Bayesian modelling

Léo Belzile

Table of contents

Welcome **1**

1 Introduction **3**
1.1 Random vectors . . . 3
1.1.1 Common distributions . . . 4
1.1.2 Marginal and conditional distributions . . . 11
1.2 Expectations . . . 17
1.3 Likelihood . . . 22

2 Bayesics **29**
2.1 Probability and frequency . . . 30
2.2 Posterior distribution . . . 31
2.3 Posterior predictive distribution . . . 40
2.4 Summarizing posterior distributions . . . 42

3 Priors **51**
3.1 Prior simulation . . . 52
3.2 Conjugate priors . . . 53
3.3 Uninformative priors . . . 59
3.4 Priors for regression models . . . 62
3.5 Informative priors . . . 62
3.6 Sensitivity analysis . . . 68

4 Monte Carlo methods **73**
4.1 Monte Carlo methods . . . 73
4.2 Markov chains . . . 86
4.2.1 Discrete Markov chains . . . 89

5 Metropolis–Hastings algorithm **95**

6 Gibbs sampling **111**
6.1 Data augmentation and auxiliary variables . . . 115

7 Computational strategies and diagnostics **123**
7.1 Convergence diagnostics and model validation 124
7.1.1 Posterior predictive checks 127
7.2 Information criteria 129
7.3 Computational strategies 132

8 Regression models **143**
8.1 Shrinkage priors 150
8.2 Bayesian model averaging via reversible jump 154

9 Deterministic approximations **157**
9.1 Laplace approximation and it's applications 157
9.2 Integrated nested Laplace approximation 163

10 Variational inference **169**
10.1 Model misspecification 169
10.2 Optimization of the evidence lower bound 173
10.2.1 Factorization 173
10.2.2 General derivation 176

11 References **179**

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.

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.

! Learning objectives:

At the end of the chapter, students should be able to

- review notions of joint, marginal and conditional distributions.
- identify the conditional from a joint density.
- calculate analytically the marginal distribution by completing a density function.
- calculate the density after a change of variables.
- calculate of moments, using the definition or using the tower property.
- write down the likelihood for simple settings involving independent observations and other simple scenarios (truncation, censoring, non-identical, serial dependence).

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), and Appendix A of Held and Bové (2020).

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 $\sum_{\boldsymbol{x} \in \mathsf{supp}(\boldsymbol{X})} f_{\boldsymbol{X}}(\boldsymbol{x}) = 1$.

1.1.1 Common distributions

Definition 1.2 (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.

The mean and variance of a gamma are $\mathsf{E}(Y) = \alpha/\beta$ and $\mathsf{Va}(Y) = \alpha/\beta^2$.

Definition 1.3 (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

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

The case $\mathsf{beta}(1, 1)$, also denoted $\mathsf{unif}(0, 1)$, corresponds to a standard uniform distribution. The beta distribution $Y \sim \mathsf{beta}(\alpha, \beta)$ has expectation $\mathsf{E}(Y) = \alpha/(\alpha + \beta)$ and variance $\mathsf{Va}(Y) = \alpha\beta/\{(\alpha + \beta)^2(\alpha + \beta + 1)\}$.

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

Definition 1.4 (Dirichlet distribution)**.** Let $\boldsymbol{\alpha} \in (0, \infty)^d$ denote shape parameters and consider 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 1.5 (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 1.6 (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 1.7 (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 $Y \sim \mathsf{Poisson}(\lambda)$.

Definition 1.8 (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

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

Although the terminology "normal" is frequent, we will stick to Gaussian in these course notes.

Definition 1.9 (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 Student-t univariate distribution with ν degrees of freedom, location μ and scale σ, denoted $\mathsf{Student}(\mu, \sigma, \nu)$, is

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

We also write $\mathsf{Student}_+$ to denote the truncated distribution on the positive half-line, $[0, \infty)$.

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

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

while the corresponding density is

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

The quantile function, the inverse of the distribution function, is $Q(p) = \lambda\{-\log(1-p)\}^{1/\alpha}$. The Weibull distribution includes the exponential as special case when $\alpha = 1$. The expected value of $Y \sim \mathsf{Weibull}(\lambda, \alpha)$ is $\mathsf{E}(Y) = \lambda\Gamma(1 + 1/\alpha)$.

Definition 1.11 (Generalized Pareto distribution)**.** The generalized Pareto distribution with scale $\tau > 0$ and shape $\xi \in \mathbb{R}$ has distribution and density functions equal to, respectively

$$F(x) = \begin{cases} 1 - \left(1 + \frac{\xi}{\tau}x\right)_+^{-1/\xi} & \xi \neq 0 \\ 1 - \exp(-x/\tau) & \xi = 0 \end{cases}, \quad x \geq 0;$$

$$f(x) = \begin{cases} \tau^{-1}\left(1 + \frac{\xi}{\tau}x\right)_+^{-1/\xi-1} & \xi \neq 0 \\ \tau^{-1}\exp(-x/\tau) & \xi = 0 \end{cases} \quad x \geq 0;$$

with $x_+ = \max\{x, 0\}$. The case $\xi = 0$ corresponding to the exponential distribution with rate τ^{-1}. The distribution is used to model excesses over a large threshold u, as extreme value theory dictates that, under broad conditions, $Y - u \mid Y > u \sim \mathsf{gen.Pareto}(\tau_u, \xi)$ as u tends to the endpoint of the support of Y, regardless of the underlying distribution of Y. See Example 2.6 for an application of this model.

Definition 1.12 (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 $Y =_d aX + b$ 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

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

Definition 1.13 (Exponential family)**.** A univariate distribution is an exponential family if it's density or mass function can be written for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $y \in \mathbb{R}$ 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.

If we have an independent and 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.

We term **conjugate family** families of distribution on $\boldsymbol{\Theta}$ with parameters $\boldsymbol{\chi}, \gamma$ if their density is proportional to

$$\exp\left\{\sum_{k=1}^{K} Q_k(\boldsymbol{\theta})\chi_k + \gamma D(\boldsymbol{\theta})\right\}$$

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.

Exponential families play a crucial role due to the fact that the vector of sufficient statistics $\boldsymbol{t}$ for a random sample allows for data compression. They feature prominently in generalized linear models.

Example 1.1 (Gaussian as exponential family)**.** We can rewrite the density of $\mathsf{Gauss}(\mu, \sigma^2)$ as

$$f(y;\mu,\sigma^2) = (2\pi)^{-1/2} \exp\left\{\frac{-y^2 + 2y\mu - \mu^2}{2\sigma^2} - \log\sigma\right\},$$

so taking $Q_1(\mu,\sigma^2) = \mu/\sigma^2$ and $Q_2(\mu,\sigma^2) = 1/\sigma^2$ and $t_1(y) = y$ and $t_2(y) = -y^2/2$. The Gaussian-inverse-gamma distribution is a conjugate family.

Example 1.2 (Binomial as exponential family)**.** 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 $Q_1(p) = \text{logit}(p) = \log\{p/(1-p)\}$ with $t_1(y) = y$. The canonical link function for Bernoulli gives rise to logistic regression model. The binomial distribution is thus an exponential family. The beta distribution is conjugate to the binomial.

Example 1.3 (Poisson as exponential family)**.** Consider $Y \sim \mathsf{Poisson}(\mu)$ with mass function

$$f(y;p) = \exp\left\{-\mu + y\log\mu - \log\Gamma(x+1)\right\}.$$

and so the canonical parameter is $Q_1(p) = \log\mu$ with the gamma distribution as conjugate family.

Proposition 1.1 (Change of variable formula)**.** *Consider an injective (one-to-one) differentiable function $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^d$, with inverse $\boldsymbol{g}^{-1}$. Then, if $\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{X})$,*

$$\Pr(\boldsymbol{Y} \leq \boldsymbol{y}) = \Pr\{\boldsymbol{g}(\boldsymbol{X}) \leq \boldsymbol{y}\} = \Pr\{\boldsymbol{X} \leq \boldsymbol{x} = \boldsymbol{g}^{-1}(\boldsymbol{y})\}.$$

Using the chain rule, we get that the density of $\boldsymbol{Y}$ may be written as

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}\left\{\boldsymbol{g}^{-1}(\boldsymbol{y})\right\}\left|\mathbf{J}_{\boldsymbol{g}^{-1}}(\boldsymbol{y})\right| = f_{\boldsymbol{X}}(\boldsymbol{x})\left|\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{x})\right|^{-1}$$

where $\mathbf{J}_g(\boldsymbol{x})$ is the Jacobian matrix with (i,j)th element $\partial[\boldsymbol{g}(\boldsymbol{x})]_i/\partial x_j$.

Example 1.4 (Location-scale transformation of Gaussian vectors)**.** Consider d independent standard Gaussian variates $X_j \sim \mathsf{Gauss}(0,1)$ for $j = 1,\ldots,d$, with joint density function

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi)^{-d/2}\exp\left(-\frac{\boldsymbol{x}^\top\boldsymbol{x}}{2}\right).$$

Consider the transformation $\boldsymbol{Y} = \mathbf{A}\boldsymbol{X} + \boldsymbol{b}$, with $\mathbf{A}$ an invertible matrix. The inverse transformation is $\boldsymbol{g}^{-1}(\boldsymbol{y}) = \mathbf{A}^{-1}(\boldsymbol{y}-\boldsymbol{b})$. The Jacobian $\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{x})$ is simply $\mathbf{A}$, so the joint density of $\boldsymbol{Y}$ is

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = (2\pi)^{-d/2}|\mathbf{A}|^{-1}\exp\left\{-\frac{(\boldsymbol{y}-\boldsymbol{b})^\top\mathbf{A}^{-\top}\mathbf{A}^{-1}(\boldsymbol{y}-\boldsymbol{b})}{2}\right\}.$$

Since $|\mathbf{A}^{-1}| = |\mathbf{A}|^{-1}$ and $\mathbf{A}^{-\top}\mathbf{A}^{-1} = (\mathbf{A}\mathbf{A}^\top)^{-1}$, we recover that $\boldsymbol{Y} \sim \mathsf{Gauss}_d(\boldsymbol{b}, \mathbf{A}\mathbf{A}^\top)$.

Example 1.5 (Inverse gamma distribution)**.** Consider $Y \sim \mathsf{gamma}(\alpha, \beta)$ and the reciprocal $g(x) = 1/x$. The Jacobian of the transformation is $|g'(y)| = 1/y^2$. The density of the inverse gamma $\mathsf{inv.gamma}(\alpha, \beta)$ with shape $\alpha > 0$ and scale $\beta > 0$ is thus

$$f_Y(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{-\alpha-1} \exp(-\beta/y), \qquad y > 0.$$

The expected value and variance are $\mathsf{E}(Y) = \beta/(\alpha - 1)$ for $\alpha > 1$ and $\mathsf{Va}(Y) = \beta^2/\{(\alpha - 1)^2(\alpha - 2)\}$ for $\alpha > 2$.

Proposition 1.2 (Simulation of Gaussian vectors)**.** *Example 1.4 shows that the Gaussian distribution is 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*

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

Example 1.6 (Dirichlet vectors from Gamma random variables)**.** Consider $\boldsymbol{X}$ a d vector of independent gamma random variables $\mathsf{gamma}(\alpha_i, 1)$. Then, if $Z = X_1 + \cdots + X_d$, we have $(X_1, \ldots, X_{d-1})/Z \sim \mathsf{Dirichlet}(\boldsymbol{\alpha})$ and $Z \sim \mathsf{gamma}(1, \alpha_1 + \cdots + \alpha_d)$.

Proof. The joint density for $\boldsymbol{X}$ is

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{j=1}^{d} \frac{x_j^{\alpha_j - 1} \exp(-x_j)}{\Gamma(\alpha_j)}.$$

Let $\boldsymbol{g}(\cdot)$ be a d place function with ith element $g_i(\boldsymbol{x}) = x_j/(x_1 + \cdots x_d)$ for $j = 1, \ldots, d-1$ and $g_d = x_1 + \cdots x_d$ and write the transformation as $g(\boldsymbol{X}) = (\boldsymbol{Y}^\top, Z)^\top$ with $\boldsymbol{y} = (y_1, \ldots, y_{d-1})^\top$ and the redundant coordinate $y_d = 1 - y_1 - \cdots y_{d-1}$ to simplify the notation. The inverse transformation yields $x_j = z y_j$ for $j = 1, \ldots, d-1$ and $x_d = z(1 - y_1 - \cdots - y_{d-1})$. The Jabobian matrix is

$$\mathbf{J}_{\boldsymbol{g}^{-1}}(\boldsymbol{y}, z) = \begin{pmatrix} z\mathbf{I}_{d-1} & \boldsymbol{y} \\ \mathbf{0}_{d-1}^\top & y_d \end{pmatrix}.$$

The absolute value of the determinant is then $z^{d-1}y_d$. Using the change of variable formula, the joint density is

$$\begin{aligned} f_{\boldsymbol{Y},Z}(\boldsymbol{y}, z) &= \prod_{j=1}^{d} \frac{(zy_j)^{\alpha_j - 1}\exp(-zy_j)}{\Gamma(\alpha_j)} \times z^{d-1}y_d \\ &= z^{\alpha_1+\cdots+\alpha_d-1}\exp(-z)\prod_{j=1}^{d}\frac{y_j^{\alpha_j-1}}{\Gamma(\alpha_j)}. \end{aligned}$$

Since the density factorizes, we find the result upon multiplying and dividing by the normalizing constant $\Gamma(\alpha_1 + \cdots + \alpha_d)$, which yields both the Dirichlet for $\boldsymbol{Y}$ and a gamma for Z. □

1.1.2 Marginal and conditional distributions

Definition 1.14 (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 1.15 (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 $f_{\boldsymbol{X}}$ and $f_{\boldsymbol{Y}}$ denotes the marginal density of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, $f_{\boldsymbol{X}|\boldsymbol{Y}}$ the conditional of $\boldsymbol{X}$ given $\boldsymbol{Y}$ and $f_{\boldsymbol{X},\boldsymbol{Y}}$ the joint density. Bayes' theorem states that for $\boldsymbol{y}$ in the support of $\boldsymbol{Y}$,*

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

which follows since $f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x};\boldsymbol{y})f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})$ and likewise $f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y};\boldsymbol{x})f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})$.

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

$$\begin{aligned}\Pr(X = x \mid Y = y) &= \frac{\Pr(Y = y \mid X = x)\Pr(X = x)}{\Pr(Y = y)} \\ &= \frac{\Pr(Y = y \mid X = x)\Pr(X = x)}{\sum_{x\in\mathcal{X}}\Pr(Y = y \mid X = x)\Pr(X = x)}.\end{aligned}$$

Example 1.7 (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 iteratively in light of new information.

Example 1.8 (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 1.9 (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 1.10 (Gaussian-gamma model)**.** 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. The Student-t can be viewed as a scale mixture of Gaussian, where the variance is inverse gamma distributed.

Example 1.11 (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 1.16 (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 1.3 (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.

1.2 Expectations

The expected value of some function of a random vector $g(\boldsymbol{Y})$, where $\boldsymbol{Y}$ has density $f_{\boldsymbol{Y}}$, is

$$\mathsf{E}\{g(\boldsymbol{Y})\} = \int g(\boldsymbol{y}) f_{\boldsymbol{Y}}(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}$$

and can be understood as a weighted integral of g with weight $f_{\boldsymbol{Y}}$; the latter does not exist unless the integral is finite.

Taking $g(\boldsymbol{y}) = \boldsymbol{y}$ yields the **expected value** of the random variable $\mathsf{E}(\boldsymbol{Y})$. We define the covariance matrix of $\boldsymbol{Y}$ as

$$\mathsf{Va}(\boldsymbol{Y}) = \mathsf{E}\left[\{\boldsymbol{Y} - \mathsf{E}(\boldsymbol{Y})\}\{\boldsymbol{Y} - \mathsf{E}(\boldsymbol{Y})\}^\top\right],$$

which reduces in the unidimensional setting to

$$\mathsf{Va}(Y) = \mathsf{E}\{Y - \mathsf{E}(Y)\}^2 = \mathsf{E}(Y^2) - \mathsf{E}(Y)^2.$$

More generally, the $k \times m$ covariance matrix between two random vectors $\boldsymbol{Y}$ of size k and $\boldsymbol{X}$ of size m is

$$\mathsf{Co}(\boldsymbol{Y}, \boldsymbol{X}) = \mathsf{E}\left[\{\boldsymbol{Y} - \mathsf{E}(\boldsymbol{Y})\}\{\boldsymbol{X} - \mathsf{E}(\boldsymbol{X})\}^\top\right],$$

If $\boldsymbol{Y}$ is d-dimensional and $\mathbf{A}$ is $p \times d$ and $\boldsymbol{b}$ is a p vector, then

$$\begin{aligned}\mathsf{E}(\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{b}) &= \boldsymbol{A}\mathsf{E}(\boldsymbol{Y}) + \boldsymbol{b},\\ \mathsf{Va}(\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{b}) &= \boldsymbol{A}\mathsf{Va}(\boldsymbol{Y})\boldsymbol{A}^\top.\end{aligned}$$

The expected value (theoretical mean) of the vector $\boldsymbol{Y}$ is thus calculated componentwise using each marginal density, i.e.,

$$\mathsf{E}(\boldsymbol{Y}) = \boldsymbol{\mu} = \begin{pmatrix} \mathsf{E}(Y_1) & \cdots & \mathsf{E}(Y_n) \end{pmatrix}^\top$$

whereas the second moment of $\boldsymbol{Y}$ is encoded in the $n \times n$ **covariance** matrix

$$\mathsf{Va}(\boldsymbol{Y}) = \boldsymbol{\Sigma} = \begin{pmatrix} \mathsf{Va}(Y_1) & \mathsf{Co}(Y_1, Y_2) & \cdots & \mathsf{Co}(Y_1, Y_n) \\ \mathsf{Co}(Y_2, Y_1) & \mathsf{Va}(Y_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \mathsf{Co}(Y_n, Y_1) & \mathsf{Co}(Y_n, Y_2) & \cdots & \mathsf{Va}(Y_n) \end{pmatrix}$$

The ith diagonal element of $\boldsymbol{\Sigma}$, $\sigma_{ii} = \sigma_i^2$, is the variance of Y_i, whereas the off-diagonal entries $\sigma_{ij} = \sigma_{ji}$ $(i \neq j)$ are the covariance of pairwise entries, with

$$\begin{aligned} \mathsf{Co}(Y_i, Y_j) &= \int_{\mathbb{R}^2} (y_i - \mu_i)(y_j - \mu_j) f_{Y_i, Y_j}(y_i, y_j) \mathrm{d}y_i \mathrm{d}y_j \\ &= \mathsf{E}_{Y_i, Y_j} \left[\{Y_i - \mathsf{E}_{Y_i}(Y_i)\} \left\{ Y_j - \mathsf{E}_{Y_j}(Y_j) \right\} \right] \end{aligned}$$

The covariance matrix $\boldsymbol{\Sigma}$ is thus symmetric. It is customary to normalize the pairwise dependence so they do not depend on the component variance. The linear **correlation** between Y_i and Y_j is

$$\rho_{ij} = \mathsf{Cor}(Y_i, Y_j) = \frac{\mathsf{Co}(Y_i, Y_j)}{\sqrt{\mathsf{Va}(Y_i)}\sqrt{\mathsf{Va}(Y_j)}} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}.$$

Proposition 1.4 (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 thus necessarily larger than that of the conditional distribution.

Example 1.12. 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 1.13 (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 $\mu = \exp(\beta\mathbf{X})$ 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 1.5 (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* $\dim(\boldsymbol{\Sigma}_{22}) = (d-k) \times (d-k)$. *The following relationships hold:*

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

Proof. By writing explicitly the relationship $\mathbf{Q}\boldsymbol{\Sigma} = \mathbf{I}_n$, we get

$$\begin{aligned}\mathbf{Q}_{11}\boldsymbol{\Sigma}_{11} + \mathbf{Q}_{12}\boldsymbol{\Sigma}_{21} &= \mathbf{I}_k \\ \mathbf{Q}_{21}\boldsymbol{\Sigma}_{12} + \mathbf{Q}_{22}\boldsymbol{\Sigma}_{22} &= \mathbf{I}_{p-k} \\ \mathbf{Q}_{21}\boldsymbol{\Sigma}_{11} + \mathbf{Q}_{22}\boldsymbol{\Sigma}_{21} &= \mathbf{O}_{p-k,k} \\ \mathbf{Q}_{11}\boldsymbol{\Sigma}_{12} + \mathbf{Q}_{12}\boldsymbol{\Sigma}_{22} &= \mathbf{O}_{k,p-k}.\end{aligned}$$

Recall that we can only invert matrices whose double indices are identical and that both $\mathbf{Q}$ and $\boldsymbol{\Sigma}$ are symmetric, so $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^\top$. One easily obtains $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} = -\mathbf{Q}_{11}^{-1}\mathbf{Q}_{12}$ making

use of the last equation. Then, $\mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} = \mathbf{Q}_{11}^{-1}$ by substituting $\mathbf{Q}_{12}$ from the last equation into the first.

For the last result, take $\boldsymbol{B} \coloneqq \left(\begin{smallmatrix} \mathbf{I} & -\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1} \\ \mathbf{O} & \mathbf{I} \end{smallmatrix}\right)$, noting that $\det(\boldsymbol{B}) = \det\left(\boldsymbol{B}^\top\right) = 1$. Computing the quadratic form $\boldsymbol{B}\mathbf{\Sigma}\boldsymbol{B}^\top$, we get $\det(\mathbf{\Sigma}) = \det(\mathbf{\Sigma}_{22})\det(\mathbf{\Sigma}_{1|2})$ where $\mathbf{\Sigma}_{1|2} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$. □

Proposition 1.6 (Conditional distribution of Gaussian vectors)**.** *Let* $\boldsymbol{Y} \sim \mathsf{Gauss}_d(\boldsymbol{\mu}, \mathbf{\Sigma})$ *and consider the partition*

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

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*

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

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

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 $\mathbf{\Sigma}$.

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,

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

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

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

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

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.

□

1.3 Likelihood

Definition 1.17 (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

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

Definition 1.18 (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 1.14 (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 1.15 (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

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

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

Example 1.16 (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

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

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

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

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

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

[1] Using for example a symbolic calculator.

