq_len(B + warmup)){
  ind <- pmax(1, i - warmup)
  # Compute the proposal mean for the Newton step
  prop_mean <- c(curr + damping *
     Amat %*% logpost_grad(curr, data = qdata))
  # prop <- prop_sd * c(rnorm(npar) %*% cholA) + prop_mean
  prop <- c(mvtnorm::rmvnorm(
    n = 1,
    mean = prop_mean,
    sigma = prop_var))
  # Compute the reverse step
  curr_mean <- c(prop + damping *
     Amat %*% logpost_grad(prop, data = qdata))
  # log of ratio of bivariate Gaussian densities
  logmh <- mvtnorm::dmvnorm(
    x = curr, mean = prop_mean,
    sigma = prop_var,
    log = TRUE) -
    mvtnorm::dmvnorm(
      x = prop,
```

```
      mean = curr_mean,
      sigma = prop_var,
      log = TRUE) +
  logpost(prop, data = qdata) -
    logpost(curr, data = qdata)
  if(logmh > log(runif(1))){
    curr <- prop
    acceptance <- acceptance + 1L
  }
  attempts <- attempts + 1L
  # Save current value
  chains[ind,] <- curr
  if(i %% 100 & i < warmup){
    out <- hecbayes::adaptive(
      attempts = attempts,
      acceptance = acceptance,
      sd.p = prop_sd,
      target = 0.574)
    prop_sd <- out$sd
    acceptance <- out$acc
    attempts <- out$att
    prop_var <- diag(prop_sd) %*% Amat %*% diag(prop_sd)
  }
}
```

MALA requires critically a good mass matrix, especially if the gradient is very large at the starting values (often the case when the starting value is far from the mode). Given the precision of the original observations, we did not need to modify anything to deal with the parameter constraints $0 \leq \mu \leq 1$ and $\sigma > 0$, outside of encoding them in the log posterior function.

The posterior mean for the standard deviation is 0.64, which suggests overdispersion.

4.3 Gibbs sampling

The Gibbs sampling algorithm builds a Markov chain by iterating through a sequence of conditional distributions. Consider a model with $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$. We consider a single (or $m \leq p$ blocks of parameters), say $\boldsymbol{\theta}^{[j]}$, such that, conditional on the remaining components

of the parameter vector $\boldsymbol{\theta}^{-[j]}$, the conditional posterior $p(\boldsymbol{\theta}^{[j]} \mid \boldsymbol{\theta}^{-[j]}, \boldsymbol{y})$ is from a known distribution from which we can simulate draws

At iteration t, we can update each block in turn: note that the kth block uses the partially updated state

$$\boldsymbol{\theta}^{-[k]\star} = (\boldsymbol{\theta}_t^{[1]}, \ldots, \boldsymbol{\theta}_t^{[k-1]}, \boldsymbol{\theta}_{t-1}^{[k+1]}, \boldsymbol{\theta}_{t-1}^{[m]})$$

which corresponds to the current value of the parameter vector after the updates. To check the validity of the Gibbs sampler, see the methods proposed in Geweke (2004).

The Gibbs sampling can be viewed as a special case of Metropolis–Hastings where the proposal distribution q is $p(\boldsymbol{\theta}^{[j]} \mid \boldsymbol{\theta}^{-[j]\star}, \boldsymbol{y})$. The particularity is that all proposals get accepted because the log posterior of the partial update, equals the proposal distribution, so

$$R = \frac{p(\boldsymbol{\theta}_t^{[j]\star} \mid \boldsymbol{\theta}^{-[j]\star}, \boldsymbol{y})}{p(\boldsymbol{\theta}_{t-1}^{[j]\star} \mid \boldsymbol{\theta}^{-[j]\star}, \boldsymbol{y})} \frac{p(\boldsymbol{\theta}_{t-1}^{[j]\star} \mid \boldsymbol{\theta}^{-[j]\star}, \boldsymbol{y})}{p(\boldsymbol{\theta}_t^{[j]\star} \mid \boldsymbol{\theta}^{-[j]\star}, \boldsymbol{y})} = 1.$$

Regardless of the order (systematic scan or random scan), the procedure remains valid. The Gibbs sampling is thus an automatic algorithm: we only need to derive the conditional posterior distributions of the parameters and run the sampler, and there are no tuning parameter involved. If the parameters are strongly correlated, the changes for each parameter will be incremental and this will lead to slow mixing and large autocorrelation, even if the values drawn are all different. Figure 4.8 shows 25 steps from a Gibbs algorithm for a bivariate target.

As a toy illustration, we use Gibbs sampling to simulate data from a d-dimensional multivariate Gaussian target with mean $\boldsymbol{\mu}$ and equicorrelation covariance matrix $\boldsymbol{\Sigma} = (1-\rho)\mathbf{I}_d + \rho \mathbf{1}_d \mathbf{1}_d^\top$ with inverse

$$\mathbf{Q} = \boldsymbol{\Sigma}^{-1} = (1-\rho)^{-1}\left\{\mathbf{I}_d - \rho \mathbf{1}_d \mathbf{1}_d / (1 + (d-1)\rho)\right\},$$

for known correlation coefficient ρ. While we can easily sample independent observations, the exercise is insightful to see how well the methods works as the dimension increases, and when the correlation between pairs becomes stronger.

Consider $\boldsymbol{Y} \sim \mathsf{Gauss}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and a partition $(\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top)^\top$: the conditional distribution of the k subvector $\boldsymbol{Y}_1$ given the $d-k$ other components $\boldsymbol{Y}_2$ is, in terms of either the covariance (first line) or the precision (second line), Gaussian where

$$\begin{aligned}\boldsymbol{Y}_1 \mid \boldsymbol{Y}_2 = \boldsymbol{y}_2 &\sim \mathsf{Gauss}_k\left\{\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right\} \\ &\sim \mathsf{Gauss}_k\left\{\boldsymbol{\mu}_1 - \mathbf{Q}_{11}^{-1}\mathbf{Q}_{12}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2), \mathbf{Q}_{11}^{-1}\right\}.\end{aligned}$$

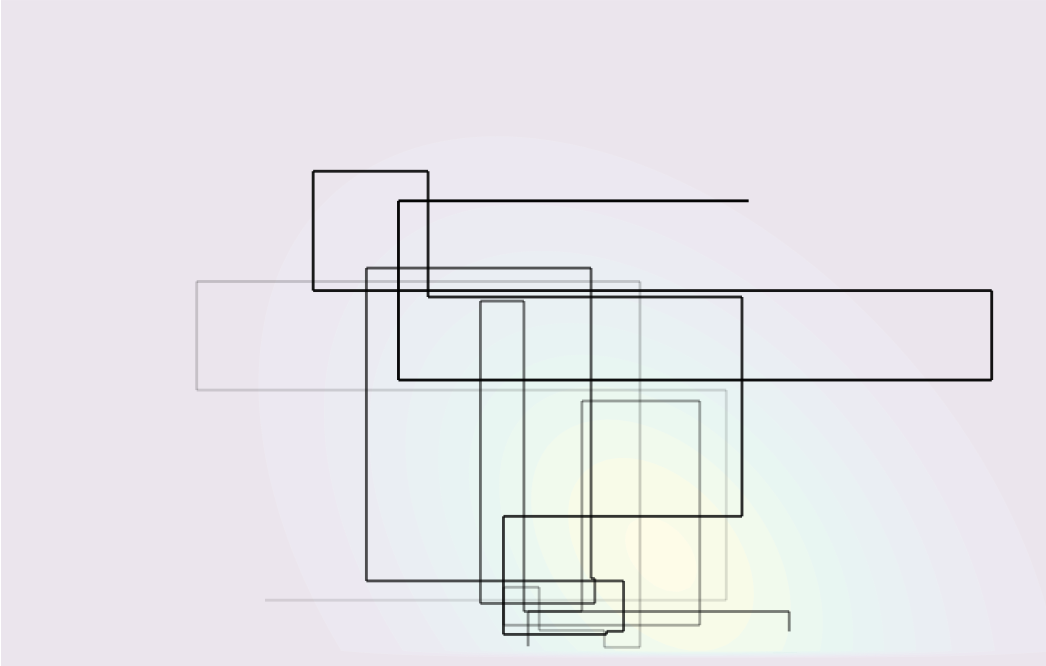

Figure 4.8: Sampling trajectory for a bivariate target using Gibbs sampling.