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

```
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 1.17 (First-order autoregressive process)**.** Consider an 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.

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

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

The AR(1) stationary process Y_t, marginally, has mean μ and unconditional variance $\sigma^2/(1-\phi^2)$. If we use the recursive definition, we find

$$Y_t = \mu(1-\phi) + \varepsilon_t + \phi\{\mu + \phi(Y_{t-2} - \mu) + \varepsilon_{t-1}\} = \mu + \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}$$

whence $\mathsf{E}(Y_t) = \mu$ and

$$\mathsf{Va}(Y_t) = \mathsf{Va}\left(\sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}\right) = \sum_{j=0}^{\infty} \phi^{2j} \mathsf{Va}(\varepsilon_{t-j}) = \frac{\sigma^2}{(1-\phi^2)}$$

where the geometric series converges if $\phi < 1$ and diverges otherwise.

If innovations $\{\varepsilon_t\}$ 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 likelihood can then be written as

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

! Summary

- Many families of distributions are closed under conditioning and marginalization (elliptical distribution such as Gaussian or Student, Dirichlet or categorical).
- Location-scale families can be standardized.
- Distributions are characterized by their moments. Statistical models typically specify a parametric form for the latter, notably regression models.
- Hierarchical models specify a joint distribution through marginal and conditionals.
- The law of iterated expectation and variance (tower property) can be used to retrieve the first two moments more easily.
- Families of distribution can be identified from their kernel (density or mass function) and support, up to normalizing constants. It suffices to identify the parameters of the distribution to retrieve the model.
- Marginalization from a joint density involves integration. We can often rewrite the integral in terms of an unnormalized density, and use knowledge of the marginalization constant.
- The likelihood depends on the distribution of the data. It encodes the information about the data.
- The score (gradient vector of the log likelihood) can be set to zero to find the maximum likelihood estimator in regular models.
- The curvature of the log likelihood (Hessian matrix) gives an idea of the information. The observed information is the sample version, the expected or Fisher

information obtained by replacing data observations with their expectation.
- In independent samples, the information grows linearly with the sample size and is $\mathrm{O}(n)$. For identically distributed samples, we can compute the Fisher information for a single observation.

2 Bayesics

The Bayesian paradigm is an inferential framework that is widely used in data science. It 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.

! Learning objectives:

At the end of the chapter, students should be able to

- define key notions (prior, marginal likelihood, Bayes factor, credible intervals, etc.) of Bayesian inference.
- distinguish between questions that relate to the posterior distribution versus the posterior predictive.
- calculate and compute numerically summaries of posterior distributions (point estimators and credible intervals) given a posterior sample.
- explain some of the conceptual differences between frequentist and Bayesian inference.

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.) From Definition 1.3, 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 $\boldsymbol{\Theta}$.

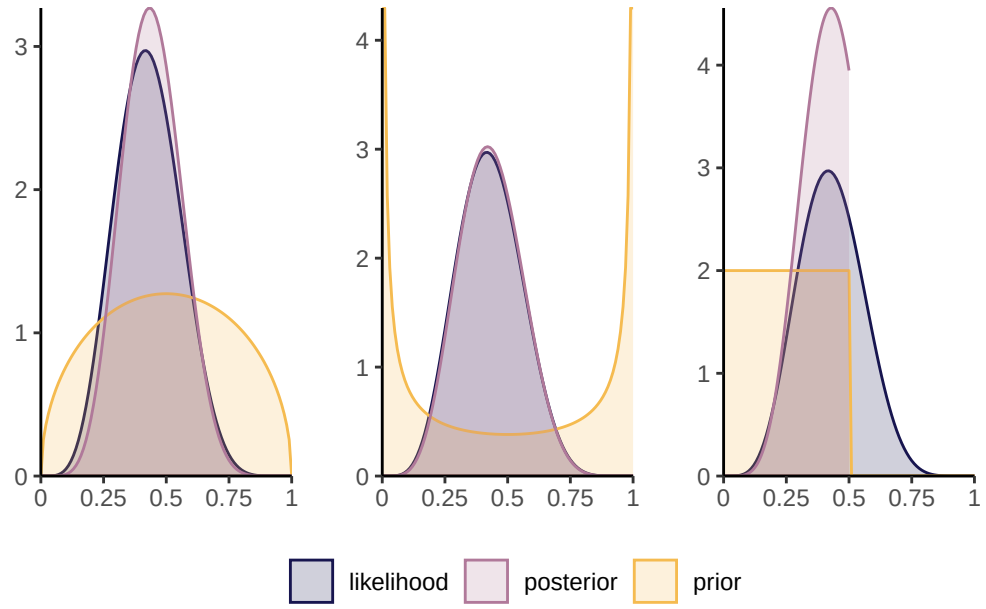

Figure 2.1: Scaled binomial likelihood for six successes out of 14 trials, with $\texttt{beta}(3/2, 3/2)$ prior (left), $\texttt{beta}(1/4, 1/4)$ (middle) and truncated uniform on $[0, 1/2]$ (right), with the corresponding posterior distributions.

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

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

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

Definition 2.1 (Bayes factor and model comparison)**.** The marginal likelihood enters in the comparison of different models. Suppose that we have models $\mathcal{M}_m$ $(m = 1, \ldots, M)$ to be compared, with parameter vectors $\boldsymbol{\theta}^{(m)}$ and data vector $\boldsymbol{y}$. Consider $p_m = \Pr(\mathcal{M}_m)$ the prior probability of the different models under consideration, with $p_1 + \cdots + p_M = 1$. The posterior odds for Models $\mathcal{M}_i$ vs $\mathcal{M}_j$ are

$$\underbrace{\frac{\Pr(\mathcal{M}_i \mid \boldsymbol{y})}{\Pr(\mathcal{M}_j \mid \boldsymbol{y})}}_{\text{posterior odds}} = \underbrace{\frac{p(\boldsymbol{y} \mid \mathcal{M}_i)}{p(\boldsymbol{y} \mid \mathcal{M}_j)}}_{\text{Bayes factor}} \underbrace{\frac{\Pr(\mathcal{M}_i)}{\Pr(\mathcal{M}_j)}}_{\text{prior odds}}$$

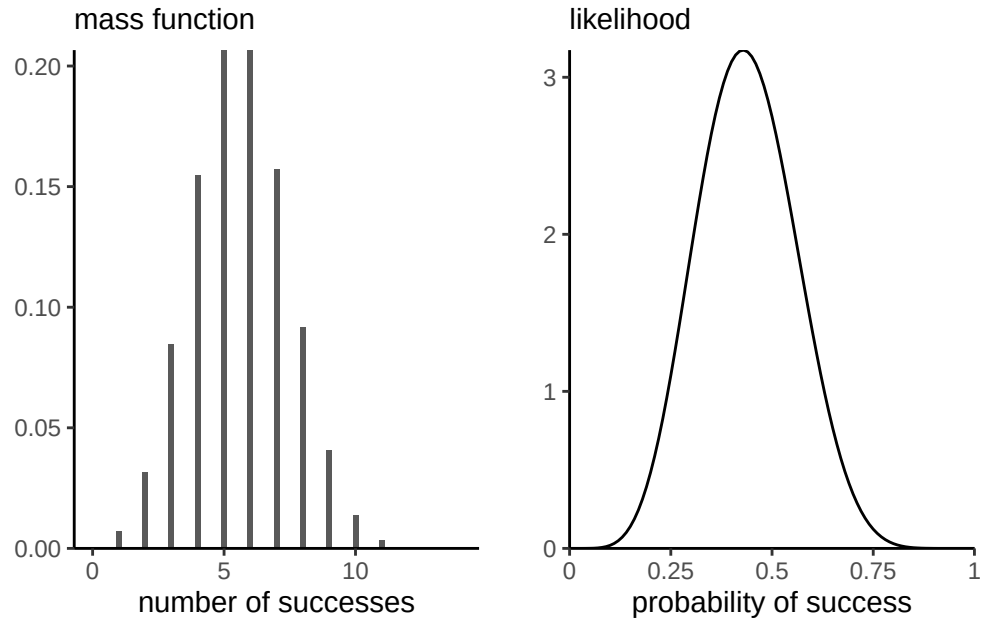

Figure 2.2: Binomial mass function (left) and scaled likelihood function (right).

where the first term on the right hand side is the Bayes factor for model i vs j, denoted BF_{ij}. The Bayes factor is the ratio of marginal likelihoods, as

$$p(\boldsymbol{y} \mid \mathcal{M}_i) = \int p(y \mid \boldsymbol{\theta}^{(i)}, \mathcal{M}_i) p(\boldsymbol{\theta}^{(i)} \mid \mathcal{M}_i) \mathrm{d}\boldsymbol{\theta}^{(i)}.$$

Values of $\mathsf{BF}_{ij} > 1$ correspond to model $\mathcal{M}_i$ being more likely than $\mathcal{M}_j$.

While Bayes factors are used for model comparison, the answers depend very strongly on the prior $p(\boldsymbol{\theta}^{(i)} \mid \mathcal{M}_i)$ specified and the latter must be proper as a general rule for the ratio to be well-defined.

The Bayes factor require that we compare the same data, but both likelihood and priors could be different from one model to the next.

Example 2.2 (Bayes factor for the binomial model)**.** The marginal likelihood for the $Y \mid P = p \sim \mathsf{binom}(n, p)$ model with prior $P \sim \mathsf{beta}(\alpha, \beta)$ is

$$p_Y(y) = \binom{n}{y} \frac{\mathrm{beta}(\alpha + y, \beta + n - y)}{\mathrm{beta}(\alpha, \beta)}.$$

where $\text{beta}(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ is the beta function, expressed in terms of gamma functions.

Consider three models with $Y \mid P^{(i)} = p, \mathcal{M}_i \sim \text{binom}(n, p)$ for $i = 1, 2, 3$ and uniform, point mass and beta priors $P^{(1)} \sim \text{unif}(0, 1)$, $P^{(2)} \sim \text{beta}(3/2, 3/2)$ and $P^{(3)} \sim 1_{p=0.5}$. For $\mathcal{M}_3$, the marginal likelihood is simply equal to the binomial distribution with $p = 0.5$.

If $n = 14$, but we let instead the number of success varies, the models that put more mass closer to the ratio y/n will be favored. The uniform prior in model $\mathcal{M}_1$ will have a higher Bayes factor than model $\mathcal{M}_2$ or $\mathcal{M}_3$ for values closer to $p = 0$ or $p = 1$, but there is mild evidence as shown in Figure 2.3.

```
# Log of marginal posterior for binom with beta prior (default is uniform)
log_marg_post_beta <- function(n, y, alpha = 1, beta = 1){
  lchoose(n, y) + lbeta(alpha + y, beta + n - y) - lbeta(alpha, beta)
}
# Log of Bayes factor
logBF2vs3 <- function(y, n){ # model 2 (beta(1.5,1.5) vs 3 (point mass at 0.5)
  log_marg_post_beta(n = n, y = y, alpha = 1.5, beta = 1.5) - dbinom(x = y, size = n, prob =
}
```

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* $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{k-1}$.

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*

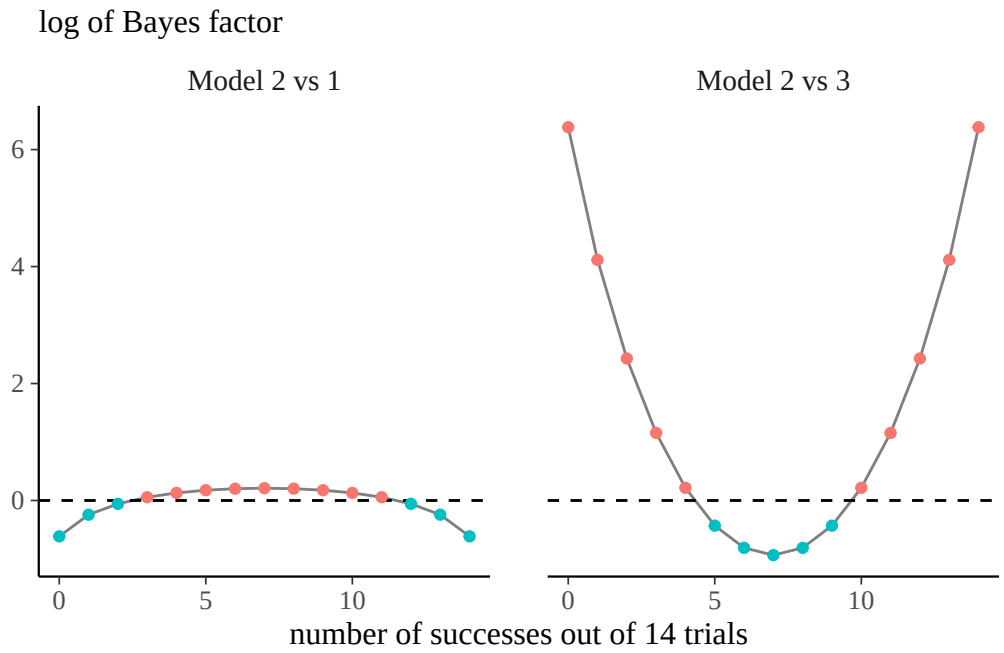

Figure 2.3: Log of Bayes factors for comparison of binomial models with $n = 14$ trials as a function of the number of successes n. Values larger than zero (on log scale) indicate preference for Model 2.

the full posterior. Figure 2.4 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 2.3 (Numerical integration)**.** 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 (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.

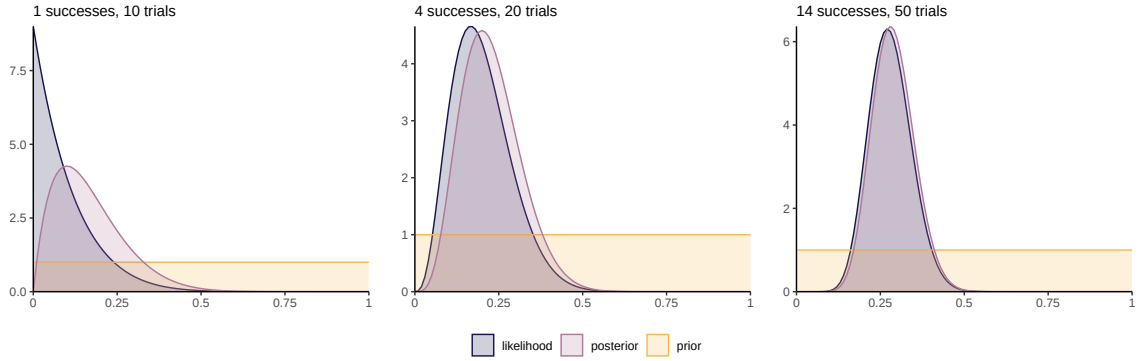

Figure 2.4: Beta posterior and binomial likelihood with a uniform prior for increasing number of observations (from left to right) out of a total of 100 trials.

```
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.

Example 2.4 (Importance of selling format)**.** Duke and Amir (2023) consider the difference between integrated and sequential format for sales. The `sellingformat` dataset contains $n = 397$ observations split into two groups: quantity-integrated decision (decide the amount to buy) and quantity-sequential (first select buy, then select the amount). Participants of the study were randomly allocated to either of these two format and their decision, either buy, `1`, or do not buy `0`, is recorded.

Table 2.1: Aggregated data from Duke and Amir (2023), experiment 1. Number of participants who did not (`0`) or did buy (`1`) products as a function of experimental condition.

	0	1
quantity-integrated	152	46
quantity-sequential	176	23

We consider the number of purchased out of the total, treating records as independent Bernoulli observations with a flat (uniform prior).

With a beta-binomial model, the posterior for the probability of buying is $\texttt{beta}(47, 153)$ for quantity-integrated and $\texttt{beta}(24, 177)$ for quantity-sequential. We can compute the posterior of the odds ratio,

$$O = \frac{\Pr(Y = 1 \mid \texttt{integrated})}{\Pr(Y = 0 \mid \texttt{integrated})} \frac{\Pr(Y = 0 \mid \texttt{sequential})}{\Pr(Y = 1 \mid \texttt{sequential})},$$

by simulating independent draws from the posteriors of each condition and computing the odds ratio.

```
data(sellingformat, package = "hecbayes")
contingency <- with(sellingformat, table(format, purchased))
# Posterior draws of the parameters
post_p_int <- rbeta(n = 1e4, shape1 = 47, shape2 = 153)
post_p_seq <- rbeta(n = 1e4, shape1 = 24, shape2 = 177)
# Reparametrization
post_odds_int <- (post_p_int / (1 - post_p_int))
post_odds_seq <- (post_p_seq / (1 - post_p_seq))
post_oddsratio <- post_odds_int / post_odds_seq
```

Figure 2.5 shows the posterior of the probability of buying for each group, and the odds. It is clear that the integrated format leads to much more sales in the experiment, with a posterior ratio exceeding 1 with probability 99.89%.

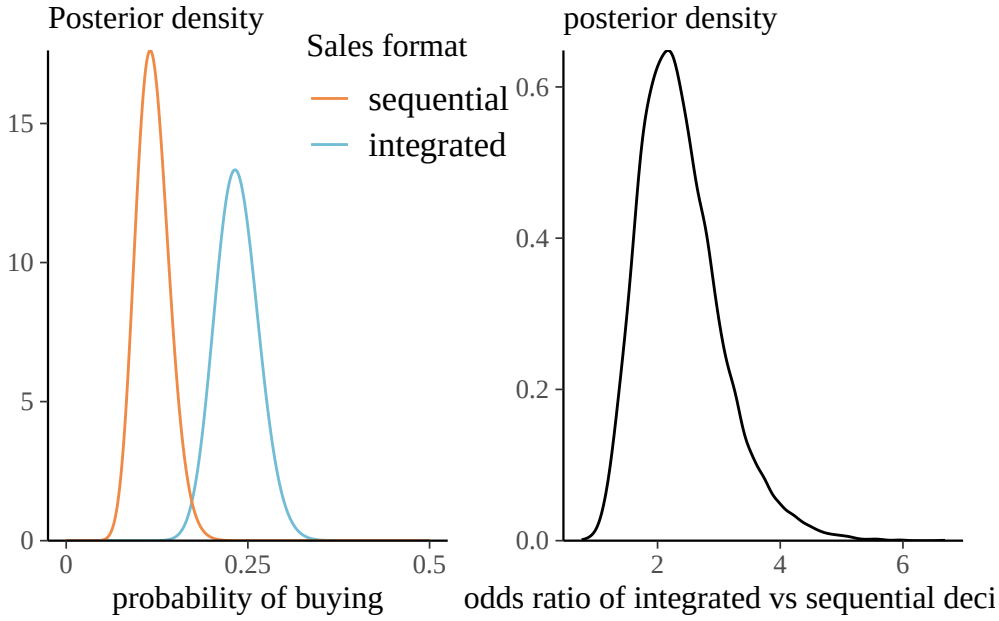

Figure 2.5: Posterior curves per group (left) and odds ratio (right)