```
# Create a 20 dimensional equicorrelation
d <- 20
Q <- hecbayes::equicorrelation(d = d, rho = 0.9, precision = TRUE)
B <- 1e4
chains <- matrix(0, nrow = B, ncol = d)
mu <- rep(2, d)
# Start far from mode
curr <- rep(-3, d)
for(i in seq_len(B)){
  # Random scan, updating one variable at a time
  for(j in sample(1:d, size = d)){
    # sample from conditional Gaussian given curr
    curr[j] <- hecbayes::rcondmvnorm(
      n = 1,
      value = curr,
      ind = j,
      mean = mu,
```

```
      precision = Q)
  }
  chains[i,] <- curr # save values after full round of update
}
```

As the dimension of the parameter space increases, and as the correlation between components becomes larger, the efficiency of the Gibbs sampler degrades: Figure 4.9 shows the first component for updating one-parameter at a time for a multivariate Gaussian target in dimensions $d = 20$ and $d = 3$, started at four deviation away from the mode. The chain makes smaller steps when there is strong correlation, resulting in an inefficient sampler.

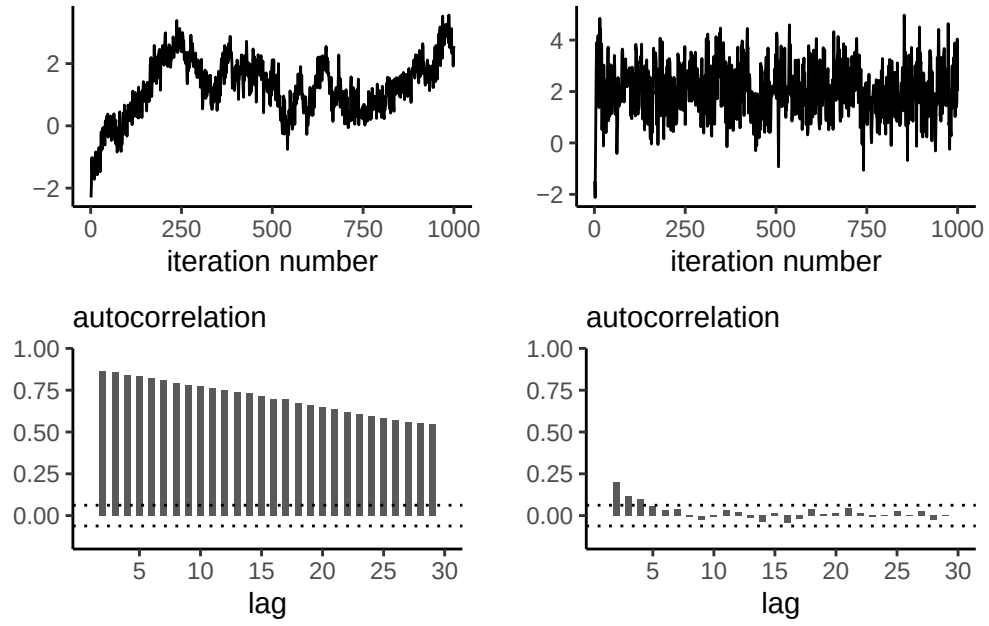

Figure 4.9: Trace plots (top) and correlograms (bottom) for the first component of a Gibbs sampler with $d = 20$ equicorrelated Gaussian variates with correlation $\rho = 0.9$ (left) and $d = 3$ with equicorrelation $\rho = 0.5$ (right).

The main bottleneck in Gibbs sampling is determining all of the relevant conditional distributions, which often relies on setting conditionally conjugate priors. In large models with multiple layers, full conditionals may only depend on a handful of parameters.

Example 4.4. Consider a Gaussian model $Y_i \sim \mathsf{Gauss}(\mu, \tau)$ $(i = 1, \ldots, n)$ are independent, and where we assign priors $\mu \sim \mathsf{Gauss}(\nu, \omega)$ and $\tau \sim \mathsf{InvGamma}(\alpha, \beta)$.

The joint posterior is not available in closed form, but the independent priors for the mean and variance of the observations are conditionally conjugate, since the joint posterior

$$p(\mu, \tau \mid \boldsymbol{y}) \propto \tau^{-n/2} \exp\left\{-\frac{1}{2\tau}\left(\sum_{i=1}^n y_i^2 - 2\mu\sum_{i=1}^n y_i + n\mu^2\right)\right\} \times \exp\left\{-\frac{(\mu-\nu)^2}{2\omega}\right\} \times \tau^{-\alpha-1}\exp(-\beta/\tau)$$

gives us

$$p(\mu \mid \tau, \boldsymbol{y}) \propto \exp\left\{-\frac{1}{2}\left(\frac{\mu^2 - 2\mu\overline{y}}{\tau/n} + \frac{\mu^2 - 2\nu\mu}{\omega}\right)\right\}$$
$$p(\tau \mid \mu, \boldsymbol{y}) \propto \tau^{-n/2-\alpha-1}\exp\left[-\left\{\frac{\sum_{i=1}^n (y_i-\mu)^2}{2} + \beta\right\}/\tau\right]$$

so we can simulate in turn

$$\mu_t \mid \tau_{t-1}, \boldsymbol{y} \sim \mathsf{Gauss}\left(\frac{n\overline{y}\omega + \tau\nu}{\tau + n\omega}, \frac{\omega\tau}{\tau + n\omega}\right)$$
$$\tau_t \mid \mu_t, \boldsymbol{y} \sim \mathsf{inv.gamma}\left\{\frac{n}{2} + \alpha, \frac{\sum_{i=1}^n (y_i-\mu)^2}{2} + \beta\right\}.$$

Remark (Gibbs sampler and proper posterior). Gibbs sampling cannot be used to determine if the posterior is improper. If the posterior is not well defined, the Markov chains may seem to stabilize even though there is no proper target.