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 2.6, 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)
```

Given the $\mathsf{Be}(a, b)$ posterior with $a = y + \alpha$ and $b = n - y + \beta$, the predictive distribution of Y_{new} for fixed n_{new} number of trials is beta-binomial with mass function

$$\begin{aligned} p(y_{\text{new}} \mid y) &= \int_0^1 \binom{n_{\text{new}}}{y_{\text{new}}} \frac{\theta^{a+y_{\text{new}}-1}(1-\theta)^{b+n_{\text{new}}-y_{\text{new}}-1}}{\mathrm{Be}(a,b)} \mathrm{d}\theta \\ &= \binom{n_{\text{new}}}{y_{\text{new}}} \frac{\mathrm{Be}(a + y_{\text{new}}, b + n_{\text{new}} - y_{\text{new}})}{\mathrm{Be}(a,b)} \end{aligned}$$

Example 2.5 (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

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

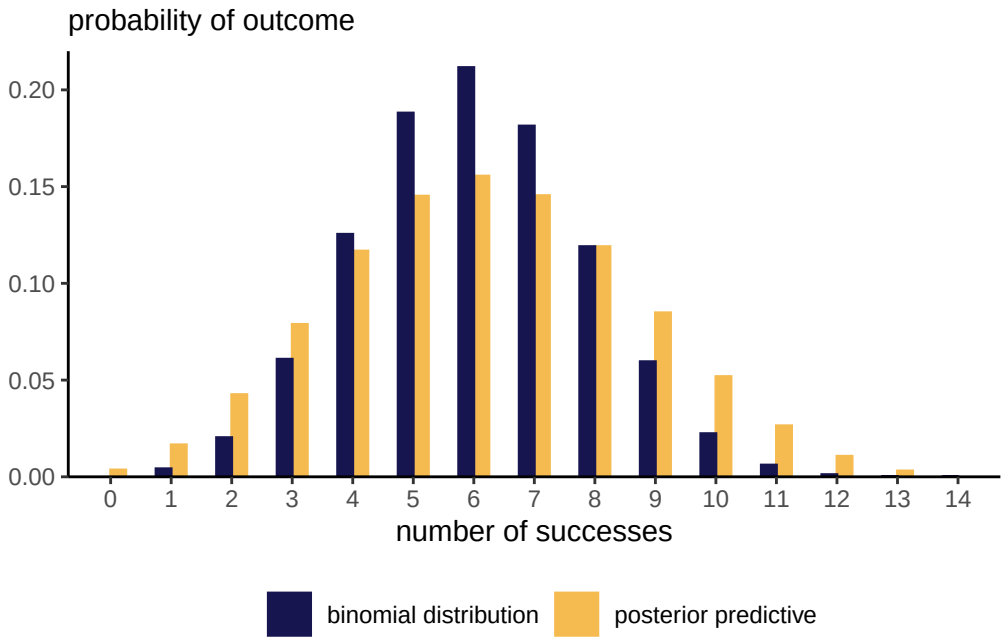

Figure 2.6: Beta-binomial posterior predictive distribution with corresponding binomial mass function evaluated at the maximum likelihood estimator.

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 2.2 (Loss function)**.** A loss function $c(\boldsymbol{\theta}, \boldsymbol{\upsilon})$ is a mapping from $\mathbb{R}^p \to \mathbb{R}^k$ 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,

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

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.

For example consider the quadratic loss function which is differentiable. Provided we can interchange differential operator and integral sign,

$$\begin{aligned}0 &= \int_{\mathbb{R}} \frac{\partial (\upsilon - \theta)^2}{\partial \upsilon} p(\theta \mid \boldsymbol{y}) \mathrm{d}\theta \\ &= \int_{\mathbb{R}} \frac{\partial 2(\upsilon - \theta)}{\partial \upsilon} p(\theta \mid \boldsymbol{y}) \mathrm{d}\theta \\ &= 2\upsilon - 2\mathsf{E}(\theta)\end{aligned}$$

which is minimized when $\widehat{\upsilon} = \mathsf{E}_{\Theta|\boldsymbol{Y}}(\theta)$.

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 2.7. 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.

Example 2.6 (Value-at-risk for 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

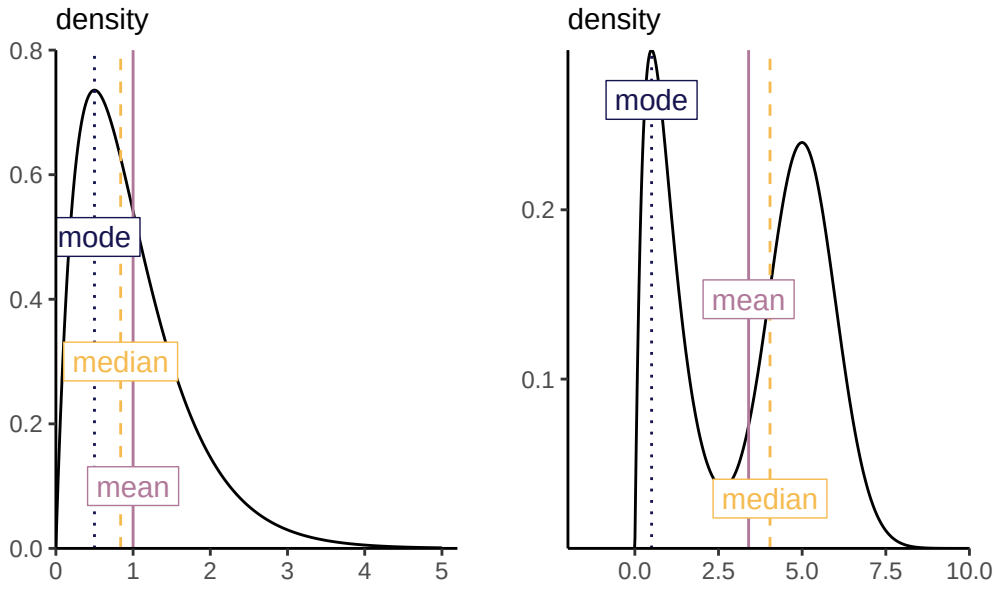

Figure 2.7: Point estimators from a right-skewed distribution (left) and from a multimodal distribution (right).

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 $Z \sim \mathsf{gen.Pareto}(\tau, \xi)$; see Definition 1.11

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.

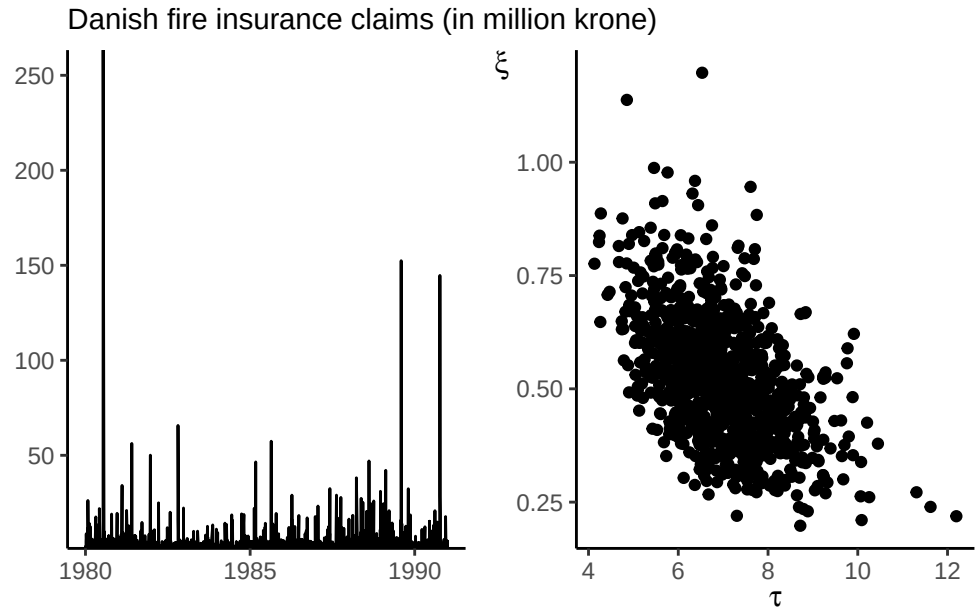

Figure 2.8: Time series of Danish fire claims exceeding a million krone (left) and posterior samples from the scale τ and shape ξ of the generalized Pareto model fitted to exceedances above 10 million krone (right).

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. The log likelihood for the full model for $y_i > u$ is

$$\ell(\tau, \xi, \zeta_u) \propto -109 \log \tau + \sum_{i=1}^{109} (1 + 1/\xi) \log \left(1 + \xi \frac{y_i - 10}{\tau}\right)_+ + 109 \log \zeta_u + 2058 \log(1 - \zeta_u),$$

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.

We consider the modelling of insurance losses exceeding $u = 10$ millions krones using a generalized Pareto distribution to the `danish` fire insurance data with some prior; see Definition 1.11 for the model. The model has three parameters: the scale τ, the shape ξ and the probability of exceeding the threshold ζ_u.

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 can be written in terms of the generalized Pareto survival function times the probability of exceedance above the threshold,

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

and solving for q_α gives

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

We obtained, using tools that will be discussed in Example 4.3, a matrix `post_samp` that contains exact samples from the posterior distribution of (τ, ξ, ζ_u). To obtain the posterior distribution of the α quantile, q_α, it thus 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 2.9 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 2.9, 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])
}
```

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

Definition 2.3. 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 2.10 shows two extreme cases in which these intervals differ: 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.

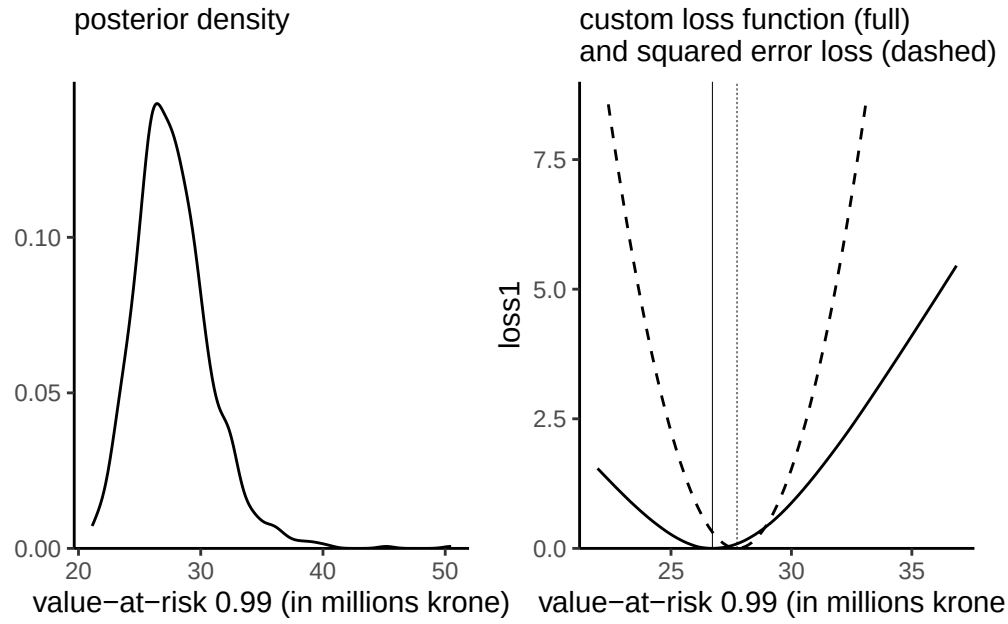

Figure 2.9: Posterior density (left) and losses functions for the 0.99 value-at-risk for the Danish fire insurance data. The vertical lines denote point estimates of the quantiles that minimize the loss functions.

```
set.seed(2023)
postsamp <- rbeta(n = 1000, shape1 = 0.5, shape2 = 0.2)
alpha <- 0.11
# Compute equitailed interval bounds
quantile(postsamp, probs = c(alpha/2, 1-alpha/2))
```

```
     5.5%     94.5%
0.0246807 0.9999980
```

```
qbeta(p = c(alpha/2, 1-alpha/2), shape1 = 0.5, shape2 = 0.2)
```

```
[1] 0.02925205 0.99999844
```

```
# Highest posterior density intervals
hdiD <- HDInterval::hdi(density(postsamp), credMass = 1-alpha, allowSplit = TRUE)
```

The equitailed intervals for a known posterior can be obtained directly from the quantile function or via Monte Carlo simply by querying sample quantiles. The HPD region is more complicated to obtain and requires dedicated software, which in the above case may fail to account for the support of the posterior!

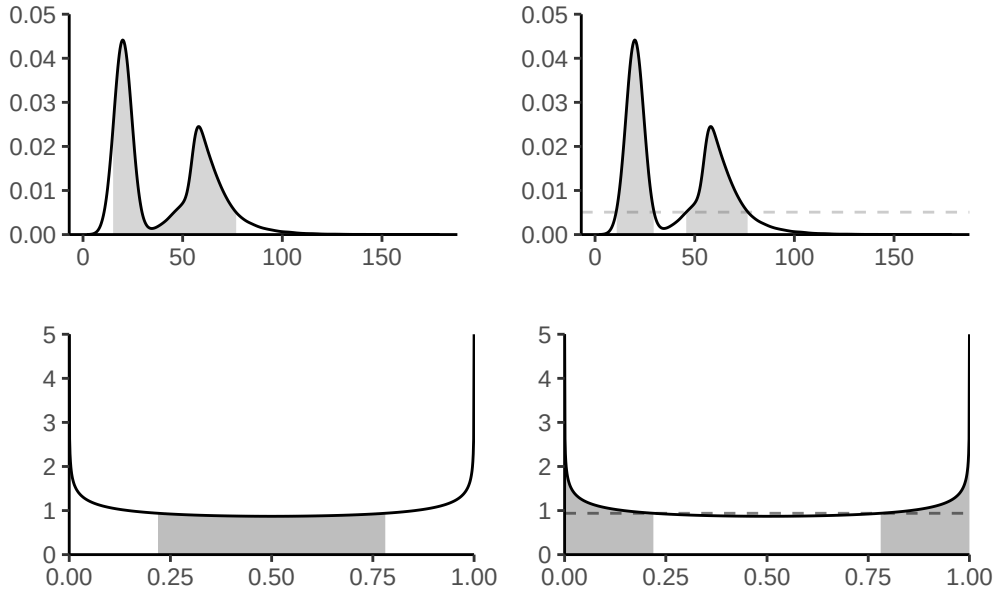

Figure 2.10: Density plots with 89% (top) and 50% (bottom) equitailed or central credible (left) and highest posterior density (right) regions for two data sets, highlighted in grey. The horizontal lign gives the posterior density value determining the cutoff for the HDP region.

! Summary:

- Bayesians treat both parameter and observations as random, the former due to uncertainty about the true value. Inference is performed conditional on observed data, and summarized in the posterior.
- Bayesians specify both a prior for the parameter, and a likelihood specifying

the data generating mechanism. It is thus an extension of likelihood-based inference.
- Information from the prior is updated in light of new data, which is encoded by the likelihood. Sequential updating leads.
- Under weak conditions on the prior, large-sample behaviour of Bayesian and frequentist.
- Bayesian inference is complicated by the fact that there is more often than not no closed-form expression for the posterior distribution. Evaluation of the normalizing constant, the so-called marginal likelihood, is challenging, especially in high dimensional settings.
- Rather than hypothesis testing, Bayesian methods rely on the posterior distribution of parameters, or on Bayes factor for model comparisons.
- The posterior predictive distribution, used for model assessment and prediction, and it has a higher variance than the data generating distribution from the likelihood, due to parameter uncertainty.
- Bayesians typically will have approximations to the posterior distribution, or samples drawn from it.
- Loss functions can be used to summarize a posterior distribution into a numerical summary of interest, which may vary depending on the objective.
- Uncertainty is reflected by credible sets or credible intervals, which encode the posterior probability that the true value θ belongs to the set.

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.

Oftentimes, we will specify independent priors in multiparameter models, but the posterior of these will not be independent.

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.

! Learning objectives:

At the end of the chapter, students should be able to

- define and distinguish between improper, weak and informative priors.
- propose conjugate priors for exponential families.
- assess by using the prior predictive distribution the compatibility of the prior with the model.
- use moment matching to specify the parameters of a prior distribution.

- perform sensitivity analysis by running a model with different priors and assessing changes to the posterior distribution.

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 $\int_{\boldsymbol{\Theta}} f(y_{\text{new}}; \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$: we can simulate outcomes from it by first drawing parameter values $\boldsymbol{\theta}_0$ from the prior, then sampling new observations from the distribution $f(y_{\text{new}}; \boldsymbol{\theta}_0)$ with those parameters values and keeping only y_{new}. If there are sensible bounds for the range of the response, we could restrict the prior range and shape until values abide to 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

[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.

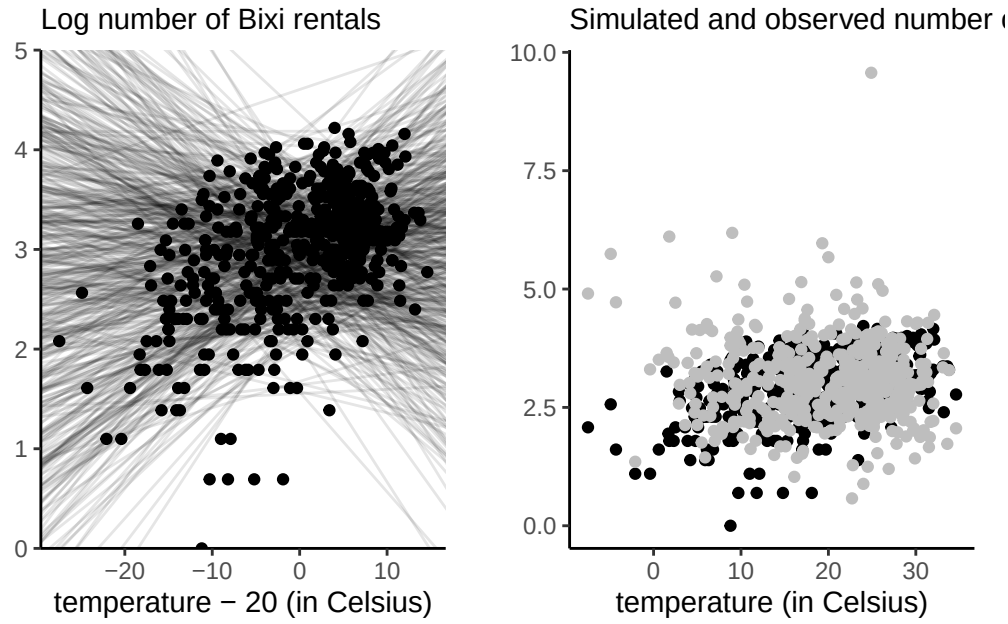

Figure 3.1: Prior draws of the linear regressions with observed data superimposed (left), and draws of observations from the prior predictive distribution (in gray) against observed data (right).

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 priors are chosen for computational convenience and because interpretation is conve-

nient, as the parameters of the posterior will often be some weighted average of prior and likelihood component.

Definition 3.1 (Conjugate priors)**.** 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 (see Definition 1.13, including the binomial, Poisson, exponential, Gaussian distributions) admit conjugate priors.

distribution	unknown parameter	conjugate prior
$Y \sim \mathsf{expo}(\lambda)$	λ	$\lambda \sim \mathsf{gamma}(\alpha, \beta)$
$Y \sim \mathsf{Poisson}(\mu)$	μ	$\mu \sim \mathsf{gamma}(\alpha, \beta)$
$Y \sim \mathsf{binom}(n, \theta)$	θ	$\theta \sim \mathsf{Be}(\alpha, \beta)$
$Y \sim \mathsf{Gauss}(\mu, \sigma^2)$	μ	$\mu \sim \mathsf{Gauss}(\nu, \omega^2)$
$Y \sim \mathsf{Gauss}(\mu, \sigma^2)$	σ	$\sigma^2 \sim \mathsf{inv.gamma}(\alpha, \beta)$
$Y \sim \mathsf{Gauss}(\mu, \sigma^2)$	μ, σ	$\mu \mid \sigma^2 \sim \mathsf{Gauss}(\nu, \omega\sigma^2)$, $\sigma^2 \sim \mathsf{inv.gamma}(\alpha, \beta)$

Example 3.2 (Conjugate prior for the binomial model)**.** Since the density of the binomial is of the form $p^y(1-p)^{n-y}$ and it belongs to an exponential family (Example 1.2), 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)**.** We saw in Example 1.3 that the Poisson distribution is an exponential family. 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 1.13 demonstrated.

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 $\beta = 104.44$ for the hyperparameters, setting $\alpha/\beta = 0.0158$ and $\alpha/\beta^2 = 0.0123^2$ and solving for the two unknown parameters.

Table 3.2: 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

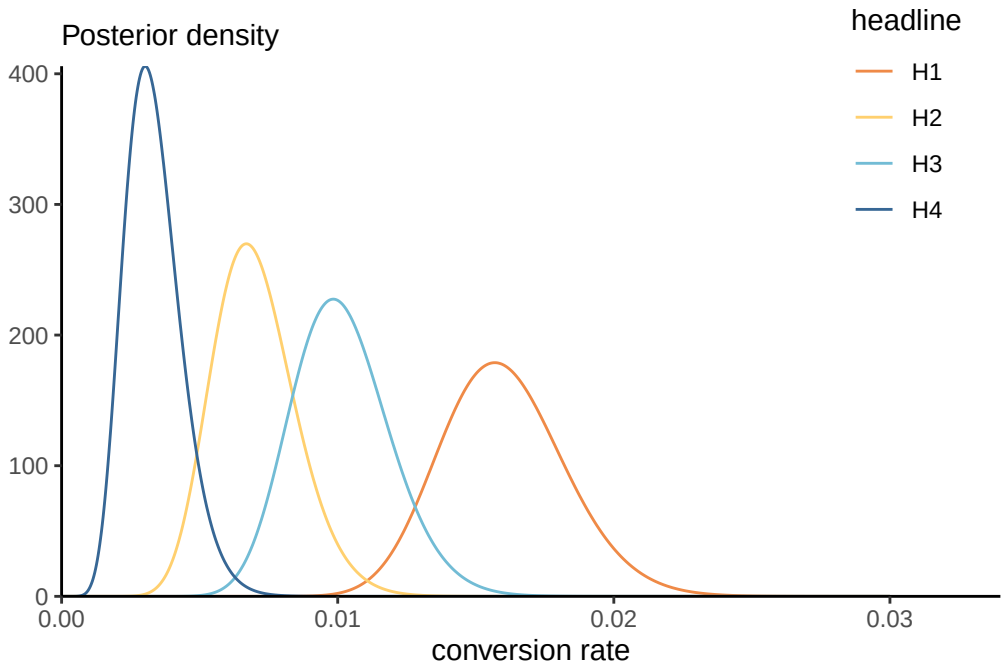

Figure 3.2: Gamma posterior for conversion rate of the different Upworthy Sesame Street headline.

We can visualize the posterior distributions. In this context, the large sample size lead to the dominance of the likelihood contribution $p(Y_i \mid \lambda_i) \sim \text{Poisson}(n_i\lambda_i)$ relative to the prior. We can see there is virtually no overlap between different rates for headers H1 (preferred) 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 2% 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.[3]

[3]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.

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 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 $\beta = 25 = \mathsf{E}_0(\lambda_j)/\mathsf{Var}_0(\lambda_j)$ and a shape of $\alpha = 0.25$ $(j=1,2)$. 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.

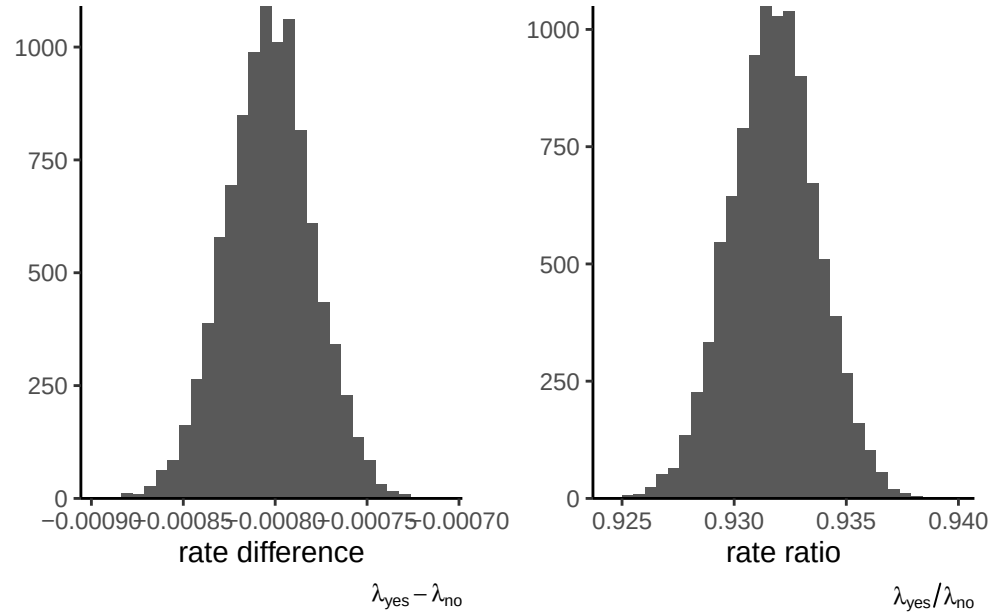

Figure 3.3: Histograms of posterior summaries for differences (left) and rates (right) based on 1000 simulations from the independent gamma posteriors.

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 ν, τ are fixed hyperparameter values. For now, we consider only inference for the conditional marginal posterior $p(\mu \mid \boldsymbol{y}, \sigma)$: discarding any term that is not a function of μ, the conditional posterior is

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

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. A prior over $\boldsymbol{\Theta}$ is **improper** if $\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \infty$.

The best example of proper 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.[4]

[4]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.

To see this, consider a bijective transformation $\theta \mapsto \vartheta$. Under the reparametrized model and suitable regularity conditions[5], the chain rule implies that

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

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 $f(y; \theta) \propto \theta^y(1-\theta)^{1-y}\mathbf{1}_{\theta \in [0,1]}$. 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 $\lambda^{-1/2}$.

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

3.4 Priors for regression models

Regression models often feature Gaussian priors on the mean coefficients $\boldsymbol{\beta}$, typically chosen to be vague with large variance. Below are some alternatives, many of which aim to enforce shrinkage towards zero, or sparsity.

Proposition 3.1 (Zellner's g prior)**.** *Consider an ordinary linear regression model for* $\boldsymbol{Y} \sim \mathsf{Gauss}_n(\beta_0\mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$*, with intercept* β_0 *and mean coefficient vector* $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ *associated to the model matrix* $\mathbf{X}$*. Zellner (1986)'s* g *prior consists in letting* $\boldsymbol{\beta} \sim \mathsf{Gauss}_p\{\mathbf{0}_p, g\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\}$*, where* $g > 0$ *is a constant.*

The ordinary least square estimator of the mean coefficients satisfies under regularity conditions on the model matrix $\widehat{\boldsymbol{\beta}} \sim \mathsf{Gauss}_p\{\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\}$ *for Gaussian data, whence we get the closed-form conditional distributions*

$$\begin{aligned}\beta_0 \mid \sigma^2, \boldsymbol{Y} &\sim \mathsf{Gauss}(\overline{y}, \sigma^2/n)\\ \boldsymbol{\beta} \mid \beta_0, \sigma^2, \boldsymbol{Y} &\sim \mathsf{Gauss}_p\left\{\frac{g}{g+1}\widehat{\boldsymbol{\beta}}, \frac{g}{g+1}\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\right\}\end{aligned}$$

where $\overline{y} = \boldsymbol{y}^\top\mathbf{1}_n/n$ *is the sample mean of the observed response vector. We can interpret* $g > 0$ *as a prior weight, with the posterior conditional mean giving weight of* n/g *to "phantom (prior) observations" with mean zero, relative to the* n *observations in the observed sample: the ratio* $g/(g+1)$ *is called* ***shrinkage factor****.*

By virtue of Proposition 1.6, the prior is also closed under conditioning, which is useful for model comparison using Bayes factors. Consider a partition $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$ *of the mean coefficients and similarly the block of columns from the model matrix, say* $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2]$ *for blocks of size* k *and* $p-k$*. If we remove* $p-k$ *regressors from the model setting* $\boldsymbol{\beta}_2 = 0$*, then the conditional is*

$$\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2 = \mathbf{0}_{p-k} \sim \mathsf{Gauss}_k\{\mathbf{0}_k, g\sigma^2(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\},$$

which is the g *prior for the submodel in which we omit the columns corresponding to* $\mathbf{X}_2$*.*

3.5 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 elicited 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 (Priors in extreme value theory)**.** The generalized extreme value distribution obtained as the limit of maximum of blocks of size m when suitably normalizes is a location-scale family with a shape parameter $\xi \in \mathbb{R}$. The latter describes the heavyness of the tail, with negative values corresponding to approximation by bounded upper tail distributions (such as the beta), $\xi = 0$ to exponential tail decay and $\xi > 0$ to polynomial tails, with finite moments of order $1/\xi$. For example, the Cauchy or Student-t distribution with one degree of freedom has infinite first moment and $\xi = 1$.

In practice, the maximum likelihood estimators do not exist if $\xi < -1$ as the model is nonregular (Smith 1985), and the cumulant of order k exists only if $\xi > -1/k$; the Fisher information matrix exists only when $\xi > -1/2$. Thus, informative priors that restrict the range of the shape, may be useful as in environmental applications the shapes would be in the vicinity of zero. Martins and Stedinger (2000) proposed a prior of the form

$$p(\xi) = \frac{(0.5+\xi)^{p-1}(0.5-\xi)^{q-1}}{\mathrm{beta}(p,q)}, \qquad \xi \in [-0.5, 0.5]$$

a shifted $\mathsf{beta}(p,q)$ prior.

On the contrary, the **maximal data information** (MDI) prior (Zellner 1971) is defined in terms of entropy,

$$p(\boldsymbol{\theta}) = \exp \mathsf{E}\{\log f(Y \mid \boldsymbol{\theta})\}.$$

It is an objective prior that reflects little about the parameter and leads to inferences that have good frequentist property.

For the generalized Pareto distribution, $p(\xi) \propto \exp(-\xi)$. In this particular case, however, it is improper without modification since $\lim_{\xi\to-\infty} \exp(-\xi) = \infty$, and the prior density increases without bound as ξ becomes smaller.

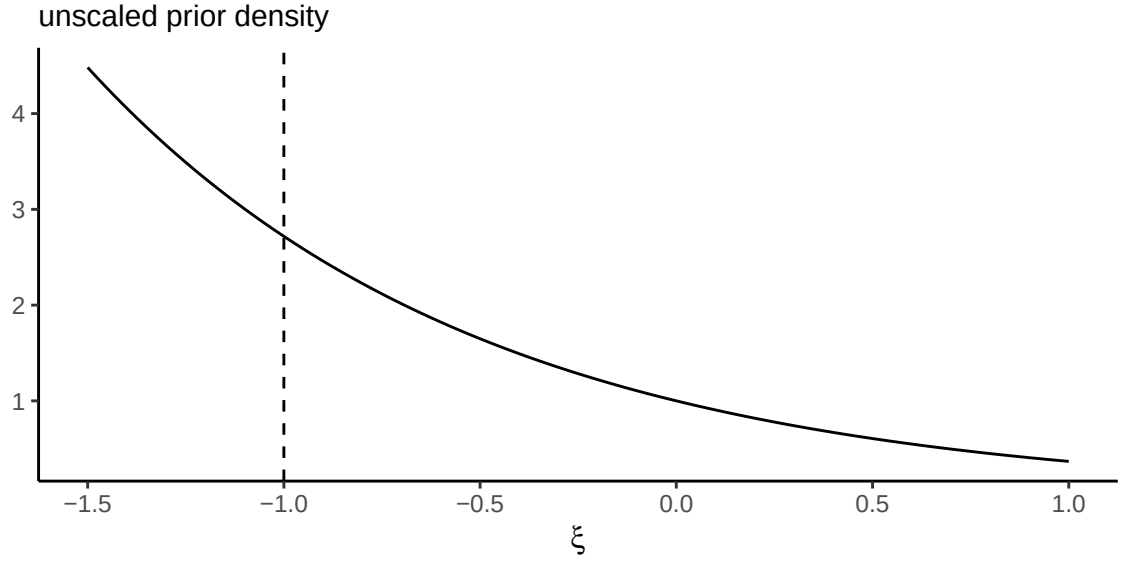

Figure 3.4: Unscaled maximum data information (MDI) prior density.

If we restrict the range of the MDI prior $p(\xi)$ to $\xi \geq -1$, then $p(\xi+1) \sim \mathsf{expo}(1)$ and the resulting posterior is proper (Northrop and Attalides 2016, Zhang.Shaby:2023). While being

"objective", it is perhaps not much suitable as it puts mass towards lower values of the shape, an undesirable feature.

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

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

Proposition 3.2 (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 (Simpson et al. 2017) or the Wasserstein distance (Bolin, Simas, and Xiong 2023) 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 distance between the prior densities is then set to $d(\zeta) = \{2\mathsf{KL}(p_\zeta \mid\mid p_0)\}^{1/2}$, where 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 divergence 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 $p(\zeta \mid \lambda) = \lambda \exp(-\lambda d(\zeta)) \left|\partial d(\zeta)/\partial \zeta\right|$. To choose λ, the authors recommend elicitation of a pair (Q_ζ, α), where Q_ζ is the quantile at level $1-\alpha$, such that $\Pr(\lambda > Q_\zeta) = \alpha$.

The construction of Wasserstein complexity priors (Bolin, Simas, and Xiong 2023) is more involved, but those priors are also parametrization-invariant and well-defined even when the Kullback–Leibler divergence limit does not exist.

Example 3.12 (Penalized complexity prior for random effects models)**.** Bolin, Simas, and Xiong (2023) consider a Gaussian prior for independent and identically random effects $\boldsymbol{\alpha}$, of the form $\alpha_j \mid \zeta \sim \mathsf{Gauss}(0, \zeta^2)$ 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, with density

$$p(\zeta \mid \lambda) = \lambda \exp(-\lambda\zeta).$$

We can elicit a high quantile Q_ζ at tail probability α for the standard deviation parameter ζ and set $\lambda = -\ln(\alpha/Q_\zeta)$.

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 ($\phi = 0$), 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.

Remark 3.1 (Variance parameters in hierarchical models). 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 $\zeta, \zeta \sim \mathsf{InvGamma}(\epsilon, \epsilon)$ 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

$$\begin{aligned} Y_{ij} &\sim \mathsf{Poisson}\{\mu = \exp(\beta_j + \alpha_i)\}, \\ \beta_j &\sim \mathsf{Gauss}(0, 100), \\ \alpha_i &\sim \mathsf{Gauss}(0, \kappa^2), \\ \kappa &\sim \mathsf{Student}_+(0, 1, 3). \end{aligned}$$

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).

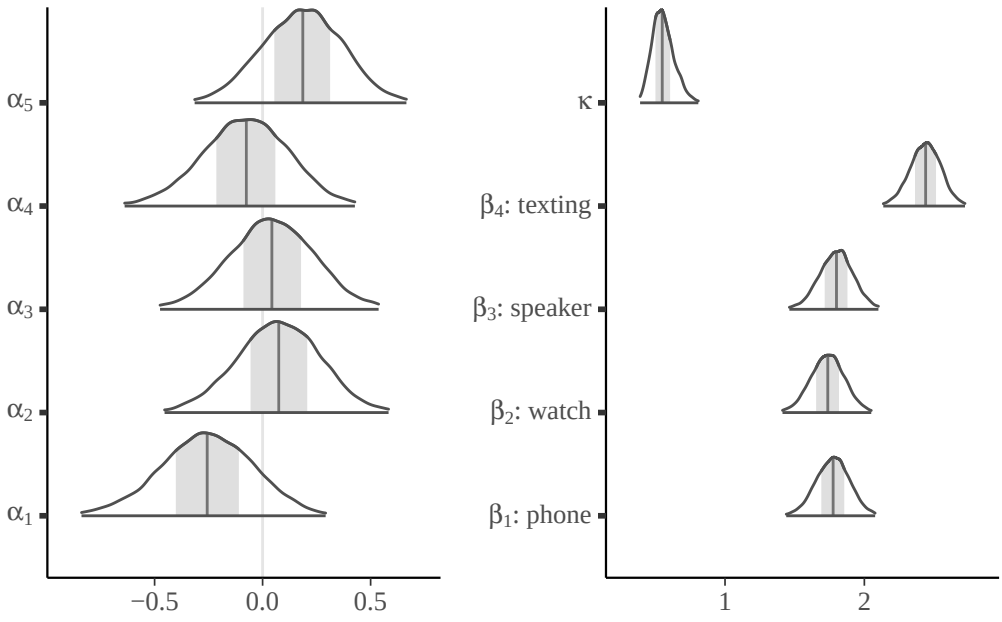

Figure 3.5: Posterior density plots with 50% credible intervals and median value for the random effects of the first five individuals (left) and the fixed effects and random effect variance (right).

3.6 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.

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.

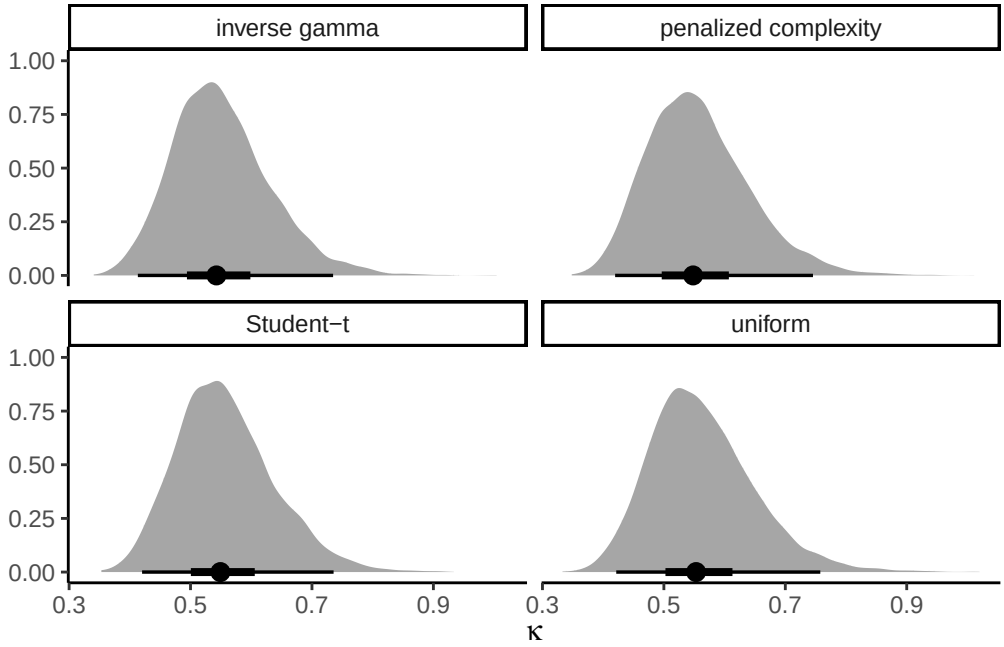

Figure 3.6: Posterior density of the scale of the random effects with uniform, inverse gamma, penalized complexity and folded Student-t with three degrees of freedom. The circle denotes the median and the bars the 50% and 95% percentile credible intervals.

Example 3.17 (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 3.7 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.

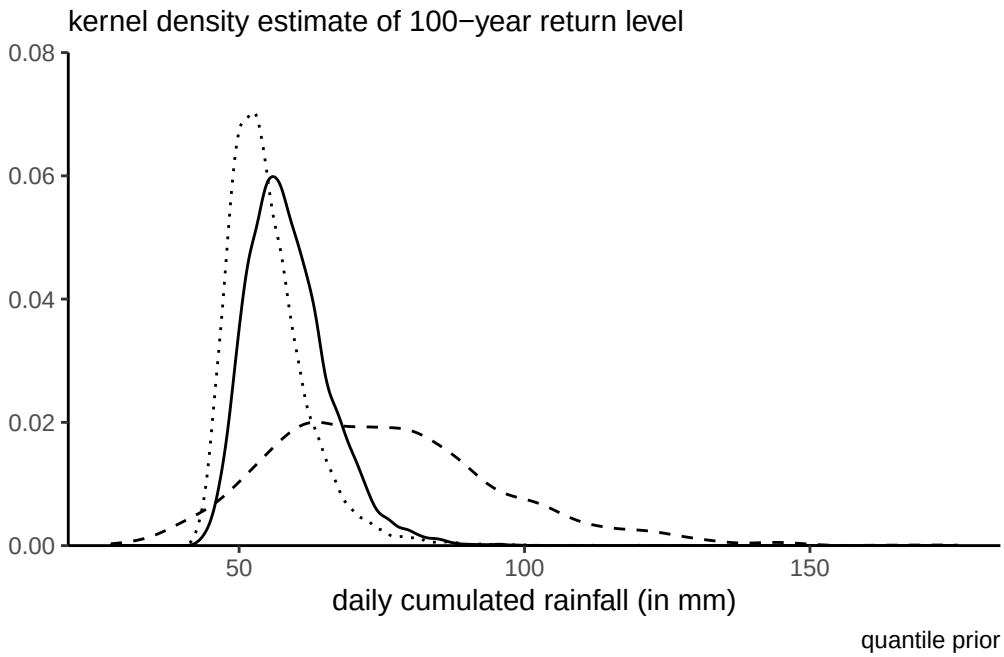

Figure 3.7: Kernel density estimates of draws from the posterior distribution of 100 year return levels with a Coles–Tawn quantile prior (full line) and from the corresponding prior predictive (dashed). The dotted line gives the posterior distribution for a maximum domain information prior on the shape with improper priors on location and scale.

! Summary:

- Priors are distributions for the parameters. In multi-parameter models, they can be specified through a joint distribution or assumed independent apriori (which does not translate into independence a posteriori).
- Priors are not invariant to reparametrization, except when they are constructed with this property (e.g., Jeffrey's prior or penalized-complexity priors).
- Improper priors may lead to improper posterior.
- Priors that restrict the domain of $\boldsymbol{\theta}$ will also restrict the posterior. These are useful to avoid regions that are implausible or impossible.
- Physical knowledge of the system can be helpful to specify sensible values of the

prior through moment matching.
- Conjugate priors facilitate derivations, but are mostly chosen for convenience.
- Generally, the prior has constant weight $\mathrm{O}(1)$, relative to $\mathrm{O}(n)$ for the likelihood. The posterior is thus dominated in most circumstances by the likelihood.
- We can compute the prior to posterior gain by comparing their density (if the prior is proper).
- For many (conjugate) priors, we can view some function of the parameter as given a prior number of observations (in Gaussian models, binomial, gamma, etc.)
- Informative priors can be used to specify expert knowledge about the system. This will impact the posterior, but often in a sensible manner, thereby regularizing or improving posterior inference.

4 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.

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. 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 Monte Carlo methods

Monte Carlo methods relies on the ability to simulate random variable. If the quantile function admits a closed-form, we can use this to simulation. Recall that if a random variable X has distribution function F, then we can define it's **generalized inverse**

$$F^{-1}(u) = \inf\{x : f(x) \geq u\}$$

and if G is continuous, then $F(X) \sim \mathsf{unif}(0,1)$. We can thus simulate data using the quantile function $F^{-1}(U)$, with $U \sim \mathsf{unif}(0,1)$.

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

Example 4.1 (Simulation of exponential variates)**.** The distribution function of $Y \sim \mathsf{expo}(\lambda)$ is $F(y) = \exp(-\lambda y)$, so the quantile function is $F^{-1}(u) = -\log(u)/\lambda$

```
n <- 1e4L
-log(runif(n)) / 2 #simulate expo(2)
```

Theorem 4.1 (Fundamental theorem of simulation)**.** *Consider a d-variate random vector $\boldsymbol{X}$, independently $U \sim \mathsf{unif}(0,1)$ and $c > 0$ any positive constant. If $(\boldsymbol{X}, U)$ is uniformly distributed on the set*

$$\mathcal{A}_f = \{(\boldsymbol{x}, u) : 0 \leq u \leq cf(\boldsymbol{x})\},$$

then $\boldsymbol{X}$ has density $f(\boldsymbol{x})$. Conversely, if $\boldsymbol{X}$ has density $f(\boldsymbol{x})$ and $U \sim \mathsf{unif}(0,1)$ independently, then $[\boldsymbol{X}, cUf(\boldsymbol{X})]$ is uniformly distributed on $\mathcal{A}_f$

We can thus view f as the marginal density of $\boldsymbol{X}$ since $f(\boldsymbol{x}) = \int_0^{f(\boldsymbol{x})} \mathrm{d}u$. If we can simulate uniformly from $\mathcal{A}_f$, then, we can discard the auxiliary variable u. See Devroye (1986), Theorem 3.1 for a proof.

The fundamental theorem of simulation underlies rejection sampling, the generalized ratio of uniform and slice sampling. The density function needs only to be known up to normalizing constant thanks to the arbitrariness of c, which will also allow us to work with unnormalized density functions.

Proposition 4.1 (Rejection sampling)**.** *Rejection sampling (also termed accept-reject algorithm) samples from a random vector with density $p(\cdot)$ by drawing candidates from a 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 $p(\cdot)$.*

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.

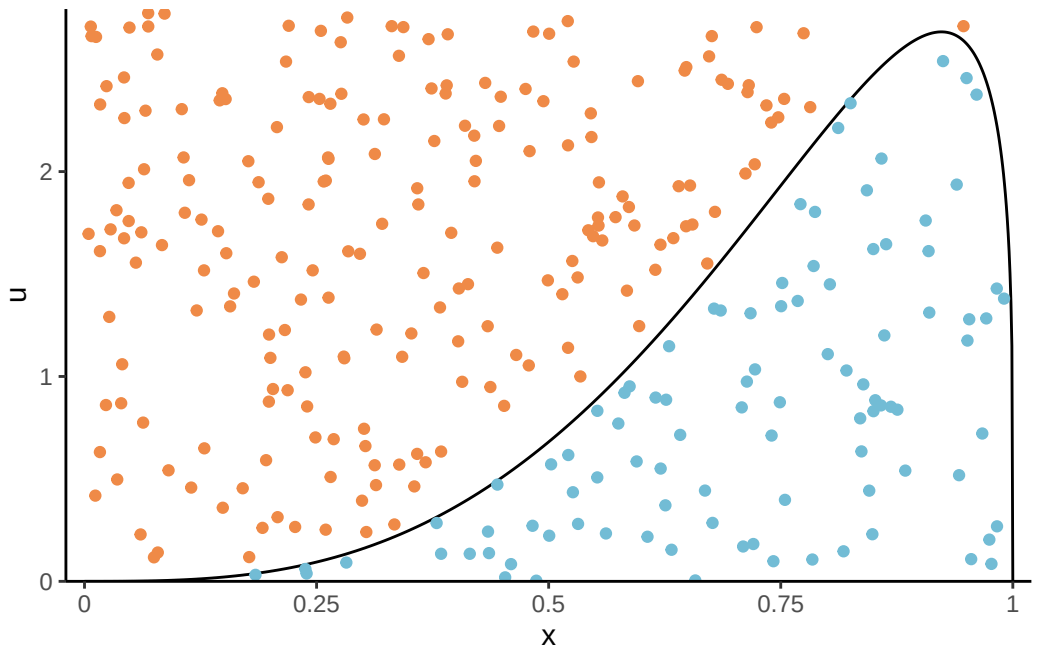

Figure 4.1: Illustration of the fundamental theorem of simulation. All points in blue below the density curve belong to $\mathcal{A}_f$.

Example 4.2 (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.

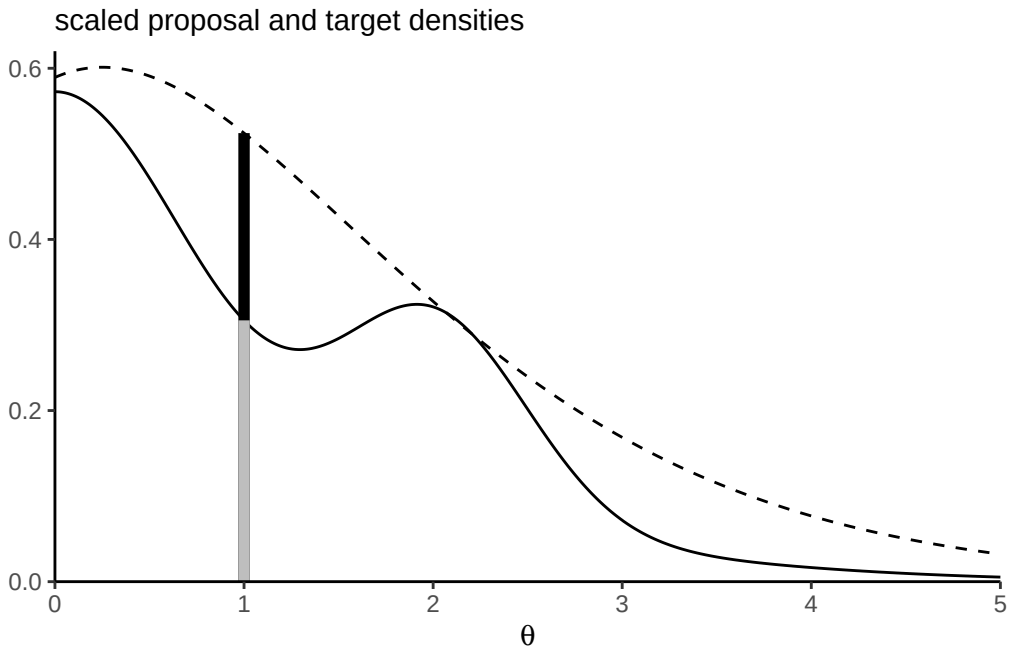

Figure 4.2: Target density (full) and scaled proposal density (dashed): the vertical segment at $\theta = 1$ shows the percentage of acceptance for a uniform slice under the scaled proposal, giving an acceptance ratio of 0.58.