Proposition 4.6 (Bayesian linear model)**.** *Consider a linear regression model with observation-specific mean* $\mu_i = \mathbf{x}_i\boldsymbol{\beta}$ $(i = 1, \ldots, n)$ *with* $\mathbf{x}_i$ *the* i*th row of the* $n \times p$ *model matrix* $\mathbf{X}$*.*

Concatenating records, $\boldsymbol{Y} \sim \mathsf{No}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{Q}_y^{-1})$*, for a known precision matrix* $\mathbf{Q}_y$*, typically* $\mathbf{I}_n$*. To construct a conjugate joint prior for* $p(\boldsymbol{\beta}, \sigma^2)$*, we consider the sequential formulation*

$$\boldsymbol{\beta} \mid \sigma^2 \sim \mathsf{Gauss}_p(\boldsymbol{\nu}_\beta, \sigma^2\mathbf{Q}_\beta^{-1}), \qquad \sigma^2 \sim \mathsf{InvGamma}(\alpha, \beta)$$

where InvGamma *denotes the inverse gamma distribution*[3]

The joint posterior is Gaussian-inverse gamma and can be factorized

$$p(\boldsymbol{\beta}, \sigma^2 \mid y) = p(\sigma^2 \mid y)p(\boldsymbol{\beta} \mid \sigma^2, y)$$

[3]This simply means that the precision σ^{-2}, the reciprocal of the variance, has a gamma distribution with shape α and rate β.

where $p(\sigma^2 \mid y) \sim \mathsf{InvGamma}(\alpha^*, \beta^*)$ *and* $p(\boldsymbol{\beta} \mid \sigma^2, y) \sim \mathsf{No}_p(\mathbf{M}\boldsymbol{m}, \sigma^2\mathbf{M})$ *with* $\alpha^* = \alpha + n/2$, $\beta^* = \beta + 0.5\boldsymbol{\nu}_\beta^\top \mathbf{Q}_\beta \boldsymbol{\nu}_\beta + \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{m}^\top \mathbf{M}\boldsymbol{m}$, $\boldsymbol{m} = \mathbf{Q}_\beta \boldsymbol{\nu}_\beta + \mathbf{X}^\top \mathbf{Q}_y \boldsymbol{y}$ *and* $\mathbf{M} = (\mathbf{Q}_\beta + \mathbf{X}^\top \mathbf{Q}_y \mathbf{X})^{-1}$; *the latter can be evaluated efficiently using Shermann–Morrisson–Woodbury identity. Given the conditionally conjugate priors, we can easily sample from the posterior using Gibbs sampling.*

4.3.1 Data augmentation and auxiliary variables

In many problems, the likelihood $p(\boldsymbol{y}; \boldsymbol{\theta})$ is intractable or costly to evaluate and auxiliary variables are introduced to simplify calculations, as in the expectation-maximization algorithm. The Bayesian analog is data augmentation (Tanner and Wong 1987), which we present succinctly: let $\boldsymbol{\theta} \in \Theta$ be a vector of parameters and consider auxiliary variables $\boldsymbol{u} \in \mathbb{R}^k$ such that $\int_{\mathbb{R}^k} p(\boldsymbol{u}, \boldsymbol{\theta}; \boldsymbol{y}) \mathrm{d}\boldsymbol{u} = p(\boldsymbol{\theta}; \boldsymbol{y})$, i.e., the marginal distribution is that of interest, but evaluation of $p(\boldsymbol{u}, \boldsymbol{\theta}; \boldsymbol{y})$ is cheaper. The data augmentation algorithm consists in running a Markov chain on the augmented state space $(\Theta, \mathbb{R}^k)$, simulating in turn from the conditionals $p(\boldsymbol{u}; \boldsymbol{\theta}, \boldsymbol{y})$ and $p(\boldsymbol{\theta}; \boldsymbol{u}, \boldsymbol{y})$ with new variables chosen to simplify the likelihood. If simulation from the conditionals is straightforward, we can also use data augmentation to speed up calculations or improve mixing. For more details and examples, see Dyk and Meng (2001) and Hobert (2011).

Example 4.5. Consider binary responses $\boldsymbol{Y}_i$, for which we postulate a probit regression model,

$$p_i = \Pr(Y_i = 1) = \Phi(\beta_0 + \beta_1 \mathrm{X}_{i1} + \cdots + \beta_p \mathrm{X}_{ip}),$$

where Φ is the distribution function of the standard Gaussian distribution. The likelihood of the probit model for a sample of n independent observations is

$$L(\boldsymbol{\beta}; \boldsymbol{y}) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1 - y_i},$$

and this prevents easy simulation. We can consider a data augmentation scheme where $Y_i = \mathsf{I}(Z_i > 0)$, where $Z_i \sim \mathsf{Gauss}(\mathbf{x}_i \boldsymbol{\beta}, 1)$, with $\mathbf{x}_i$ denoting the ith row of the design matrix.

The augmented data likelihood is

$$p(\boldsymbol{z}, \boldsymbol{y} \mid \boldsymbol{\beta}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{z} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{z} - \mathbf{X}\boldsymbol{\beta})\right\} \times \prod_{i=1}^{n} \mathsf{I}(z_i > 0)^{y_i} \mathsf{I}(z_i \leq 0)^{1 - y_i}$$

Given Z_i, the coefficients $\boldsymbol{\beta}$ are simply the results of ordinary linear regression with unit variance, so

$$\boldsymbol{\beta} \mid \boldsymbol{z}, \boldsymbol{y} \sim \mathsf{Gauss}\left\{\widehat{\boldsymbol{\beta}}, (\mathbf{X}^\top \mathbf{X})^{-1}\right\}$$

with $\widehat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top z$ is the ordinary least square estimator from the regression with model matrix $\mathbf{X}$ and response vector z. The augmented variables Z_i are conditionally independent and truncated Gaussian with

$$Z_i \mid y_i, \beta \sim \begin{cases} \text{trunc.Gauss}(\mathbf{x}_i\beta, -\infty, 0) & y_i = 0 \\ \text{trunc.Gauss}(\mathbf{x}_i\beta, 0, \infty) & y_i = 1. \end{cases}$$

and we can use the algorithms of Example 1.13 to simulate these.