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

```
[1] 0.34002
```

```
# 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. A different use for the simulations is to approximate integrals numerically. 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.

Proposition 4.2 (Generalized ratio-of-uniform)**.** *An exact simulation algorithm is described in Kinderman and Monahan (1977) and extended Wakefield, Gelfand, and Smith (1991) for random number generation in low dimensions. Consider a d-vector $\boldsymbol{X}$ with density $cf(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^+$ with support $\mathcal{S} \subseteq \mathbb{R}^d$, and $c > 0$ is the (possibly unknown) normalizing constant. Consider variables $\boldsymbol{u} = (u_0, u_1, \ldots, u_d)$ uniformly distributed over the set*

$$\mathcal{B}(r) = \left\{ (u_0, u_1, \ldots, u_d) \in \mathbb{R}^{d+1} : 0 < u_0 \leq \left[f\left(\frac{u_1}{u_0^r}, \ldots, \frac{u_d}{u_0^r} \right) \right]^{1/(rd+1)} \right\}$$

for some positive radius parameter $r \geq 0$. The measure of the set $\mathcal{B}(r) = c^{-1}(1+rd)^{-1}$. By Theorem 4.1, $(u_1/u_0^r, \ldots, u_d/u_0^r)$ is drawn from the renormalized density $cf(x)$. The challenge lies in simulating $\boldsymbol{u}$ uniformly over $\mathcal{B}_r$, but we can use accept-reject if the later is enclosed in a bounding box $\mathbb{B}$, keeping only samples that satisfy the constraints. If, over $\mathcal{S}$, $f(\boldsymbol{x})$ and $x_i^{rd+1} f(\boldsymbol{x})^r$ for $i = 1, \ldots, d$ are bounded then we can enclose $\mathcal{B}(r)$ within the $(d+1)$-dimensional bounding box

$$\mathbb{B} = \{ a_j(r) \leq u_i \leq b_j(r); j = 0, \ldots, d \},$$

with $a_0(r) = 0$. The parameters of the bounding box are

$$\begin{aligned} b_0(r) &= \sup_{\boldsymbol{x} \in \mathcal{S}} f(\boldsymbol{x})^{1/(rd+1)}, \\ a_j(r) &= \inf_{\substack{\boldsymbol{x} \in \mathcal{S} \\ x_i \leq 0}} x_i \, f(\boldsymbol{x})^{r/(rd+1)}, \\ b_j(r) &= \sup_{\substack{\boldsymbol{x} \in \mathcal{S} \\ x_i \geq 0}} x_i \, f(\boldsymbol{x})^{r/(rd+1)}, \end{aligned}$$

The probability of acceptance $p_a(d, r)$ of a point simulated uniformly over the bounding box depends on both the radius and the dimension and is

$$p_a(d, r) = c \left[(rd+1) \, b_0(r) \prod_{j=1}^{d} \{ b_j(r) - a_j(r) \} \right]^{-1}.$$

Wakefield, Gelfand, and Smith (1991) propose using $r = 1/2$ and relocating the mode of f to the origin increase the acceptance rate. Northrop proposes to use a Box–Cox transformation (Box and Cox 1964) together with a rotation in the software Northrop (2024) to improve the acceptance rate. The bounding box may exist only for certain values of r; see the `rust` package vignette for technical details and examples.

Example 4.3 (Ratio-of-uniform for insurance loss)**.** We use the ratio-of-uniform algorithm presented in Proposition 4.2 for the data from Example 2.6 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 $Y \sim \text{gen.Pareto}(\tau, \xi)$ to exceedances above 10 millions krones to the `danish` fire insurance data, using a truncated maximal data information prior $p(\tau, \xi) \propto \tau^{-1} \exp(-\xi + 1)\mathrm{I}(\xi > -1)$.

```
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, #mle
```

```
  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.

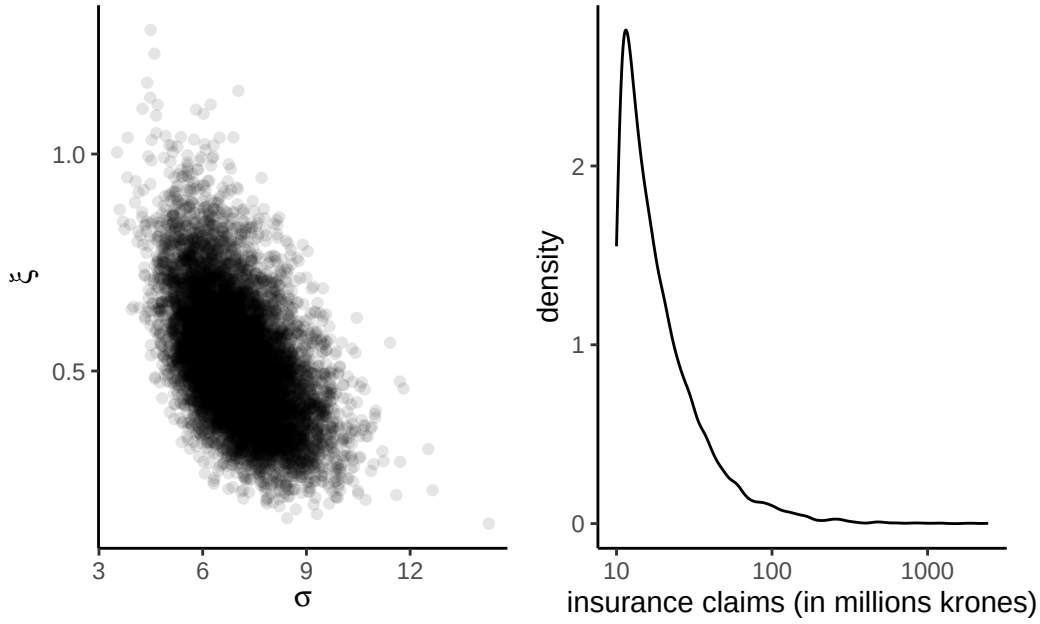

Figure 4.3: Scatterplot of posterior samples from the generalized Pareto model applied to Danish fire insurance losses above 10 millions krones, with maximal data information prior (left) and posterior predictive density on log scale (right).

Proposition 4.3 (Monte Carlo integration)**.** *Specifically, suppose we are interested in the average* $\mathsf{E}\{g(X)\}$ *of* X *with density or mass function* f *supported on* $\mathcal{X}$ *for some function* g*. Monte Carlo integration proceeds by drawing* B *independent samples* $x_1, \ldots, x_B$ *from density*

p and evaluating the empirical average of g, with

$$\mathsf{E}\{g(X)\} = \int_{\mathcal{X}} g(x)p(x)\mathrm{d}x \approx \widehat{\mathsf{E}}\{g(X)\} = \frac{1}{B}\sum_{b=1}^{B} g(x_b).$$

By the law of large number, this estimator is convergent when $B \to \infty$ provided that the expectation is finite. If the variance of $g(X)$ is finite, we can approximate the latter by the sample variance of the simple random sample and obtain the Monte Carlo standard error of the estimator

$$\mathsf{se}^2[\widehat{\mathsf{E}}\{g(X)\}] = \frac{1}{B(B-1)}\sum_{b=1}^{B}\Big[g(x_b) - \widehat{\mathsf{E}}\{g(X)\}\Big]^2.$$

We can also use a similar idea to evaluate the integral of $g(X)$ if X has density p, by drawing instead samples from q. This is formalized in the next proposition.

Proposition 4.4 (Importance sampling)**.** *Consider a random variable X with density $p(x)$ supported on $\mathcal{X}$. We can calculate the integral $\mathsf{E}_p\{g(X)\} = \int_{\mathcal{X}} g(x)p(x)\mathrm{d}x$ by considering instead draws from a density $q(\cdot)$ with nested support, $\mathcal{X} \subseteq \mathrm{supp}(q)$. Then,*

$$\mathsf{E}\{g(X)\} = \int_{\mathcal{X}} g(x)\frac{p(x)}{q(x)}q(x)\mathrm{d}x$$

and we can proceed similarly by drawing samples from q. This is most useful when the variance is finite, which happens if the integral

$$\int_{\mathcal{X}} g^2(x)\frac{p^2(x)}{q(x)}\mathrm{d}x < \infty.$$

An alternative Monte Carlo estimator, which is biased but has lower variance, is obtained by drawing independent $x_1, \ldots, x_B$ from q and taking instead the weighted average of

$$\widetilde{\mathsf{E}}\{g(X)\} = \frac{B^{-1}\sum_{b=1}^{B} w_b g(x_b)}{B^{-1}\sum_{b=1}^{B} w_b}.$$

with weights $w_b = p(x_b)/q(x_b)$. The latter equal 1 on average, so one could omit the denominator without harm. The standard error for the independent draws equals

$$\mathsf{se}^2[\widetilde{\mathsf{E}}\{g(X)\}] = \frac{\sum_{b=1}^{B} w_b^2\Big[g(x_b) - \widetilde{\mathsf{E}}\{g(X)\}\Big]^2}{\Big(\sum_{b=1}^{B} w_b\Big)^2}.$$

Example 4.4 (Importance sampling for the variance of a beta distribution)**.** Consider $X \sim \mathsf{beta}(\alpha, \alpha)$ for $\alpha > 1$ with $\mathsf{E}(X) = 0.5$ since the density is symmetric. We tackle the estimation of the variance, which can be written as $\mathsf{E}\{(X-0.5)^2\}$. While we can easily derive the theoretical expression, equal to $\mathsf{Va}(X) = \{4 \cdot (2\alpha+1)\}^{-1}$, we can also use Monte Carlo integration as proof of concept.

Rather than simulate directly from our data generating mechanism, we can use an importance sampling density $q(x)$ which puts more mass away from 0.5 where the integral is zero. Consider the equiweighted mixture of $\mathsf{beta}(\alpha, 3\alpha)$ and $\beta(3\alpha, \alpha)$, which is bimodal. Figure 4.4 shows the function we wish to integrate, the density and the importance sampling density, and the weighting function $p(x)/q(x)$ of the first 50 observations drawn from $q(x)$ with $\alpha = 1.5$. The variance ratio shows an improvement of more than 9% for the same Monte Carlo sample size.

```
B <- 2e4L
alpha <- 1.5
factor <- 3
# Mode at the mean 0.5
X0 <- rbeta(n = B, alpha, alpha)
px <- function(x){dbeta(x, alpha, alpha)}
# Importance sampling density - mixture of two betas (alpha, factor*alpha)
X1 <- ifelse(runif(B) < 0.5, rbeta(B, alpha, factor*alpha), rbeta(B, factor*alpha, alpha))
qx <- function(x){0.5*dbeta(x, alpha, factor*alpha) + 0.5*dbeta(x, factor*alpha, alpha)}
# Function to integrate - gives variance of a symmetric beta distribution
g <- function(x){(x - 0.5)^2}
# Weights for importance sampling
w <- px(X1)/qx(X1)
# Monte Carlo integration
mc_est <- mean(g(X0))
mc_var <- var(g(X0))/B
# Importance sampling weighted mean and variance
is_est <- weighted.mean(g(X1), w = w) # equivalent to mean(g(X1)*w)/mean(w)
is_var <- sum(w^2*(g(X1) - is_est)^2)/ (sum(w)^2)
# True value for the beta variance
th_est <- 1/(4*(2*alpha+1))
# Point estimates and differences
round(c(true = th_est, "monte carlo" = mc_est, "importance sampling" = is_est),4)
```

```
                true         monte carlo importance sampling
              0.0625              0.0622              0.0627
```

```
# Ratio of std. errors for means
mc_var/is_var # value > 1 means that IS is more efficient
```

```
[1] 1.087118
```

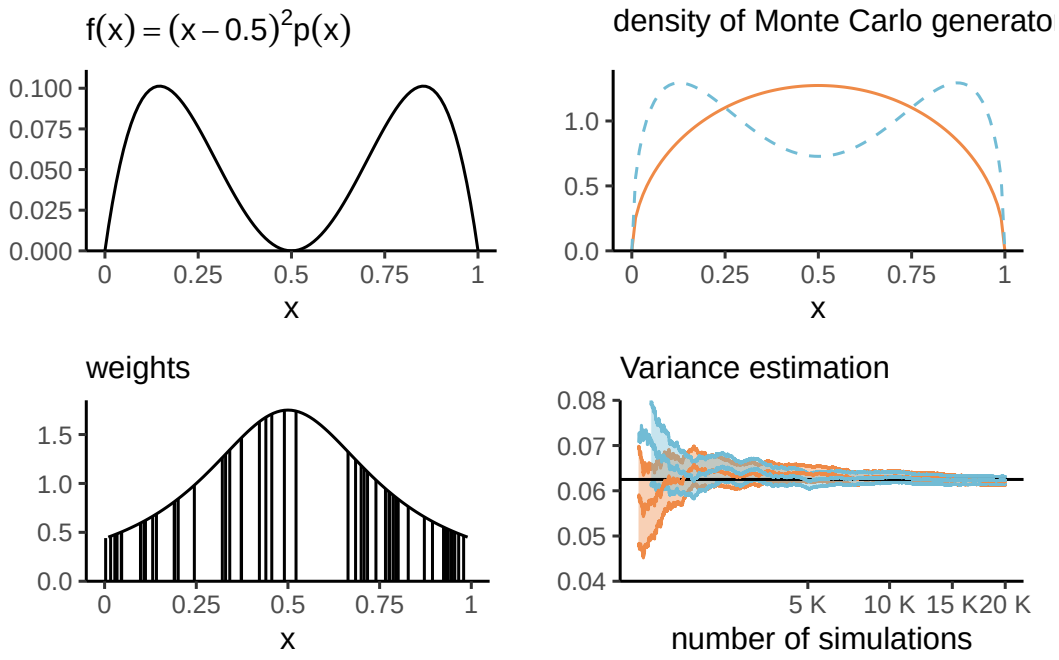

Figure 4.4: Monte Carlo integration with importance sampling for the variance of a symmetric beta distribution. Top left: variance function $f(x) = (x - 0.5)^2$. Top right: density of $\mathsf{beta}(\alpha, \alpha)$ (orange) and importance sampling mixture distribution (blue). Bottom left: weighting function and weights for 50 first importance sampling draws. Bottom right: sample paths of Monte Carlo mean with Wald 95% confidence intervals.

Example 4.5 (Expectations of functions of a gamma variate)**.** 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 the variance as $\mathsf{E}(X^2) - \{\mathsf{E}(X)\}^2$.

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: for our examples, 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 4.5 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 4.5 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.

Example 4.6 (Tail probability of a Gaussian distribution)**.** Consider estimation of the probability that a standard Gaussian random variable exceeds $a = 4$, which is $p = 1 - \Phi(a)$. We can use standard numerical approximations to the distribution function implemented in any software package, which shows this probability is roughly one in 3.1574×10^4.

If we consider a truncated Gaussian above a, then the integral of $\mathsf{I}(x > a)$ is one (since the truncated Gaussian is a valid density). Thus, we can estimate rather the normalizing constant by simulating standard Gaussian and comparing this with the importance sampling estimator, using the knowledge of the value of the integral to derive rather the normalizing constant. Monte Carlo integration from with $B = 10^6$ is demonstrated using the following code

```
a <- 4
B <- 1e6L # 1 million draws
exact <- pnorm(a, lower.tail = FALSE)
# Vanilla Monte Carlo
X <- rnorm(B)
mc <- mean(X >= 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.

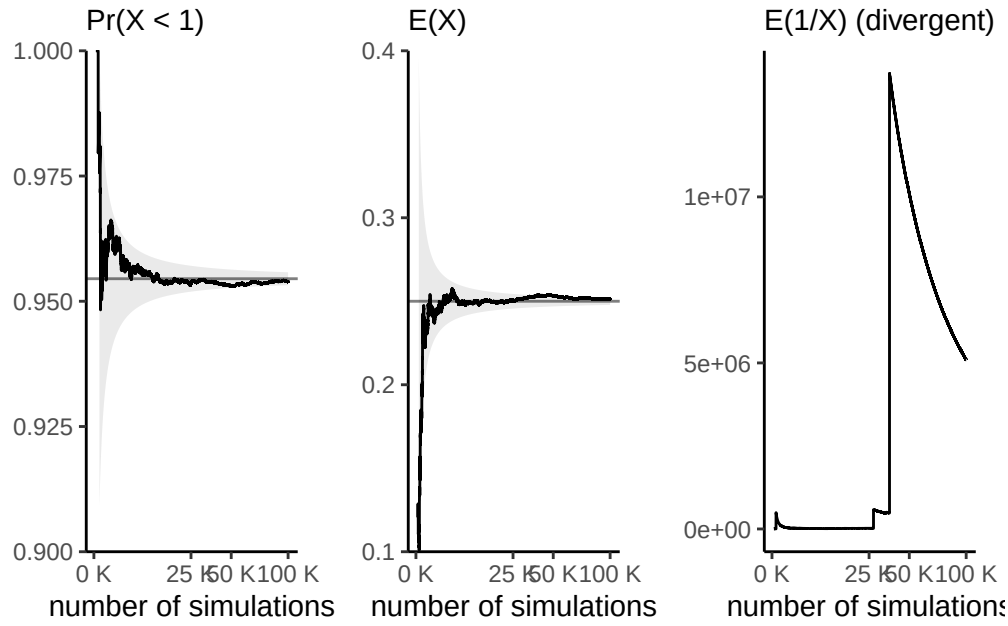

Figure 4.5: Running mean trace plots for $g(x) = \mathrm{I}(x < 1)$ (left), $g(x) = x$ (middle) and $g(x) = 1/x$ (right) for a Gamma distribution with shape 0.5 and rate 2, as a function of the Monte Carlo sample size.

```
# Importance sampling with Rayleigh
Y <- sqrt(a^2 + 2*rexp(B))
drayleigh <- function(x, a){ x*exp((a^2-x^2)/2)}
is <- mean(dnorm(Y)/drayleigh(Y, a = a))
# Relative error
c(mc = (mc - exact)/exact, is = (is - exact)/exact)
```