```
probit_regression <- function(y, x, B = 1e4L, burnin = 100){
  y <- as.numeric(y)
  n <- length(y)
  # Add intercept
  x <- cbind(1, as.matrix(x))
  xtxinv <- solve(crossprod(x))
  # Use MLE as initial values
  beta.curr <- coef(glm(y ~ x - 1, family=binomial(link = "probit")))
  # Containers
  Z <- rep(0, n)
  chains <- matrix(0, nrow = B, ncol = length(beta.curr))
  for(b in seq_len(B + burnin)){
    ind <- max(1, b - burnin)
    Z <- TruncatedNormal::rtnorm(
      n = 1,
      mu = as.numeric(x %*% beta.curr),
      lb = ifelse(y == 0, -Inf, 0),
      ub = ifelse(y == 1, Inf, 0),
      sd = 1)
    beta.curr <- chains[ind,] <- as.numeric(
      mvtnorm::rmvnorm(
        n = 1,
        mean = coef(lm(Z ~ x - 1)),
        sigma = xtxinv))
  }
return(chains)
}
```

Example 4.6 (Bayesian LASSO)**.** The Laplace distribution with mean μ and scale σ, which

has density

$$f(x;\mu,\sigma) = \frac{1}{2\sigma}\exp\left(-\frac{|x-\mu|}{\sigma}\right),$$

can be expressed as a scale mixture of Gaussians, where $Y \sim \mathsf{La}(\mu,\sigma)$ is equivalent to $Z \mid \tau \sim \mathsf{Gauss}(\mu,\tau)$ and $\tau \sim \mathsf{Exp}\{(2\sigma)^{-1}\}$. With the improper prior $p(\mu,\sigma) \propto \sigma^{-1}$ and with n independent and identically distributed Laplace variates, the joint posterior can be written

$$\begin{aligned} p(\boldsymbol{\tau},\mu,\sigma \mid \boldsymbol{y}) &\propto \left(\prod_{i=1}^{n}\tau_i\right)^{-1/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\tau_i}\right\} \\ &\quad\times\frac{1}{\sigma^{n+1}}\exp\left(-\frac{1}{2\sigma}\sum_{i=1}^{n}\tau_i\right) \end{aligned}$$

and $\mu \mid \cdots$ and $\sigma \mid \cdots$ are, as usual, Gaussian and inverse gamma, respectively. The variances, τ_j, are conditionally independent of one another with

$$p(\tau_j \mid \mu,\sigma,y_j) \propto \tau_j^{-1/2}\exp\left\{-\frac{1}{2}\frac{(y_j-\mu)^2}{\tau_j} - \frac{1}{2}\frac{\tau_j}{\sigma}\right\}$$

so with $\xi_j = 1/\tau_j$, we have

$$p(\xi_j \mid \mu,\sigma,y_j) \propto \xi_j^{-3/2}\exp\left\{-\frac{1}{2\sigma}\frac{\xi_j(y_j-\mu)^2}{\sigma} - \frac{1}{2}\frac{1}{\xi_j}\right\}$$

and we recognize the latter as a Wald (or inverse Gaussian) distribution, whose density function is

$$\begin{aligned} f(y;\nu,\lambda) &= \left(\frac{\lambda}{2\pi y^3}\right)^{1/2}\exp\left\{-\frac{\lambda(y-\nu)^2}{2\nu^2 y}\right\}, \quad y>0 \\ &\overset{y}{\propto} y^{-3/2}\exp\left\{-\frac{\lambda}{2}\left(\frac{y}{\nu}+\frac{1}{y}\right)\right\} \end{aligned}$$

for location $\nu > 0$ and shape $\lambda > 0$, where $\xi_i \sim \mathsf{Wald}(\nu_i,\lambda)$ with $\nu_i = \{\sigma/(y_i-\mu)^2\}^{1/2}$ and $\lambda = \sigma^{-1}$.

Park and Casella (2008) use this hierarchical construction to defined the Bayesian LASSO. With a model matrix $\mathbf{X}$ whose columns are standardized to have mean zero and unit standard deviation, we may write

$$\begin{aligned} \boldsymbol{Y} \mid \mu,\boldsymbol{\beta},\sigma^2 &\sim \mathsf{Gauss}_n(\mu\mathbf{1}_n + \mathbf{X}\boldsymbol{\beta},\sigma\mathbf{I}_n) \\ \beta_j \mid \sigma,\tau &\sim \mathsf{Gauss}(0,\sigma\tau) \\ \tau &\sim \mathsf{Exp}(\lambda/2) \end{aligned}$$

If we set an improper prior $p(\mu, \sigma) \propto \sigma^{-1}$, the resulting conditional distributions are all available and thus the model is amenable to Gibbs sampling.

The Bayesian LASSO places a Laplace penalty on the regression coefficients, with lower values of λ yielding more shrinkage. Figure 4.10 shows a replication of Figure 1 of Park and Casella (2008), fitted to the `diabetes` data. Note that, contrary to the frequentist setting, none of the posterior draws of β are exactly zero.

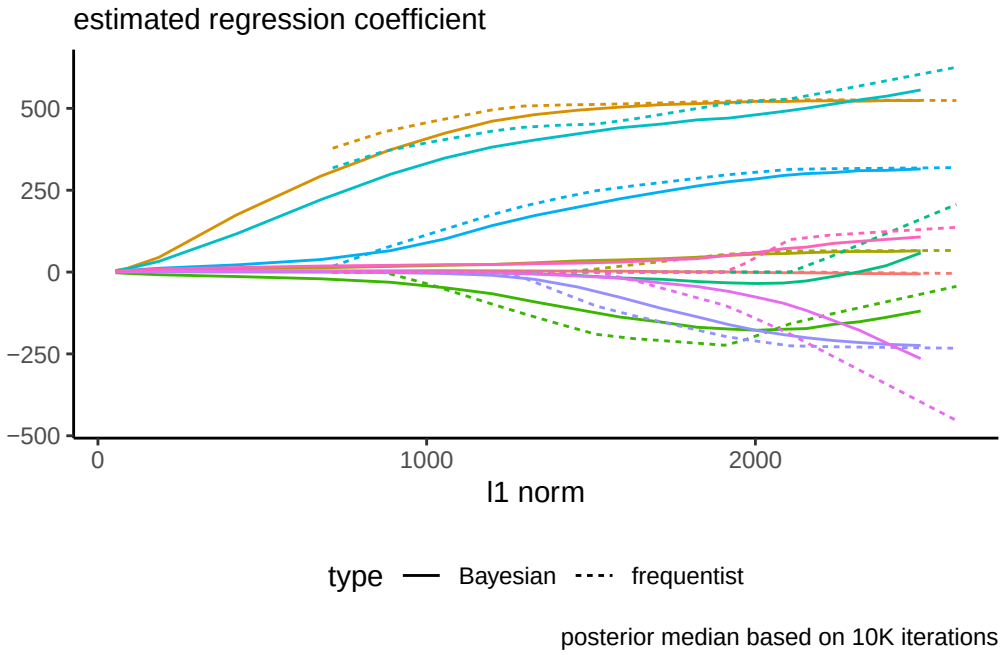

Figure 4.10: Traceplot of β coefficients (penalized maximum likelihood estimates and median aposteriori as a function of the l_1 norm of the coefficients, with lower values of the latter corresponding to higher values of the penalty λ.

Many elliptical distributions can be cast as scale mixture models of spherical or Gaussian variables; see, e.g., Section 10.2 of Albert (2009) for a similar derivation with a Student-t distribution.

Example 4.7 (Mixture models)**.** In clustering problems, we can specify that observations arise from a mixture model with a fixed or unknown number of coefficients: the interest lies then in estimating

A K-mixture model is a weighted combination of models frequently used in clustering or to

model subpopulations with respective densities f_k, with density

$$f(x; \boldsymbol{\theta}, \boldsymbol{\omega}) = \sum_{k=1}^{K} \omega_k f_k(x; \boldsymbol{\theta}_k), \qquad \omega_1 + \cdots \omega_K = 1.$$

Since the density involves a sum, numerical optimization is challenging. Let C_i denote the cluster index for observation i: if we knew the value of $C_i = j$, the density would involve only f_j. We can thus use latent variables representing the group allocation to simplify the problem and run an EM algorithm or use the data augmentation. In an iterative framework, we can consider the complete data as the tuples (X_i, Z_i), where $Z_i = \mathsf{I}(C_i = k)$.

With the augmented data, the conditional distribution of $Z_i \mid X_i, \boldsymbol{\omega}, \boldsymbol{\theta} \sim \mathsf{Multinom}(1, \boldsymbol{\gamma}_{ik})$ where

$$\gamma_{ik} = \frac{\omega_k f_k(X_i \boldsymbol{\theta}_k)}{\sum_{j=1}^{K} f_j(X_i \boldsymbol{\theta}_k)}.$$

Given suitable priors for the probabilities $\boldsymbol{\omega}$ and $\boldsymbol{\theta} \equiv \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\}$, we can use Gibbs sampling updating $\boldsymbol{Z}$, $\boldsymbol{\omega}$ and $\boldsymbol{\theta}$ in turn.

4.4 Bayesian workflow and diagnostics for Markov chains

For a given problem, there are many different Markov chain Monte Carlo algorithms that one can implement: they will typically be distinguished based on the running time and the efficiency (with algorithms providing chains that have low autocorrelation being better). Many visual diagnostics and standard tests can be used to diagnose lack of convergence, or inefficiency. The purpose of this section is to review these in turn.

The Bayesian workflow is a coherent framework for model construction, estimation and validation. It typically involves multiple iterations tuning, adapting and modifying both the models and the algorithms in the hope of achieving a model that is useful (Gelman et al. 2020); see also Michael Betancourt for excellent visualizations.

To illustrate these, we revisit the model from Example 3.15 with a penalized complexity prior for the individual effect α_i and vague normal priors. We also fit a simple Poisson model with only the fixed effect, taking $Y_{ij} \sim \mathsf{Poisson}\{\exp(\beta_j)\}$ with $\beta_j \sim \mathsf{Gauss}(0, 100)$ has much too little variability relative to the observations.

4.4.1 Trace plots

It is useful to inspect visually the Markov chain, as it may indicate several problems. If the chain drifts around without stabilizing around the posterior mode, then we can suspect

that it hasn't reached it's stationary distribution (likely due to poor starting values). In such cases, we need to disregard the dubious draws from the chain by discarding the so-called warm up or **burn in** period. While there are some guarantees of convergence in the long term, silly starting values may translate into tens of thousands of iterations lost wandering around in regions with low posterior mass. Preliminary optimization and plausible starting values help alleviate these problems. Figure 4.11 shows the effect of bad starting values on a toy problem where convergence to the mode is relatively fast. If the proposal is in a flat region of the space, it can wander around for a very long time before converging to the stationary distribution.

If we run several chains, as in Figure 4.11, with different starting values, we can monitor convergence by checking whether these chains converge to the same target. A **trace rank** plots, shown on right panel of Figure 4.11, compares the rank of the values of the different chain at a given iteration: with good mixing, the ranks should switch frequently and be distributed uniformly across integers.

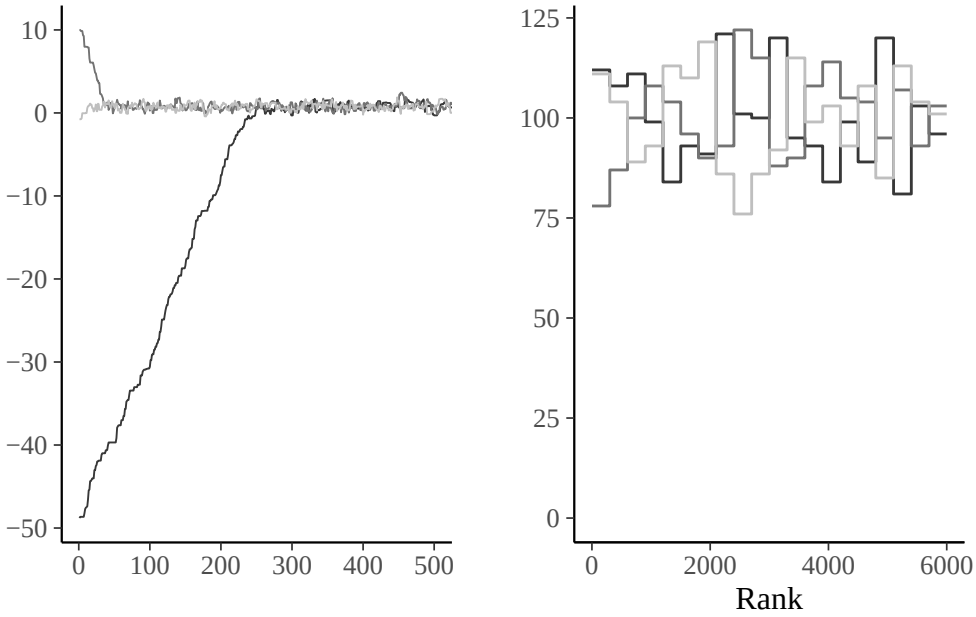

Figure 4.11: Traceplots of three Markov chains for the same target with different initial values for the first 500 iterations (left) and trace rank plot after discarding these (right).

4.4.2 Diagnostics of convergence

Generally, one would run a MCMC algorithm. The first iterations, used during the burn in period to tune proposal variances and allow the chains to converge to the stationary distribution, are discarded. If visual inspection of the chains reveal that some of the chains for one or more parameters are not stationary until some iteration, we will discard all of these in addition.

The target of inference is functional (i.e., one-dimensional summaries of the chain): we need to have convergence of the latter, but also sufficient effective sample size for our averages to be accurate (at least to two significant digits).

For the Poisson example, the effective sample size for the β for the multilevel model is a bit higher than 1000 with $B = 5000$ iterations, whereas we have for the simple naive model is 10^4 for $B = 10000$ draws, suggesting superefficient sampling. The dependency between $\boldsymbol{\alpha}$ and β is responsible for the drop in accuracy.

The `coda` (convergence diagnosis and output analysis) **R** package contains many tests. For example, the Geweke Z-score compares the averages for the beginning and the end of the chain: rejection of the null implies lack of convergence, or poor mixing.

Running multiple Markov chains can be useful for diagnostics. The Gelman–Rubin diagnostic $\widehat{R}$, also called potential scale reduction statistic, is obtained by considering the difference between within-chains and between-chains variance. Suppose we run M chains for B iterations, post burn in. Denoting by θ_{bm} the bth draw of the mth chain, we compute the global average $\overline{\theta} = B^{-1}M^{-1}\sum_{b=1}^{B}\sum_{m=1}^{m}\theta_{bm}$ and similarly the chain sample average and variances, respectively $\overline{\theta}_m$ and $\widehat{\sigma}^2_m$ $(m = 1, \ldots, M)$. The between-chain variance and within-chain variance estimator are

$$\mathsf{Va}_{\text{between}} = \frac{B}{M-1}\sum_{m=1}^{M}(\overline{\theta}_m - \overline{\theta})^2$$
$$\mathsf{Va}_{\text{within}} = \frac{1}{M}\sum_{m=1}^{m}\widehat{\sigma}^2_m$$

and we can compute

$$\widehat{R} = \left(\frac{\mathsf{Va}_{\text{within}}(B-1) + \mathsf{Va}_{\text{between}}}{B\mathsf{Va}_{\text{within}}}\right)^{1/2}$$

The potential scale reduction statistic must be, by construction, larger than 1 in large sample. Any value larger than this is indicative of problems of convergence. While the Gelman–Rubin diagnostic is frequently reported, and any value larger than 1 deemed problematic, it is not enough to have approximately $\widehat{R} = 1$ to guarantee convergence, but large values are

usually indication of something being amiss. Figure 4.12 shows two instances where the chains are visually very far from having the same average and this is reflected by the large values of $\widehat{R}$.