```
           mc            is
-2.119405e-02 -2.927613e-05
```

4.2 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. The treatment of Markov chains in this chapter is rather loose and non-formal. Readers can refer to Chapter 6 of Robert and Casella (2004) for a more rigourous exposition.

Definition 4.1 (Discrete-time stochastic process)**.** 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.

Definition 4.2 (Stationarity)**.** A stochastic (i.e., random) process is (strongly) stationary if the distribution of $\{X_1, \ldots, X_n\}$ is the same as that of $\{X_{t+1}, \ldots X_{n+t}\}$ for any value of t and given n.

It is weakly stationary if the expected value is constant, meaning $\mathsf{E}(X_t) = \mu$ for all time points t, and the covariance at lag h, $\mathsf{Cov}(X_t, X_{t+h}) = \gamma_h$, does not depend on t. Strong stationarity implies weak stationarity.

Definition 4.3 (Markov property)**.** A stochastic process is markovian 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.

Definition 4.4 (Ergodicity)**.** Let $\{Y_t\}$ is a weakly stationary sequence with mean $\mathsf{E}(Y_t) = \mu$ and $\gamma_h = \mathsf{Cov}(Y_t, Y_{t+h})$. Then, if the autocovariance series is convergent, meaning

$$\sum_{t=0}^{\infty} |\gamma_h| < \infty,$$

then $\{Y_t\}$ is ergodic for the mean and $\overline{Y} \xrightarrow{\text{p}} \mu$. In other words, the ergodic theorem is a law of large numbers for stochastic processes that allows for serial dependence between observations, provided the latter is not too large.

Ergodicity means that two segments of a time series far enough apart act as independent.

Proposition 4.5 (Ergodicity and transformations)**.** *Any transformation $g(\cdot)$ of a stationary and ergodic process $\{Y_t\}$ retains the ergodicity properties, so $\overline{g} = T^{-1}\sum_{t=1}^{T} g(Y_t) \to \mathsf{E}\{g(Y_t)\}$ as $T \to \infty$.*

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

Example 4.7 (Stationarity and AR(1))**.** Consider a Gaussian $\mathsf{AR}(1)$ model with conditional mean and variance $\mathsf{E}_{Y_t|Y_{t-1}}(Y_t) = \mu + \phi(Y_{t-1} - \mu)$ and $\mathsf{Va}_{Y_t|Y_{t-1}}(Y_t) = \sigma^2$. Using the law of iterated expectation and variance, if the process is weakly stationary, then $\mathsf{E}_{Y_t}(Y_t) = \mathsf{E}_{Y_{t-1}}(Y_{t-1})$

$$\begin{aligned}\mathsf{E}_{Y_t}(Y_t) &= \mathsf{E}_{Y_{t-1}}\left\{\mathsf{E}_{Y_t|Y_{t-1}}(Y_t)\right\}\\ &= \mu(1-\phi) + \phi\mathsf{E}_{Y_{t-1}}(Y_{t-1})\end{aligned}$$

and so the unconditional mean is μ. For the variance, we have

$$\begin{aligned}\mathsf{E}_{Y_t}(Y_t) &= \mathsf{E}_{Y_{t-1}}\left\{\mathsf{Va}_{Y_t|Y_{t-1}}(Y_t)\right\} + \mathsf{Va}_{Y_{t-1}}\left\{\mathsf{E}_{Y_t|Y_{t-1}}(Y_t)\right\}\\ &= \sigma^2 + \mathsf{Va}_{Y_{t-1}}\left\{\mu + \phi(Y_{t-1}-\mu)\right\}\\ &= \sigma^2 + \phi^2\mathsf{Va}_{Y_{t-1}}(Y_{t-1}).\end{aligned}$$

and we recover the formulas from Example 1.17.

The covariance at lag k, in terms of innovations, gives

$$\gamma_k = \mathsf{Co}(Y_t, Y_{t-k}) = \mathsf{Va}(\phi Y_{t-1}, Y_{t-k}) + \mathsf{Va}(\varepsilon_t, Y_{t-k}) = \phi\gamma_{k-1}$$

so by recursion $\gamma_k = \phi^k\mathsf{Va}(Y_t)$.

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 $p(Y_t \mid Y_{t-1})$.

When can we use output from a Markov chain in place of independent Monte Carlo draws? The assumptions laid out in the ergodic theorem, which provides guarantees for the convergence of sample average, are that the chain is irreducible. If the chain is also acyclic, the chain has a unique stationary distribution.

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 $p(X_t \mid X_{t-1})$. This results in correlated draws, due to the reliance on the previous observation.

Proposition 4.6 (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*

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

In the independent case, the covariance is zero so we get the sum of variances. If the process is stationary, the covariance 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 $\mathsf{Cor}(Y_t, Y_{t+k}) = \rho_k$. Under technical conditions[4], a central limit theorem applies and we get an asymptotic variance for the mean of

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

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.*

Definition 4.5 (Effective sample size)**.** Loosely speaking, the effective sample size is the equivalent number of observations if the marginal posterior draws were independent. We define it as

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

where ρ_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.

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

4.2.1 Discrete Markov chains

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* i*th 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 $\mathbf{P}_1$ *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: $\mathbf{P}_2$ *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.*

$$\mathbf{P}_1 = \begin{pmatrix} 0.5 & 0.3 & 0.2 \\ 0 & 0.4 & 0.6 \\ 0 & 0.5 & 0.5 \end{pmatrix}, \qquad \mathbf{P}_2 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

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, consider a transition $\mathbf{P}_3$ *on* $1,\dots,5$ *defined as*

$$\mathbf{P}_3 = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 0 & 0 \\ 0 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 0 \\ 0 & 0 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ 0 & 0 & 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

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* $\mathbf{P}_1$, $(1/3, 1/3, 1/3)$ *for* $\mathbf{P}_2$ *and* $(1,2,2,2,1)/8$ *for* $\mathbf{P}_3$.

While the existence of a stationary distribution require aperiodicity, the latter is not really important from a computational perspective as ergodicity holds without it.

Figure 4.7 shows the path of the random walk driven by $\mathbf{P}_3$ *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 the sample proportions to converge to the stationary distribution proportions.*

Since we will be dealing with continuous random variables in later chapters, we use transition kernels rather than transition matrices, but the intuition will carry forward.

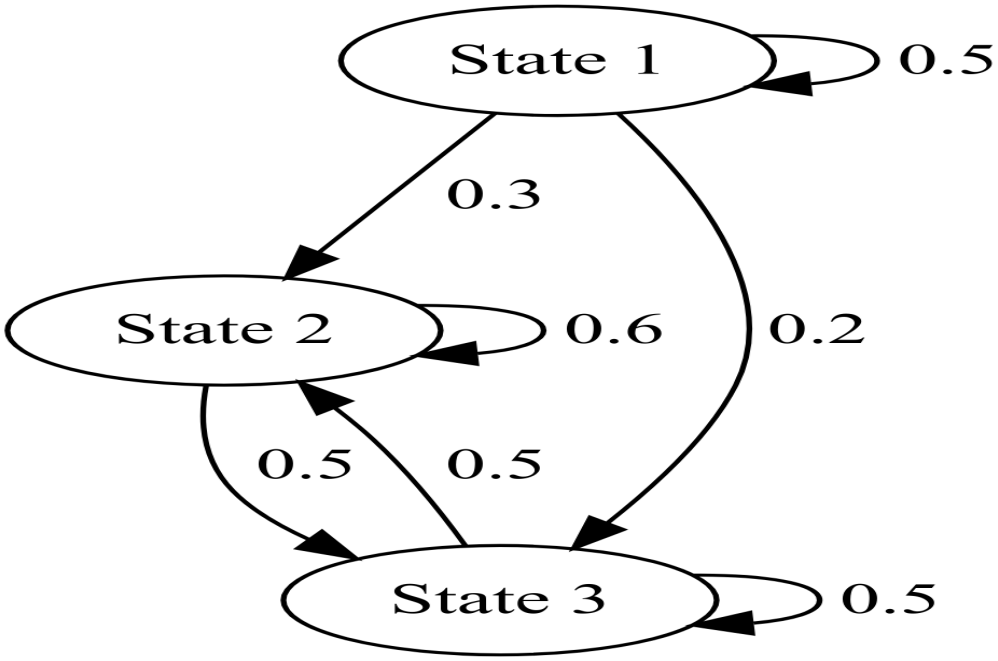

Figure 4.6: Graphical representation of the transition matrix $\mathbf{P}_1$.

Definition 4.6 (Transition kernel)**.** A transition kernel $K(\boldsymbol{\theta}^{\text{cur}}, \boldsymbol{\theta}^{\text{prop}})$ proposes a move from the current value $\boldsymbol{\theta}^{\text{cur}}$ to a proposal $\boldsymbol{\theta}^{\text{prop}}$.

Example 4.8 (Effective sample size of first-order autoregressive process)**.** The lag-k correlation of the stationary autoregressive process of order 1 is $\rho_k = \phi^k$, so summing the series gives an effective sample size for B draws of $B(1-\phi)/(1+\phi)$. The price to pay for having correlated samples is inefficiency: the higher the autocorrelation, the larger the variability of our mean estimators.

We can see from Figure 4.8 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 variance estimates for the mean with an $\mathsf{AR}(1)$ process with $\phi = 0.75$ than with an iid sample, we would need 7 times as many observations: this is the prize to pay for autocorrelation.

Proposition 4.7 (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*

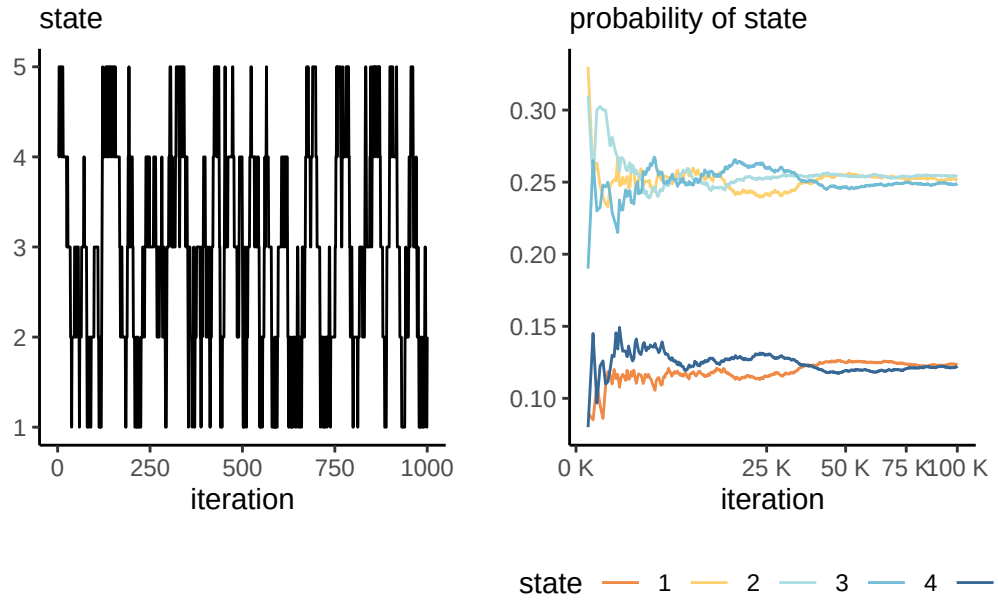

Figure 4.7: Discrete Markov chain on integers from 1 to 5, with transition matrix $\mathbf{P}_3$, with traceplot of 1000 first iterations (left) and running mean plots of sample proportion of each state visited per 100 iterations (right).

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 4.9 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

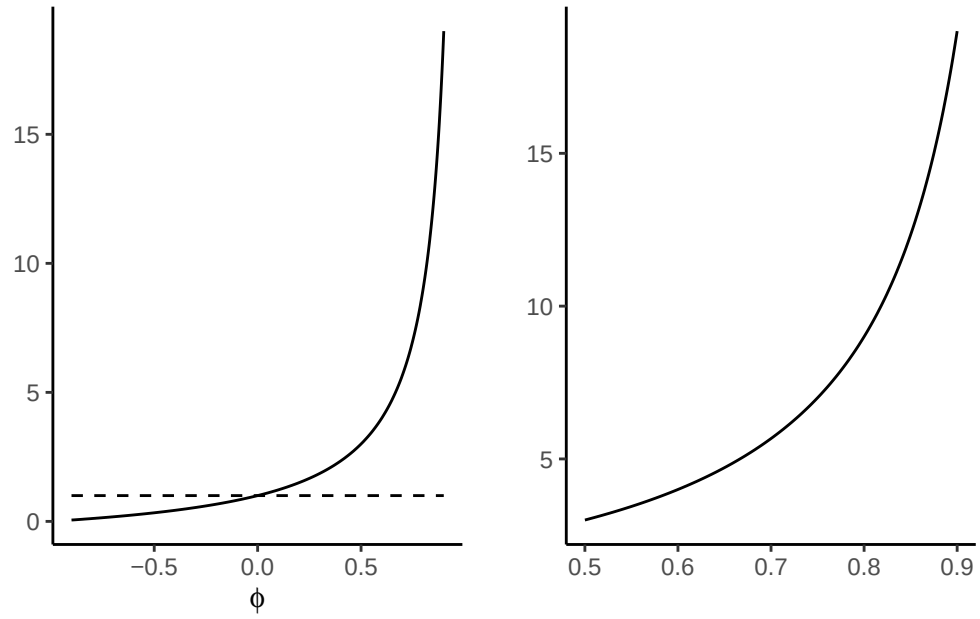

Figure 4.8: Scaled asymptotic variance of the sample mean for a stationary autoregressive first-order process with unit variance (full line) and a corresponding sample of independent observations with the same marginal variance (dashed line). The right panel gives the ratio of variances for a restricted range of positive correlation coefficients.

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.

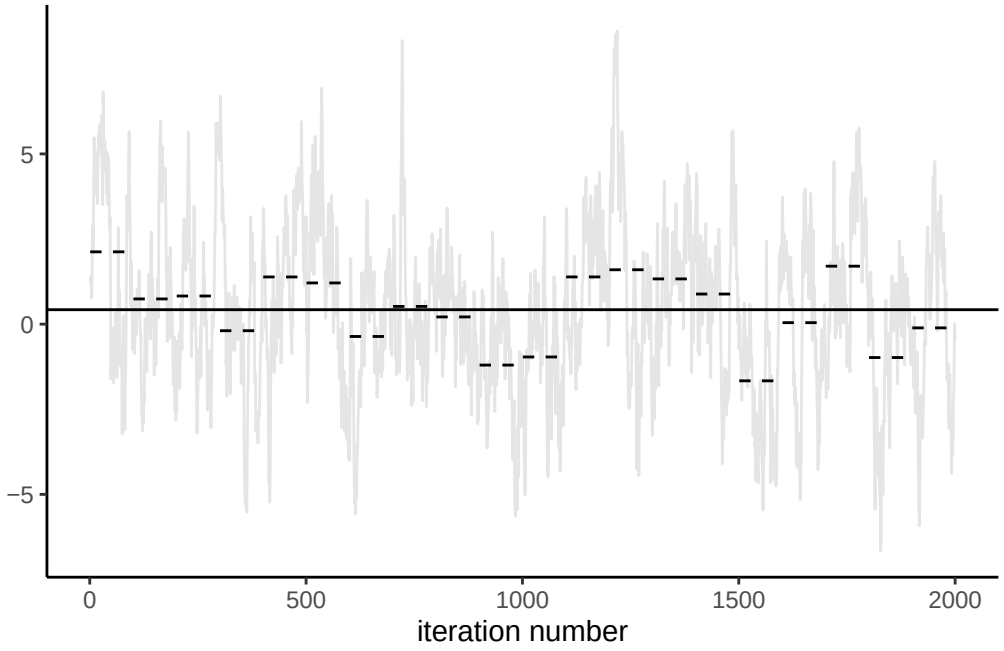

Figure 4.9: Calculation of the standard error of the posterior mean using the batch method.

5 Metropolis–Hastings algorithm

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.

> ! **Learning objectives**:
>
> At the end of the chapter, students should be able to
>
> - implement a Metropolis–Hastings algorithm to draw samples from the posterior
> - tune proposals to obtain good mixing properties.

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 5.1 (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*

$$R = \frac{p(\boldsymbol{\theta}_t^\star)}{p(\boldsymbol{\theta}_{t-1})} \frac{q(\boldsymbol{\theta}_{t-1} \mid \boldsymbol{\theta}_t^\star)}{q(\boldsymbol{\theta}_t^\star \mid \boldsymbol{\theta}_{t-1})} \tag{5.1}$$

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 following theoretical details provided for completeness only.

Definition 5.1 (Detailed balance)**.** If our target is $p(\cdot)$, then the Markov chain satisfies the **detailed balance** condition with respect to $p(\cdot)$ if

$$K(\boldsymbol{\theta}^{\text{cur}}, \boldsymbol{\theta}^{\text{prop}})p(\boldsymbol{\theta}^{\text{cur}}) = K(\boldsymbol{\theta}^{\text{prop}}, \boldsymbol{\theta}^{\text{cur}})p(\boldsymbol{\theta}^{\text{prop}}).$$

If a Markov chain satisfies the detailed balance with respect to $p(\cdot)$, then the latter is necessarily the invariant density of the Markov chain and the latter is reversible.

Proposition 5.2 (Metropolis–Hastings satisfies detailed balance)**.** *The Metropolis–Hastings algorithm has transition kernel for a move from $\boldsymbol{x}$ to a proposal $\boldsymbol{y}$*

$$K(\boldsymbol{x}, \boldsymbol{y}) = \alpha(\boldsymbol{x}, \boldsymbol{y})q(\boldsymbol{y} \mid \boldsymbol{x}) + \{1 - r(\boldsymbol{x})\}\mathsf{I}(\boldsymbol{y} = \boldsymbol{x})$$

where $r(\boldsymbol{x}) = \int \alpha(\boldsymbol{x}, \boldsymbol{y})q(\boldsymbol{y} \mid \boldsymbol{x})\mathrm{d}\boldsymbol{y}$ is the average probability of acceptance of a move from $\boldsymbol{x}$, $\mathsf{I}(\cdot = \boldsymbol{x})$ is a point mass at $\boldsymbol{x}$, and $\alpha(\cdot)$ is defined on the next slide.

One can show that the Metropolis–Hastings algorithm satisfies detailed balanced; see, e.g., Theorem 7.2 of Robert and Casella (2004).

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. It also ensures that, provided the starting value has 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 5.1.

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 5.1. 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.

```
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
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
```

```
# 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 5.2 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

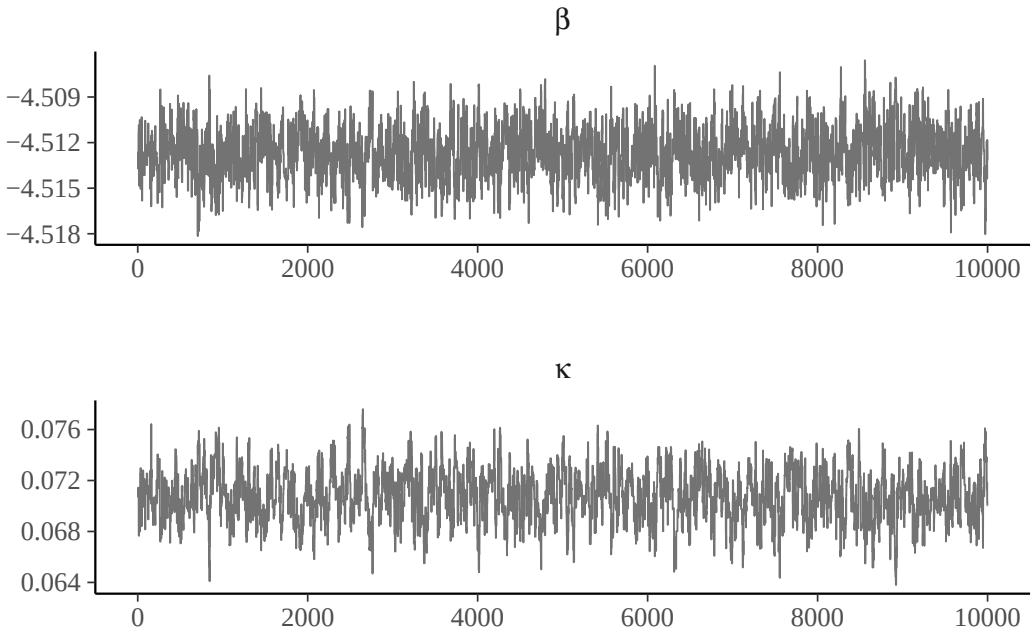

Figure 5.1: Traceplots of Markov chain of log rate and log odds rate for the Metropolis–Hastings sampler applied to the Upworthy question data.

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.

Remark 5.1 (Reparametrization). A better parametrization would simply sample two parameters with $\lambda_2 = \exp(\alpha)$, where α is the log mean of the second group, with the same prior as for β. Since the likelihood factorizes and the parameters are independent apriori, this would lead to zero correlation and lead to more efficient mixing of the Markov chain, should we wish to sample parameters in turn one at the time. A Markov chain for κ can then be obtained by substracting the values of $\alpha - \beta$ from the new draws.

The quality of the mixing of the chain (autocorrelation), depends on the proposal variance, which can obtain by trial and error. Trace plots Figure 5.1 show the values of the chain as 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

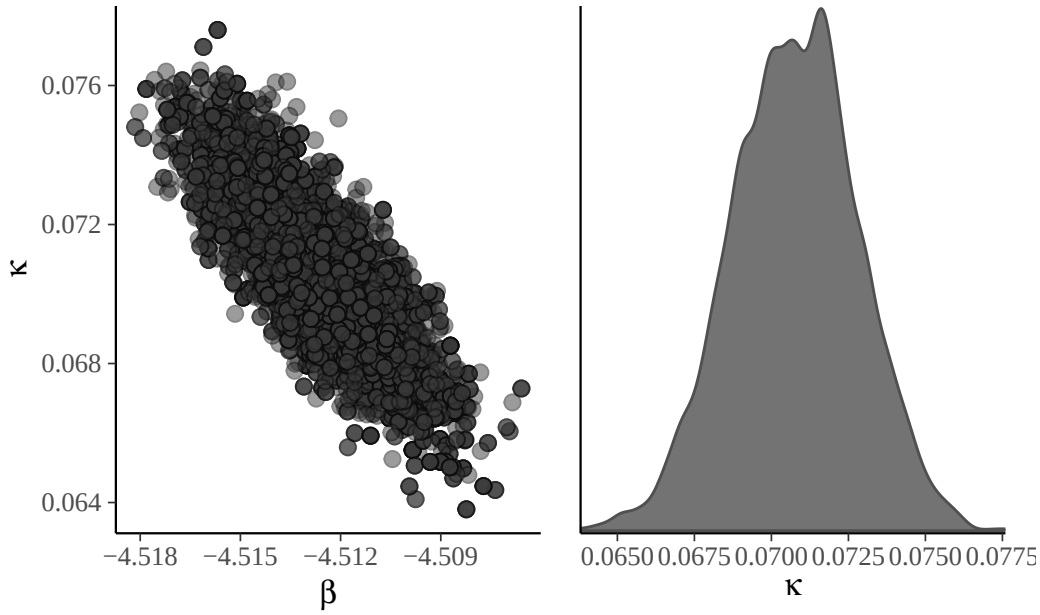

Figure 5.2: Scatterplot of posterior draws (left) and marginal density plot of log odds rate (right).

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 5.3.

One way to calibrate is to track the acceptance rate of the proposals: for the three chains in Figure 5.3, 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. 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

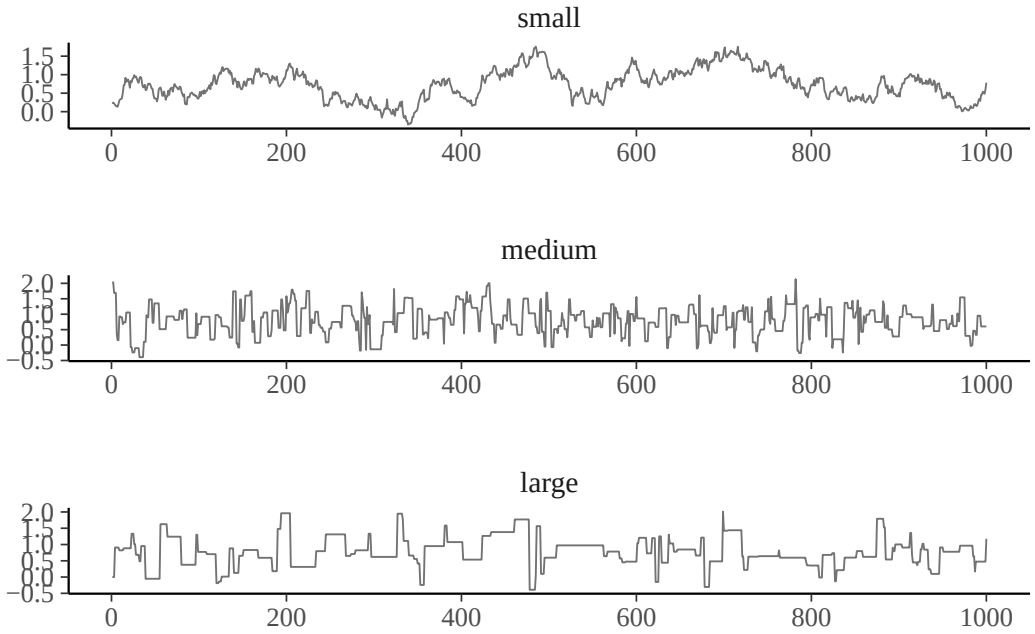

Figure 5.3: Example of traceplot with proposal variance that is too small (top), adequate (middle) and too large (bottom).