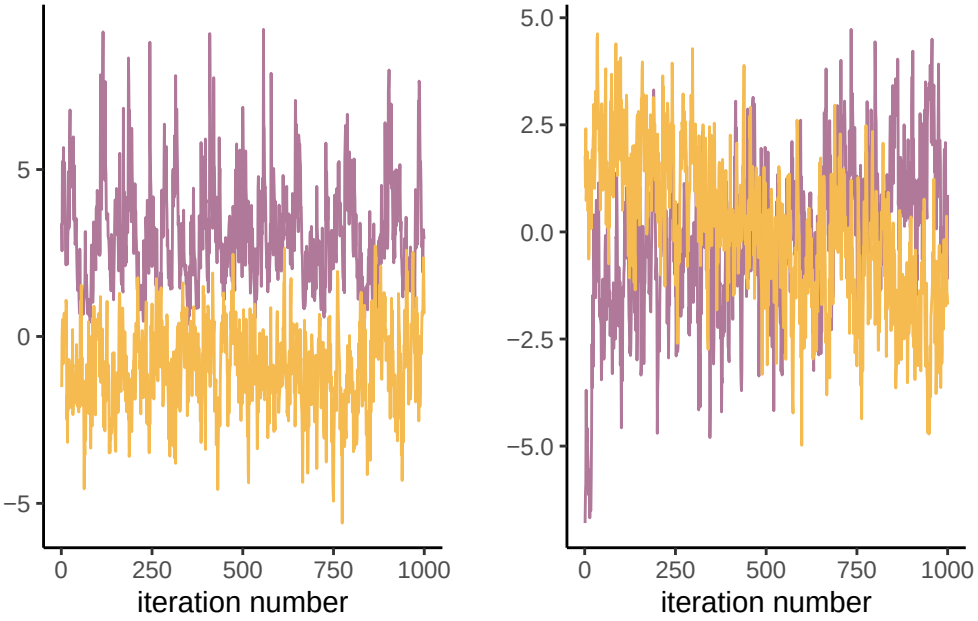

Figure 4.12: Two pairs of Markov chains: the top ones seem stationary, but with different modes. This makes the between chain variance substantial, with a value of $\widehat{R} \approx 3.4$, whereas the chains on the right hover around the same values of zero, but do not appear stable with $\widehat{R} \approx 1.6$.

More generally, it is preferable to run a single chain for a longer period than run multiple chains sequentially, as there is a cost to initializing multiple times with different starting values since we must discard initial draws. With parallel computations, multiple chains are more frequent nowadays.

MCMC algorithms are often run thinning the chain (i.e., keeping only a fraction of the samples drawn, typically every k iteration). This is wasteful as we can of course get more precise estimates by keeping all posterior draws, whether correlated or not. The only argument in favor of thinning is limited storage capacity: if we run very long chains in a model with hundreds of parameters, we may run out of memory.

4.4.3 Posterior predictive checks

Posterior predictive checks can be used to compare models of varying complexity.One of the visual diagnostics, outlined in Gabry et al. (2019), consists in computing a summary statistic of interest from the posterior predictive (whether mean, median, quantile, skewness, etc.) which is relevant for the problem at hand and which we hope our model can adequately capture.

Suppose we have B draws from the posterior and simulate for each n observations from the posterior predictive $p(\widetilde{\boldsymbol{y}} \mid \boldsymbol{y})$: we can benchmark summary statistics from our original data $\boldsymbol{y}$ with the posterior predictive copies $\widetilde{\boldsymbol{y}}_b$. Figure 4.13 shows this for the two competing models and highlight the fact that the simpler model is not dispersed enough. Even the more complex model struggles to capture this additional heterogeneity with the additional variables. One could go back to the drawing board and consider a negative binomial model.

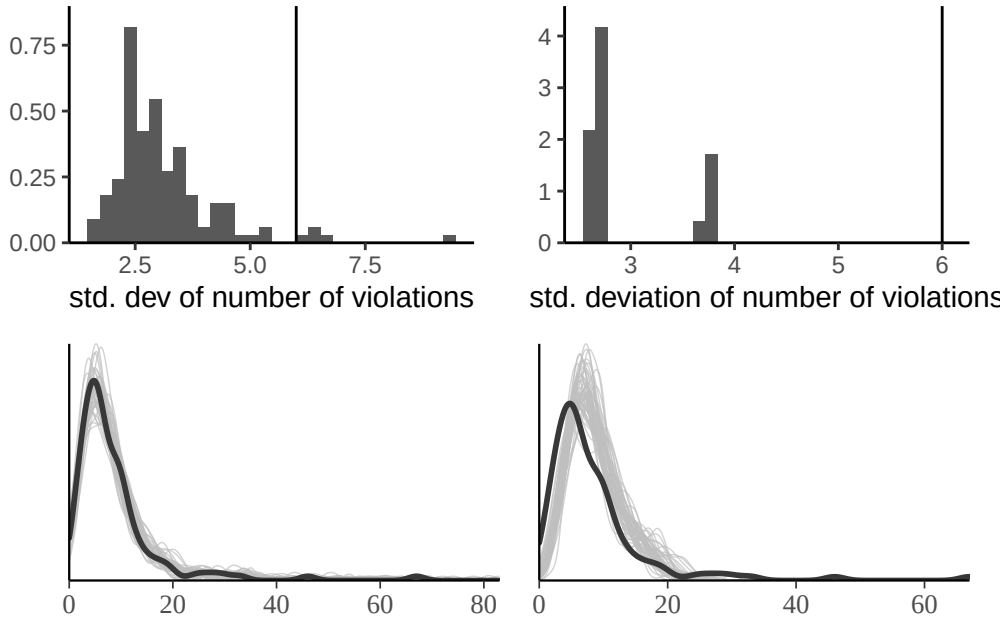

Figure 4.13: Posterior predictive checks for the standard deviation (top) and density of posterior draws (bottom) for hierarchical Poisson model with individual effects (left) and simpler model with only conditions (right).

4.4.4 Information criterion

The widely applicable information criterion (Watanabe 2010) is a measure of predictive performance that approximates the cross-validation loss. Consider first the log pointwise predictive density, defined as the expected value over the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$,

$$\mathsf{LPPD}_i = \mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}} \left\{ \log p(y_i \mid \boldsymbol{\theta}) \right\}.$$

The higher the value of the predictive density LPPD_i, the better the fit for that observation.

As in general information criteria, we sum over all observations, adding a penalization factor that approximates the effective number of parameters in the model, with

$$n\mathsf{WAIC} = -\sum_{i=1}^{n} \mathsf{LPPD}_i + \sum_{i=1}^{n} \mathsf{Va}_{\boldsymbol{\theta}|\boldsymbol{y}} \{ \log p(y_i \mid \boldsymbol{\theta}) \}$$

where we use again the empirical variance to compute the rightmost term. When comparing competing models, we can rely on their values of WAIC to discriminate about the predictive performance. To compute WAIC, we need to store the values of the log density of each observation, or at least minimally compute the running mean and variance accurately pointwise at storage cost $\mathrm{O}(n)$. Note that Section 7.2 of Gelman et al. (2013) define the widely applicable information criterion as $2n \times \mathsf{WAIC}$ to make on par with other information criteria, which are defined typically on the deviance scale and so that lower values correspond to higher predictive performance. For the smartwatch model, we get a value of 3.07 for the complex model and 4.51: this suggests an improvement in using individual-specific effects.

We can also look at the predictive performance. For the `diabetes` data application with the Bayesian LASSO with fixed λ, the predictive performance is a trade-off between the effective number of parameter (with larger penalties translating into smaller number of parameters) and the goodness-of-fit. Figure 4.14 shows that the decrease in predictive performance is severe when estimates are shrunk towards 0, but the model performs equally well for small penalties.

Ideally, one would measure the predictive performance using the leave-one-out predictive distribution for observation i given all the rest, $p(y_i \mid \boldsymbol{y}_{-i})$, to avoid double dipping — the latter is computationally intractable because it would require running n Markov chains with $n-1$ observations each, but we can get a good approximation using importance sampling. The `loo` package uses this with generalized Pareto smoothing to avoid overly large weights.

Once we have the collection of estimated $p(y_i \mid \boldsymbol{y}_{-i})$, we can assess the probability level of each observation. This gives us a set of values which should be approximately uniform if the model was perfectly calibrated. The probability of seeing an outcome as extreme as y_i can

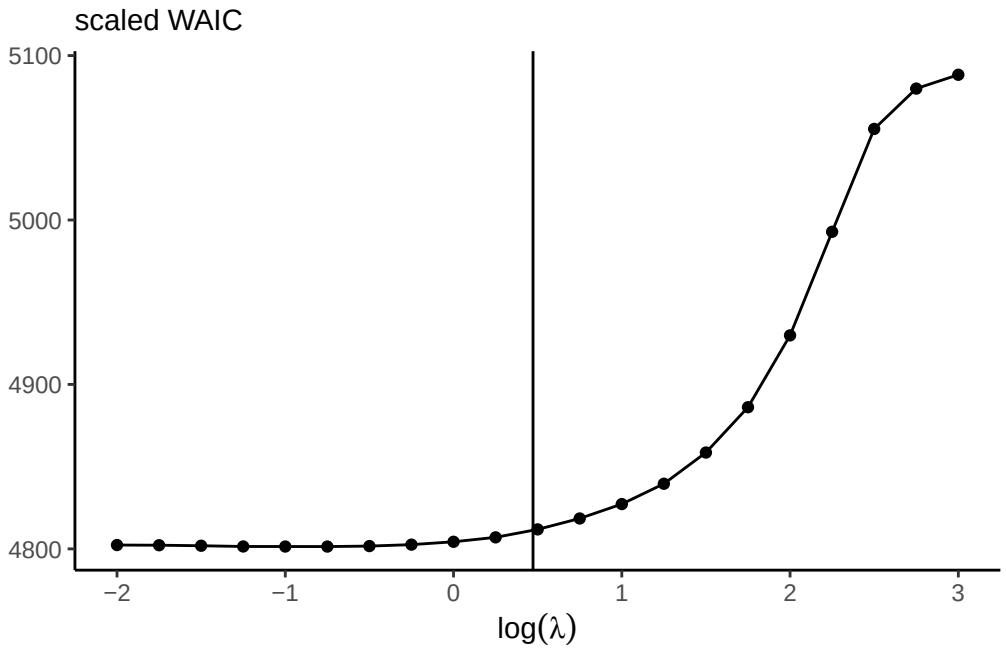

Figure 4.14: Widely applicable information criterion for the Bayesian LASSO problem fitted to the diabetes data, as a function of the penalty λ.

be obtained by simulating draws from the posterior predictive given $\boldsymbol{y}_{-i}$ and computing the scaled rank of the original observation. Values close to zero or one may indicate outliers.

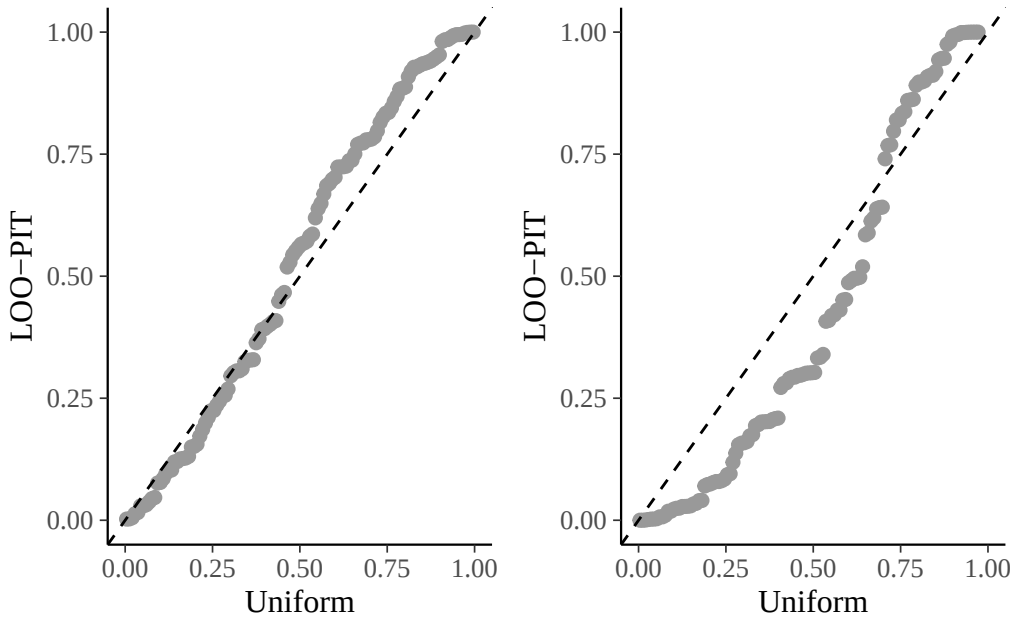

Figure 4.15: Quantile-quantile plots based on leave-one-out cross validation for model for the Poisson hierarchical model with the individual random effects (left) and without (right).

5 References

Albert, Jim. 2009. *Bayesian Computation with R*. 2nd ed. New York: springer. https://doi.org/10.1007/978-0-387-92298-0.

Alexander, Rohan. 2023. *Telling Stories with Data: With Applications in R*. Boca Raton, FL: CRC Press.

Andrieu, Christophe, and Johannes Thoms. 2008. "A Tutorial on Adaptive MCMC." *Statistics and Computing* 18 (4): 343–73. https://doi.org/10.1007/s11222-008-9110-y.

Botev, Zdravko, and Pierre L'Écuyer. 2017. "Simulation from the Normal Distribution Truncated to an Interval in the Tail." In *Proceedings of the 10th EAI International Conference on Performance Evaluation Methodologies and Tools on 10th EAI International Conference on Performance Evaluation Methodologies and Tools*, 23–29. https://doi.org/10.4108/eai.25-10-2016.2266879.

Brodeur, Mathieu, Perrine Ruer, Pierre-Majorique Léger, and Sylvain Sénécal. 2021. "Smartwatches Are More Distracting Than Mobile Phones While Driving: Results from an Experimental Study." *Accident Analysis & Prevention* 149: 105846. https://doi.org/10.1016/j.aap.2020.105846.