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 5.4 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 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

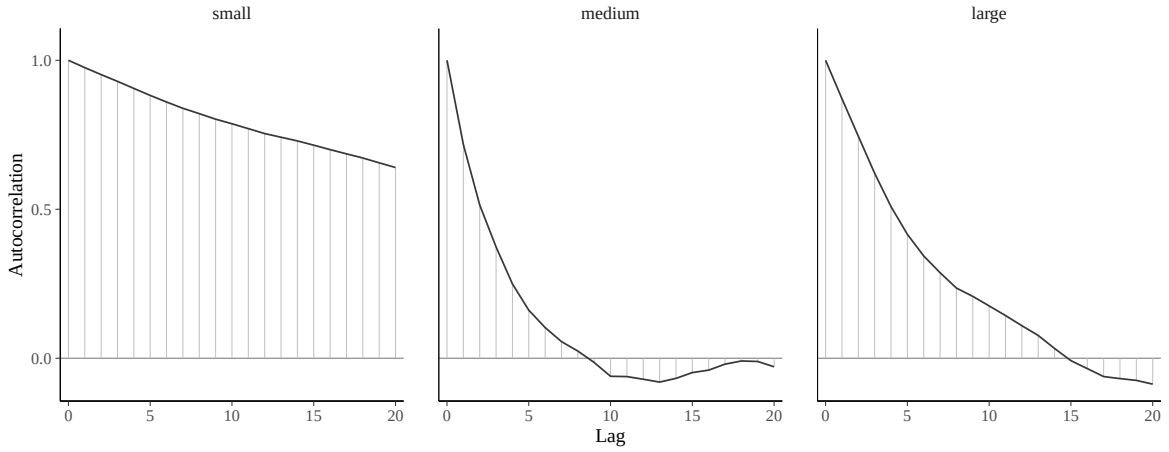

Figure 5.4: Correlogram for the three Markov chains.

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 5.3 (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 5.4 (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 5.5 (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 (H. Rue and Held 2005).

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 5.2. 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 $\sigma \sim \mathsf{Exp}(0.7)$.

```
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.

Summary:

- Metropolis–Hastings generalizes rejection sampling by building a Markov chain and providing a mechanism for sampling.
- Small proposal variance leads to high acceptance rate, but small step sizes. Large variance proposals leads to many rejections, in which case the previous value is

carried forward. Both extreme scenarios lead to large autocorrelation.

- The proposal density can be anything, but must ideally account for the support and allow for exploration of the state.
- Good initial starting values can be obtained by computing maximum a posteriori estimates.
- Initializing multiple chains at different starting values can be used to check convergence to the stationary distribution.
- Mixing will improve if strongly correlated parameters are sampled together.
- The optimal acceptance rate depends on the dimension, but guidelines for random walk Metropolis are to have 0.44 for a single parameter model and 0.234 for multivariate targets; see Neal (2011) for a heuristic derivation.
- To obtain the target acceptance rate, users must tune the variance of the proposal kernel. This is typically achieved by running the chain for some period, computing the empirical acceptance rate and increasing (respectively decreasing) the variance if the acceptance rate is too high (too low).
- Metropolis-adjusted Langevin algorithm (MALA) uses the gradient information to inform the proposal; it is akin to a Newton step.
- The detailed balance requires a function g such that $g(r) = rg(1/r)$. Taking $g(r) = \min(1, r)$ as in Metropolis–Hasting rule leads to the lowest asymptotic variance (Peskun 1973).
-

6 Gibbs sampling

! Learning objectives:

At the end of the chapter, students should be able to

- implement Gibbs sampling.
- derive the conditional distributions of a model for Gibbs sampling.
- use data augmentation to emulate 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

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

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 6.1 shows 25 steps from a Gibbs algorithm for a bivariate target.

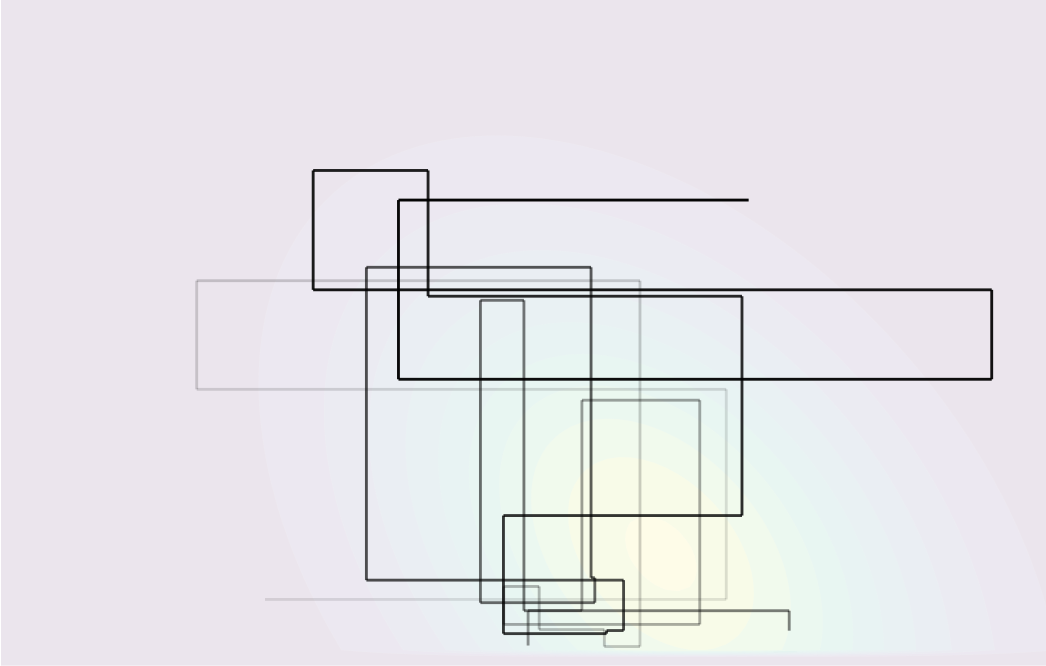

Figure 6.1: Sampling trajectory for a bivariate target using Gibbs sampling.

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

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

```
# 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 6.2 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.

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 6.1. 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 $\tau \sim \mathsf{inv.gamma}(\alpha, \beta)$.

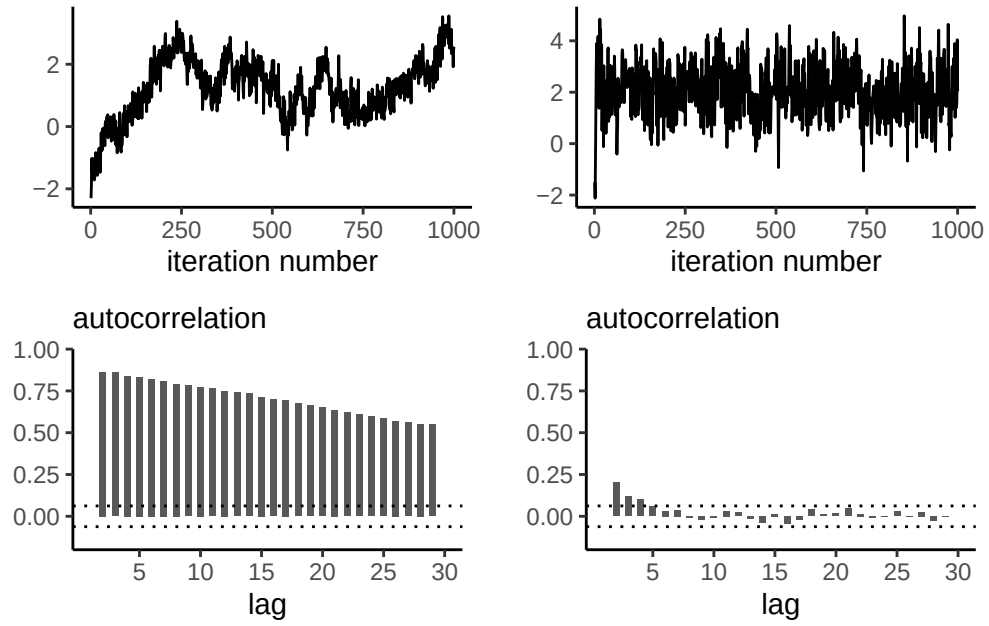

Figure 6.2: Trace plots (top) and correlograms (bottom) for the first component of a Gibbs sampler with $d = 20$ equicorrelated Gaussian variates with correlation $\rho = 0.9$ (left) and $d = 3$ with equicorrelation $\rho = 0.5$ (right).

The 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

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

gives us

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

so we can simulate in turn

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

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.

6.1 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 6.2 (Probit regression)**.** 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 $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{z}$ is the ordinary least square estimator from the regression with model matrix $\mathbf{X}$ and response vector $\boldsymbol{z}$. The augmented variables Z_i are conditionally independent and truncated Gaussian with

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

and we can use the algorithms of Example 4.2 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 6.3 (Bayesian LASSO)**.** The Laplace distribution with location μ and scale σ, has density

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

It can be expressed as a scale mixture of Gaussians, where $Y_i \sim \mathsf{Laplace}(\mu,\sigma)$ is equivalent to $Z_i \mid \tau \sim \mathsf{Gauss}(\mu,\lambda_i)$ and $\Lambda_i \sim \mathsf{expo}\{(2\sigma^2)^{-1}\}$. To see this, we first look at the Wald (or inverse Gaussian) distribution $\mathsf{Wald}(\nu,\omega)$ with location $\nu > 0$ and shape $\omega > 0$,, whose density is

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

To show that the marginal (unconditionally) is Laplace, we write the joint density and integrate out the variance term λ, make the change of variable to get the result in terms of the precision $\xi = 1/\lambda$, whence

$$\begin{aligned} p(z) &= \int_0^\infty p(z \mid \lambda)p(\lambda)\mathrm{d}\lambda \\ &= \int_0^\infty \frac{1}{(2\pi\lambda)^{1/2}}\exp\left\{-\frac{1}{2\lambda}(z-\mu)^2\right\}\frac{1}{2\sigma^2}\exp\left(-\frac{\lambda}{2\sigma^2}\right)\mathrm{d}\lambda \\ &= \frac{1}{2\sigma^2}\int_0^\infty \frac{1}{(2\pi\lambda)^{1/2}}\exp\left[-\frac{1}{2}\left\{\frac{(z-\mu)^2}{\lambda}+\frac{\lambda}{\sigma^2}\right\}\right]\mathrm{d}\lambda \\ &= \frac{1}{2\sigma^2}\int_0^\infty \frac{1}{\xi^2}\frac{\xi^{1/2}}{(2\pi)^{1/2}}\exp\left[-\frac{1}{2\sigma^2}\left\{\xi\sigma^2(z-\mu)^2+\frac{1}{\xi}\right\}\right]\mathrm{d}\xi \\ &= \frac{1}{2\sigma^2}\int_0^\infty \frac{1}{(2\pi\xi^3)^{1/2}}\exp\left[-\frac{\omega}{2}\left\{\frac{\xi}{\nu^2}+\frac{1}{\xi}\right\}\right]\mathrm{d}\xi \\ &= \frac{1}{2\sigma^2\omega^{1/2}}\exp\left(-\frac{\omega}{\nu}\right) \\ &= \frac{1}{2\sigma}\exp\left(-\frac{|z-\mu|}{\sigma}\right). \end{aligned}$$

upon recovering the conditional density of $\Xi \mid Z \sim \mathsf{Wald}(\nu, \omega)$ with parameters $\nu = (\sigma|z-\mu|)^{-1}$ and $\omega = \sigma^{-2}$.

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

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

With the improper prior $p(\mu, \sigma^2) \propto 1/\sigma^2$ and with n independent and identically distributed Laplace variates, written as a scale mixture, the model is amenable to Gibbs sampling. With $\mathbf{D}_\tau^{-1} = \mathrm{diag}(\tau_1^{-1}, \ldots, \tau_p^{-1})$ and $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \overline{y}\mathbf{1}_n$ the centered response vector, we can simulate in turn (Park and Casella 2008)

$$\begin{aligned}\mu \mid \sigma^2, \boldsymbol{y} &\sim \mathsf{Gauss}(\overline{y}, \sigma^2/n)\\ \boldsymbol{\beta} \mid \sigma^2, \boldsymbol{\tau}, \boldsymbol{y} &\sim \mathsf{Gauss}_p\left\{\left(\mathbf{X}^\top\mathbf{X} + \mathbf{D}_\tau^{-1}\right)^{-1}\mathbf{X}\widetilde{\boldsymbol{y}}, \sigma^2\left(\mathbf{X}^\top\mathbf{X} + \mathbf{D}_\tau^{-1}\right)^{-1}\right\}\\ \sigma^2 \mid \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{y} &\sim \mathsf{inv.gamma}\left\{\frac{n-1+p}{2}, \frac{(\widetilde{\boldsymbol{y}} - \mathbf{X}\boldsymbol{\beta})^\top(\widetilde{\boldsymbol{y}} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}^\top\mathbf{D}_\tau^{-1}\boldsymbol{\beta}}{2}\right\},\\ \tau_j^{-1} \mid \boldsymbol{\beta}, \sigma^2 &\sim \mathsf{Wald}\left(\frac{\lambda^{1/2}\sigma}{|\beta_j|}, \lambda\right)\end{aligned}$$

where the last three conditional distributions follow from marginalizing out μ.

The Bayesian LASSO places a Laplace penalty on the regression coefficients, with lower values of λ yielding more shrinkage. Figure 6.3 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 $\boldsymbol{\beta}$ are exactly zero.

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 6.4 (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 the relative weights of the components, and their location and scale.

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

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

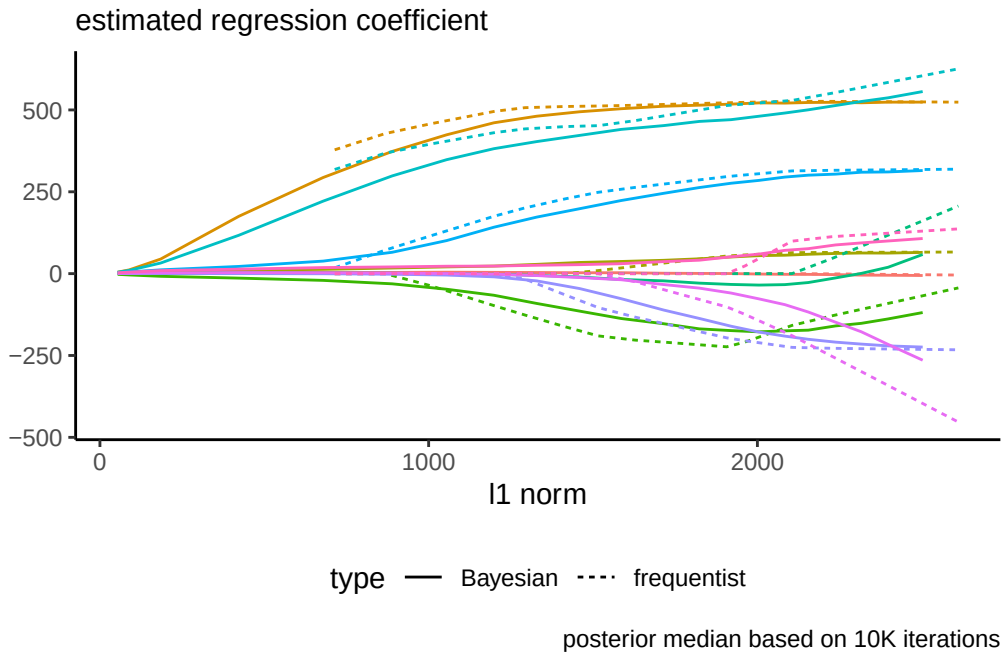

Figure 6.3: Traceplot of β coefficients (penalized maximum likelihood estimates and median aposteriori as a function of the l_1 norm of the coefficients, with lower values of the latter corresponding to higher values of the penalty λ.

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 likelihood becomes

$$\prod_{i=1}^{n}\prod_{k=1}^{K}\{\omega_k f_k(x;\boldsymbol{\theta}_k)\}^{Z_i},$$

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

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

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, assigning a conjugate Dirichlet prior for $\boldsymbol{\omega}$.

Example 6.5 (Mixture model for geyser)**.** We consider a Gaussian mixture model for waiting time between two eruptions of the Old Faithful geyser in Yellowstone. The distribution is of the form

$$f_i(x) = p_i\phi_1(x_i; \mu_1, \tau_1^{-1}) + (1-p_i)\phi_2(x_i; \mu_2, \tau_2^{-1}).$$

where $\phi(\cdot; \mu, \tau^{-1})$ is the density function of a Gaussian with mean μ and precision τ. We assign conjugate priors with $p_i \sim \mathsf{beta}(a_1, a_2)$, $\mu_j \sim \mathsf{Gauss}(c, d^{-1})$ and $\tau_j \sim \mathsf{gamma}(b_1, b_2)$. For the hyperpriors, we use $a_1 = a_2 = 1, b_1 = 1, b_2 = 0.1, c = 60$, and $d = 1/40$.

```
data(faithful)
n <- nrow(faithful)
y <- faithful$waiting
# Fix hyperpriors
a1 <- 2; a2 <- 2; c <- 60; d <- 1/40; b1 <- 1; b2 <- 0.01
# Assign observations at random to groups
set.seed(80601)
cut <- runif(1, 0.1, 0.9)*diff(range(y)) + min(y)
group <- as.integer(y > cut)
p <- sum(group == 0L)/n
mu <- c(mean(y[group == 0]), mean(y[group == 1]))
prec <- 1/c(var(y[group == 0]), var(y[group == 1]))
# Storage and number of replications
B <- 1e4L
theta <- matrix(nrow = B, ncol = 5L)
# Step 1: assign variables to clusters
for(b in 1:B){
  d1 <- dnorm(y, mean = mu[1], sd = 1/sqrt(prec[1])) # group 0
  d2 <- dnorm(y, mean = mu[2], sd = 1/sqrt(prec[2])) # group 1
  # Data augmentation: group labels
  group <- rbinom(n = n, size = rep(1, n), prob = (1-p)*d2/(p*d1 + (1-p)*d2))
  # Step 2: update probability of cluster
  p <- rbeta(n = 1, shape1 = n - sum(group) + a1, sum(group) + a2)
  for(j in 1:2){
    yg <- y[group == (j-1L)]
    ng <- length(yg)
    prec_mu <- prec[j] * ng + d
    mean_mu <- (sum(yg)*prec[j] + c*d)/prec_mu
    mu[j] <- rnorm(n = 1, mean = mean_mu, sd = 1/sqrt(prec_mu))
    prec[j] <- rgamma(n = 1,
```

```
                    shape = b1 + ng/2,
                    rate = b2 + 0.5*sum((yg-mu[j])^2))
  }
  theta[b, ] <- c(p, mu, prec)
}
# Discard initial observations (burn in)
theta <- theta[-(1:100),]
```

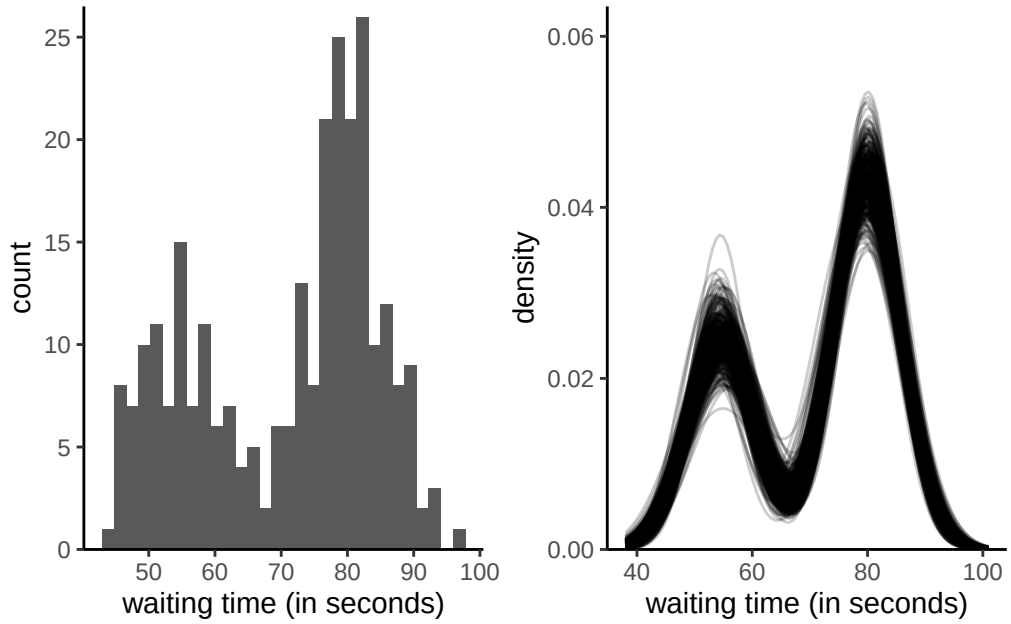

Figure 6.4: One-dimensional density mixture for the Old Faithful data, with histogram of data (left) and posterior density draws (right).

Remark 6.1 (Label switching in mixture models). If we run a MCMC algorithm to sample from a mixture models, the likelihood is invariant to permutation of the group labels, leading to identifiability issues when the chain swaps modes, when running multiple Markov chains with symmetric priors or using tempering algorithms. Two chains may thus reach the same stationary distribution, with group labels swapped. It is sometimes necessary to impose ordering constraints on the mean parameters $\boldsymbol{\mu}$, although this isn't necessarily easy to generalize beyond the univariate setting. See Jasra, Holmes, and Stephens (2005) and Stephens (2002) for more details.

Summary:

- Gibbs sampling is a special case of Metropolis–Hastings algorithm, where we sample from the conditional distributions given other parameters.
- Use of (conditionally) conjugate priors enables Gibbs sampling.
- The fact that any Gibbs step is accepted with probability one does not mean the sampler is efficient: there can be significant autocorrelation in the chains.
- We can sometimes update parameters jointly, or reduce the dependence by integrating out some of the conditioning variables (marginalization).
- We can use Gibbs step for some updates within a more general algorithm.
- Even if there is no closed-form expression, we can use Monte Carlo methods to simulate parameters in a Gibbs sampler.
- In many scenarios, the likelihood is costly to evaluate or not amenable to Gibbs sampling. Data augmentation introduces additional parameters to the model in exchange for simplifying the likelihood.
- Data augmentation leads to a trade-off between complexity and efficiency (more parameters, slower mixing).
- Data augmentation is commonly used for expectation-maximisation (EM) algorithm for maximum likelihood estimation in frequentist setting.
- Special classes of models (Bayesian linear regression, mixtures, etc.) are typically fitted using Gibbs sampling.
- Probabilistic programming languages (Bugs, JAGS) rely on Gibbs sampling.

7 Computational strategies and diagnostics

The Bayesian workflow is a coherent framework for model formulation construction, inference and validation. It typically involves trying and comparing different models, adapting and modifying these models (Gelman et al. 2020); see also Michael Betancourt for excellent visualizations. In this chapter, we focus on three aspects of the workflow: model validation, evaluation and comparison.