Coles, Stuart G., and Jonathan A. Tawn. 1996. "A Bayesian Analysis of Extreme Rainfall Data." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 45 (4): 463–78. https://doi.org/10.2307/2986068.

Devroye, L. 1986. *Non-Uniform Random Variate Generation*. New York: Springer. http://www.nrbook.com/devroye/.

Dyk, David A van, and Xiao-Li Meng. 2001. "The Art of Data Augmentation." *Journal of Computational and Graphical Statistics* 10 (1): 1–50. https://doi.org/10.1198/10618600152418584.

Finetti, Bruno de. 1974. *Theory of Probability: A Critical Introductory Treatment*. Vol. 1. New York: Wiley.

Gabry, Jonah, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. 2019. "Visualization in Bayesian Workflow." *Journal of the Royal Statistical Society Series A: Statistics in Society* 182 (2): 389–402. https://doi.org/10.1111/rssa.12378.

Gelfand, Alan E., and Adrian F. M. Smith. 1990. "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association* 85 (410): 398–409. https://doi.org/10.1080/01621459.1990.10476213.

Gelman, Andrew. 2006. "Prior Distributions for Variance Parameters in Hierarchical Models (Comment on Article by Browne and Draper)." *Bayesian Analysis* 1 (3): 515–34. https:

//doi.org/10.1214/06-BA117A.
Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis*. 3rd ed. New York: Chapman; Hall/CRC. https://doi.org/10.1201/b16018.
Gelman, Andrew, Aki Vehtari, Daniel Simpson, Charles C Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. 2020. "Bayesian Workflow." *arXiv*. https://doi.org/https://doi.org/10.48550/arXiv.2011.01808.
Geman, Stuart, and Donald Geman. 1984. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6 (6): 721–41. https://doi.org/10.1109/TPAMI.1984.4767596.
Geweke, John. 2004. "Getting It Right: Joint Distribution Tests of Posterior Simulators." *Journal of the American Statistical Association* 99 (467): 799–804. https://doi.org/10.1198/016214504000001132.
Geyer, Charles J. 2011. "Introduction to Markov Chain Monte Carlo." In *Handbook of Markov Chain Monte Carlo*, edited by S. Brooks, A. Gelman, G. Jones, and X. L. Meng, 3–48. Boca Raton: CRC Press. https://doi.org/10.1201/b10905.
Gosset, William Sealy. 1908. "The Probable Error of a Mean." *Biometrika* 6 (1): 1–25. https://doi.org/10.1093/biomet/6.1.1.
Gradshteyn, I. S., and I. M. Ryzhik. 2014. *Table of Integrals, Series, and Products*. 8th ed. Academic Press. https://doi.org/10.1016/C2010-0-64839-5.
Green, Peter J. 2001. "A Primer on Markov Chain Monte Carlo." *Monographs on Statistics and Applied Probability* 87: 1–62.
Hastings, W. K. 1970. "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika* 57 (1): 97–109. https://doi.org/10.1093/biomet/57.1.97.
Hobert, James P. 2011. "The Data Augmentation Algorithm: Theory and Methodology." In *Handbook of Markov Chain Monte Carlo*, edited by S. Brooks, A. Gelman, G. Jones, and X. L. Meng, 253–94. Boca Raton: CRC Press. https://doi.org/10.1201/b10905.
Jegerlehner, Sabrina, Franziska Suter-Riniker, Philipp Jent, Pascal Bittel, and Michael Nagler. 2021. "Diagnostic Accuracy of a SARS-CoV-2 Rapid Antigen Test in Real-Life Clinical Settings." *International Journal of Infectious Diseases* 109 (August): 118–22. https://doi.org/10.1016/j.ijid.2021.07.010.
Kinderman, Albert J, and John F Monahan. 1977. "Computer Generation of Random Variables Using the Ratio of Uniform Deviates." *ACM Transactions on Mathematical Software (TOMS)* 3 (3): 257–60.
Marshall, Albert W., and Ingram Olkin. 1985. "A Family of Bivariate Distributions Generated by the Bivariate Bernoulli Distribution." *Journal of the American Statistical Association* 80 (390): 332–38. https://doi.org/10.1080/01621459.1985.10478116.
Mathieu, Edouard, Hannah Ritchie, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Joe Hasell, Bobbie Macdonald, et al. 2020. "Coronavirus Pandemic (COVID-19)." *Our World in Data*.
Matias, J. Nathan, Kevin Munger, Marianne Aubin Le Quere, and Charles Ebersole. 2021.

"The Upworthy Research Archive, a Time Series of 32,487 Experiments in U.S. Media." *Scientific Data* 8 (195). https://doi.org/10.1038/s41597-021-00934-7.

McNeil, A. J., R. Frey, and P. Embrechts. 2005. *Quantitative Risk Management: Concepts, Techniques, and Tools*. 1st ed. Princeton, NJ: Princeton University Press.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. 1953. "Equation of State Calculations by Fast Computing Machines." *The Journal of Chemical Physics* 21 (6): 1087–92. https://doi.org/10.1063/1.1699114.

Nadarajah, Saralees. 2008. "Marshall and Olkin's Distributions." *Acta Applicandae Mathematicae* 103 (1): 87–100. https://doi.org/10.1007/s10440-008-9221-7.

Park, Trevor, and George Casella. 2008. "The Bayesian Lasso." *Journal of the American Statistical Association* 103 (482): 681–86. https://doi.org/10.1198/016214508000000337.

Robert, Christian P., and George Casella. 2004. *Monte Carlo Statistical Methods*. New York, NY: Springer. https://doi.org/10.1007/978-1-4757-4145-2.

Roberts, Gareth O., and Jeffrey S. Rosenthal. 2001. "Optimal Scaling for Various Metropolis–Hastings Algorithms." *Statistical Science* 16 (4): 351–67. https://doi.org/10.1214/ss/1015346320.

Sherlock, Chris. 2013. "Optimal Scaling of the Random Walk Metropolis: General Criteria for the 0.234 Acceptance Rule." *Journal of Applied Probability* 50 (1): 1–15. https://doi.org/10.1239/jap/1363784420.

Simpson, Daniel, Håvard Rue, Andrea Riebler, Thiago G. Martins, and Sigrunn H. Sørbye. 2017. "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors." *Statistical Science* 32 (1): 1–28. https://doi.org/10.1214/16-STS576.

Sørbye, Sigrunn Holbek, and Håvard Rue. 2017. "Penalised Complexity Priors for Stationary Autoregressive Processes." *Journal of Time Series Analysis* 38 (6): 923–35. https://doi.org/10.1111/jtsa.12242.

Tanner, Martin A., and Wing Hung Wong. 1987. "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association* 82 (398): 528–40. https://doi.org/10.1080/01621459.1987.10478458.

Wakefield, J. C., A. E. Gelfand, and A. F. M. Smith. 1991. "Efficient Generation of Random Variates via the Ratio-of-Uniforms Method." *Statistics and Computing* 1 (2): 129–33. https://doi.org/10.1007/BF01889987.

Watanabe, Sumio. 2010. "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory." *Journal of Machine Learning Research* 11 (116): 3571–94. http://jmlr.org/papers/v11/watanabe10a.html.