! Learning objectives:

At the end of the chapter, students should be able to

- use output of MCMC to obtain estimates and standard errors.
- choose suitable test statistics to evaluate model adequacy.
- assess convergence using graphical tools and effective sample size.
- perform model comparisons using Bayes factor or predictive measures.
- diagnose performance of MCMC algorithms (in terms of mixing and effective sample size).
- implement strategies to improve sampling performance, including block updates, reparametrization and marginalization.

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 per iteration and the efficiency of the samplers, with algorithms providing realizations of Markov chains with lower autocorrelation being preferred. 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, and to go over tricks that can improve mixing.

Generating artificial data: Some problems and checks relate to models and the correct implementations (of the algorithms). Sometimes, the probabilistic procedure will generate draws, but it's unclear whether our numerical implementation is correct. We can sometimes see this if the output is truly misleading, but it's not always obvious. We can for example generate an "artificial" or fake data set from the model with some fixed parameter inputs to see if we can recover the parameter values used to generate these within some credible set.

Many such sanity checks can be implemented by means of simulations. Consider prior predictive checks: if the prior has a distribution from which we can generate, we can obtain prior draws from $p(\boldsymbol{\theta})$, generate data from the prior predictive $p(y \mid \boldsymbol{\theta})$ by simulating new observations from the data generating mechanism of the likelihood, and use these to obtain prior predictive by removing the likelihood component altogether: the draws from the prior predictive should then match posterior draws with only the prior.

The "data-averaged posterior" is obtained upon noting that (Geweke 2004)

$$p(\boldsymbol{\theta}) = \int \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta} \mid \boldsymbol{y}) p(\boldsymbol{y} \mid \widetilde{\boldsymbol{\theta}}) p(\widetilde{\boldsymbol{\theta}}) \mathrm{d}\widetilde{\boldsymbol{\theta}} \mathrm{d}\boldsymbol{y}$$

by forward sampling first the prior, than data for this particular value and obtaining the posterior associated with the latter.

We can test that our sampling algorithm correctly samples from the posterior distribution of interest by running the following procedure, which is however computationally intensive.

Proposition 7.1 (Simulation based calibration)**.** *Simulation-based calibration (Talts et al. 2020) proceeds with, in order*

1. $\boldsymbol{\theta}_0 \sim p(\boldsymbol{\theta})$,
2. $\boldsymbol{y}_0 \sim p(\boldsymbol{y} \mid \boldsymbol{\theta}_0)$,
3. $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_B \sim p(\boldsymbol{\theta} \mid \boldsymbol{y}_0)$.

Conditional on the simulated $\boldsymbol{y}$, the distribution of $\boldsymbol{\theta}_0$ is the same as that of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_B$. We do a dimension reduction step taking the test function $t(\cdot)$ to get the rank of the prior draw among the posterior ones, breaking ties at random if any. In the absence of ties,

$$T = \sum_{b=1}^{B} \mathrm{I}\{t(\boldsymbol{\theta}_b, \boldsymbol{y}) < t(\boldsymbol{\theta}_0, \boldsymbol{y})\},$$

These steps are repeated K times, yielding K test functions $T_1, \ldots, T_K$. We then test for uniformity using results from Säilynoja, Bürkner, and Vehtari (2022).

7.1 Convergence diagnostics and model validation

Many diagnostics rely on running multiple Markov chains for the same problem, with different starting values. In practice, it is more efficient to run a single long chain than multiple chains, because of the additional computational overhead related to burn in and warmup period. Running multiple chains however has the benefit of allowing one to compute diagnostics of convergence (by comparing chains) such as $\widehat{R}$, and to detect local modes.

Definition 7.1 (Trace plots)**.** A trace plot is a line plot of the Markov chain as a function of the number of iterations. It should be stable around some values if the posterior is unimodal and the chain has reached stationarity. The ideal shape is that of a 'fat hairy catterpilar'.

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 7.1 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.

Definition 7.2 (Trace rank plot)**.** If we run several chains, as in Figure 7.1, with different starting values, we can monitor convergence by checking whether these chains converge to the same target. A **trace rank** plot 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.

A trace rank plot is shown on right panel of Figure 7.1.

Definition 7.3 (Burn in period)**.** We term "burn in" the initial steps of the MCMC algorithm that are discarded because the chain has not reached it's stationary distribution, due to poor starting values. , but visual inspection using a trace plot may show that it is necessary to remove additional observations.

Most software will remove the first N initial values (typically one thousand). Good starting values can reduce the need for a long burn in period. 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. Geweke (1992)'s test measure whether the distribution of the resulting Markov chain is the same at the beginning and at the end through a test of equality of means.

Definition 7.4 (Warmup)**.** Warmup period refers to the initial sampling phase (potentially overlapping with burn in period) during which proposals are tuned (for example, by changing the variance proposal to ensure good acceptance rate or for Hamiltonian Monte Carlo (HMC) to tune the size of the leapfrog. These initial steps should be disregarded.

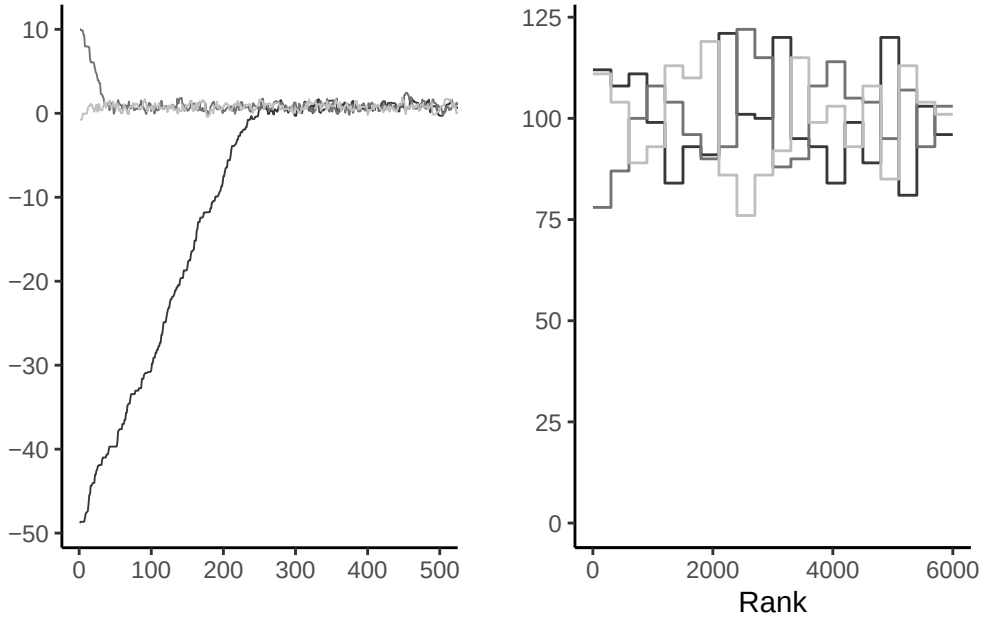

Figure 7.1: Traceplots of three Markov chains for the same target with different initial values for the first 500 iterations (left) and trace rank plot after discarding these (right).

The target of inference is a 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).

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 $\beta_j \sim \mathsf{Gauss}(0, 100)$. This model has too little variability relative to the observations and fits poorly as is.

For the Poisson example, the effective sample size for the $\boldsymbol{\beta}$ for the multilevel model is a bit higher than 1000 with $B = 5000$ iterations, whereas we have for the simple naive model is 1.028×10^4 for $B = 10000$ draws, suggesting superefficient sampling. The dependency between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is responsible for the drop in accuracy.

The `coda` (convergence diagnosis and output analysis) **R** package (Plummer et al. 2006) 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.

Proposition 7.2 (Gelman–Rubin diagnostic)**.** *The Gelman–Rubin diagnostic $\widehat{R}$, introduced in Gelman and Rubin (1992) and 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 $\widehat{\sigma}_m^2$ ($m = 1, \ldots, M$). The between-chain variance and within-chain variance estimator are*

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

$$\mathsf{Va}_{within} = \frac{1}{M}\sum_{m=1}^{M}\widehat{\sigma}_m^2$$

and we can compute

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

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 7.2 shows two instances where the chains are visually very far from having the same average and this is reflected by the large values of $\widehat{R}$.

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.

Definition 7.5 (Thinning)**.** 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.

7.1.1 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

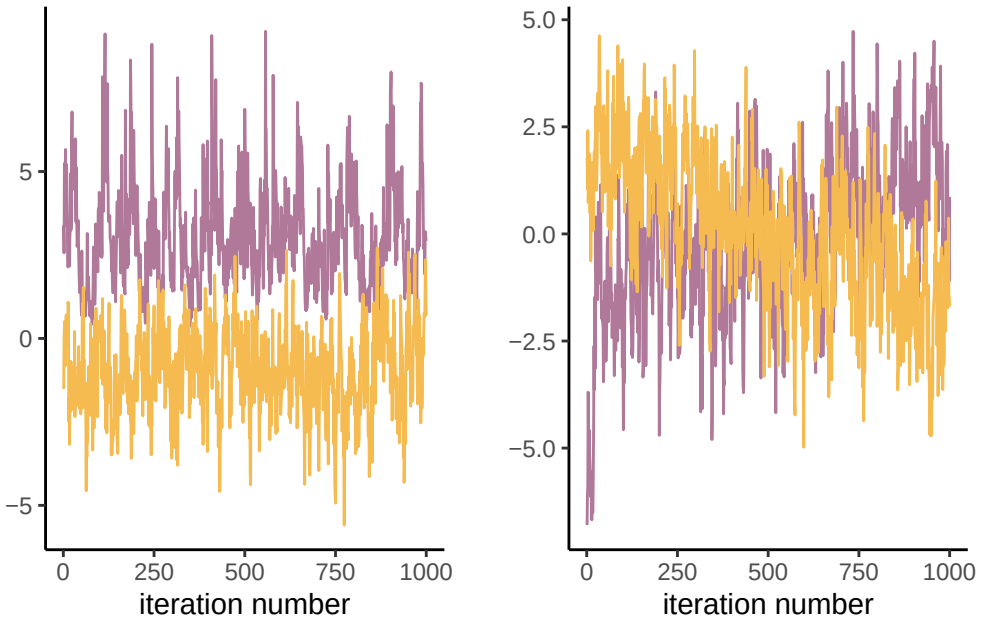

Figure 7.2: Two pairs of Markov chains: the top ones seem stationary, but with different modes. This makes the between chain variance substantial, with a value of $\widehat{R} \approx 3.4$, whereas the chains on the right hover around the same values of zero, but do not appear stable with $\widehat{R} \approx 1.6$.

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. These should be salient features of the data, and may reveal inadequate likelihood or prior information.

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 7.3 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.

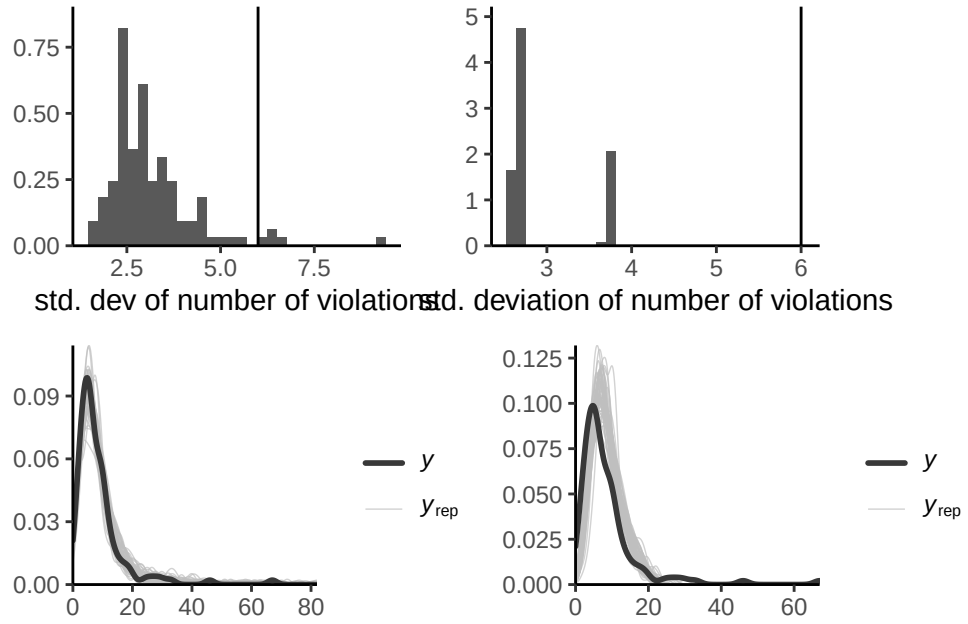

Figure 7.3: Posterior predictive checks for the standard deviation (top) and density of posterior draws (bottom) for hierarchical Poisson model with individual effects (left) and simpler model with only conditions (right).

7.2 Information criteria

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 $2n \times$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.

An older criterion which has somewhat fallen out of fashion is the **deviance** information criterion of Spiegelhalter et al. (2002). It is defined as

$$\mathsf{DIC} = -2\ell(\widetilde{\boldsymbol{\theta}}) + 2p_D$$

where p_D is the posterior expectation of the deviance relative to the point estimator of the parameter $\widetilde{\boldsymbol{\theta}}$ (e.g., the maximum a posteriori or the posterior mean)

$$p_D = \mathsf{E}\{D(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) \mid \boldsymbol{y}\} = \int 2\{\ell(\widetilde{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta})\} f(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}$$

The DIC can be easily evaluated by keeping track of the log likelihood evaluated at each posterior draw from a Markov chain Monte Carlo algorithm. The penalty term p_D is however not invariant to reparametrizations. Assuming we can derive a multivariate Gaussian approximation to the MLE under suitable regularity conditions, the DIC is equivalent in large samples to AIC. The DIC is considered by many authors as not being a Bayesian procedure; see Spiegelhalter et al. (2014) and the discussion therein.

Criteria such as LPPD and therefore WAIC require some form of exchangeability, and don't apply to cases where leave-one-out cross validation isn't adequate, for example in spatio-temporal models.

Example 7.1 (Information criteria for smartwatch and Bayesian LASSO)**.** For the smartwatch model, we get a value of 3.07 for the complex model and 4.5: this suggests an improvement in using individual-specific effects.

```
#' WAIC
#' @param loglik_pt B by n matrix of pointwise log likelihood
WAIC <- function(loglik_pt){
  -mean(apply(loglik_pt, 2, mean)) +  mean(apply(loglik_pt, 2, var))
}
```

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 7.4 shows that the decrease in predictive performance is severe when estimates are shrunk towards 0, but the model performs equally well for small penalties.

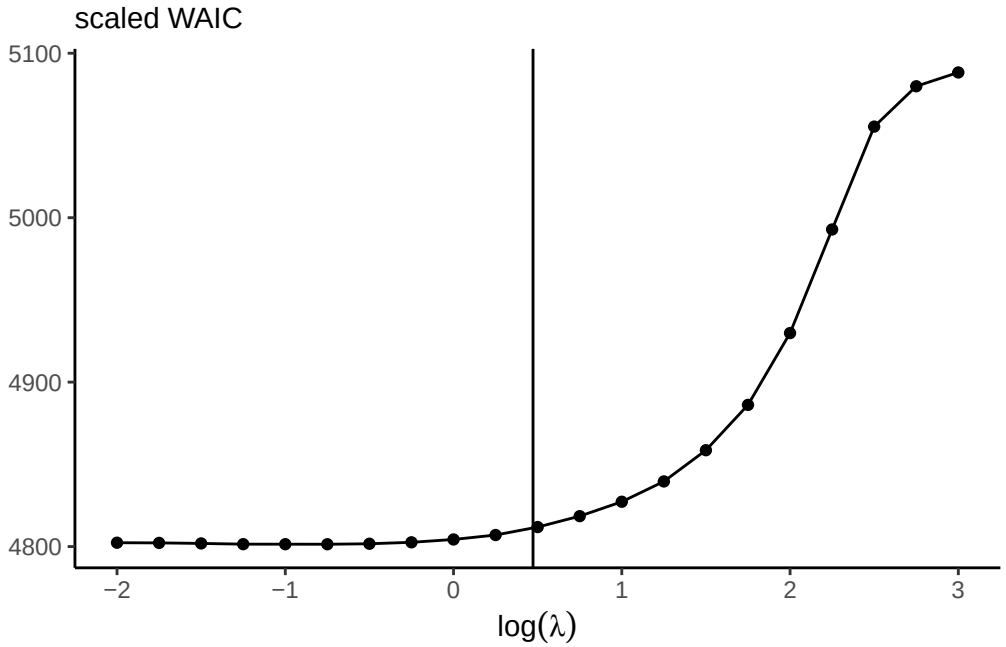

Figure 7.4: Widely applicable information criterion for the Bayesian LASSO problem fitted to the diabetes data, as a function of the penalty λ.

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 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.

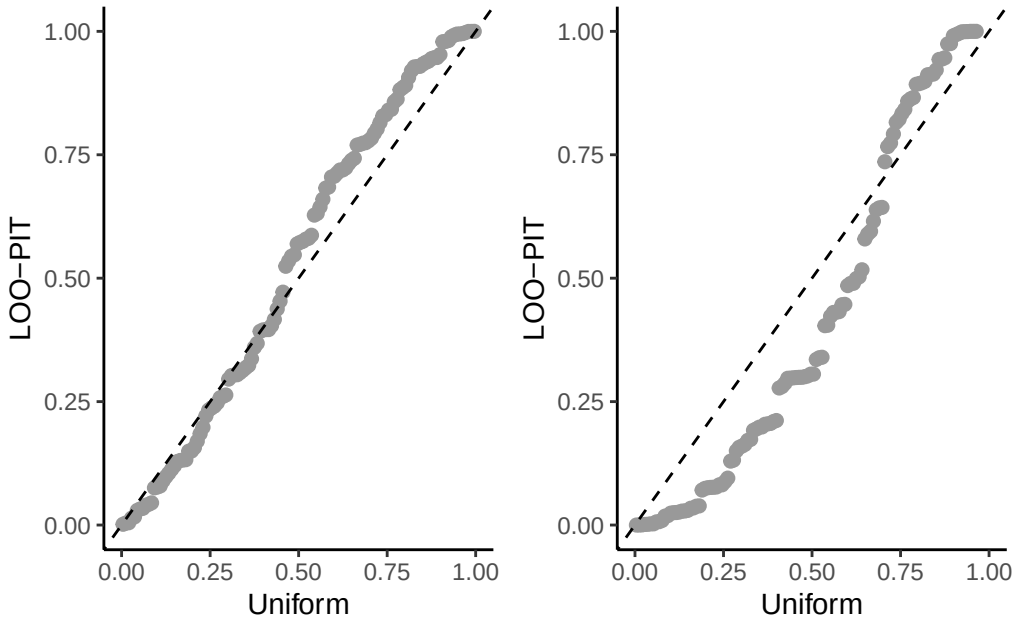

Figure 7.5: Quantile-quantile plots based on leave-one-out cross validation for model for the Poisson hierarchical model with the individual random effects (left) and without (right).

7.3 Computational strategies

The data augmentation strategies considered in Section 6.1 helps to simplify the likelihood and thereby reduce the cost of each iteration. However, latent variables are imputed conditional on current parameter values $\boldsymbol{\theta}_a$: the higher the number of variables, the more the model will concentrate around current values of $\boldsymbol{\theta}_a$, which leads to slow mixing.

There are two main strategies to deal with this problem: blocking the random effects together and simulating them jointly to improve mixing, and marginalizing out over some of the latent variables.

Example 7.2 (Marginalization in Gaussian models)**.** To illustrate this fact, consider a hierarchical Gaussian model of the form

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{B} + \boldsymbol{\varepsilon}$$

where $\mathbf{X}$ is an $n \times p$ design matrix with centered inputs, $\boldsymbol{\beta} \sim \mathsf{Gauss}(\mathbf{0}_p, \sigma^2\mathbf{I}_p)$, $\boldsymbol{B} \sim \mathsf{Gauss}_q(\mathbf{0}_q, \boldsymbol{\Omega})$ are random effects and $\boldsymbol{\varepsilon} \sim \mathsf{Gauss}_n(\mathbf{0}_n, \kappa^2\mathbf{I}_n)$ are independent white noise.

We can write

$$\begin{aligned}\boldsymbol{Y} \mid \beta, \boldsymbol{B} &\sim \mathsf{Gauss}_n(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{B}, \sigma^2\mathbf{I}_p)\\ \boldsymbol{Y} \mid \beta &\sim \mathsf{Gauss}_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{Q}^{-1}),\end{aligned}$$

where the second line corresponds to marginalizing out the random effects $\boldsymbol{B}$. This reduces the number of parameters to draw, but the likelihood evaluation is more costly due to $\mathbf{Q}^{-1}$. If, as is often the case, $\boldsymbol{\Omega}^{-1}$ and $\mathbf{Z}$ are sparse matrices, the full precision matrix can be efficiently computed using Shermann–Morisson–Woodbury identity as

$$\begin{aligned}\mathbf{Q}^{-1} &= \mathbf{Z}\boldsymbol{\Omega}^{-1}\mathbf{Z}^\top + \kappa^2\mathbf{I}_n,\\ \kappa^2\mathbf{Q} &= \mathbf{I}_n - \mathbf{Z}\boldsymbol{G}^{-1}\mathbf{Z}^\top,\\ \boldsymbol{G} &= \mathbf{Z}^\top\mathbf{Z} + \kappa^2\boldsymbol{\Omega}^{-1}\end{aligned}$$

Section 3.1 of Nychka et al. (2015) details efficient ways of calculating the quadratic form involving $\mathbf{Q}$ and it's determinant.

Proposition 7.3 (Pseudo marginal)**.** *Another option proposed by Andrieu and Roberts (2009) based on an original idea from Beaumont (2003) relies on pseudo marginalization, where integration is done via Monte Carlo sampling. Specifically, suppose that we are ultimately interested in*

$$p(\boldsymbol{\theta}) = \int p(\boldsymbol{\theta}, \boldsymbol{z})\mathrm{d}\boldsymbol{z},$$

but that for this purpose we normally sample from both parameters. Given a proposal $\boldsymbol{\theta}$ and $q_1(\boldsymbol{\theta})$ and subsequently L draws once from $q_2(\boldsymbol{z} \mid \boldsymbol{\theta})$ for the nuisance, we can approximate the marginal using, e.g., importance sampling as

$$\widehat{p}(\boldsymbol{\theta}; \boldsymbol{z}) = \frac{1}{L}\sum_{l=1}^{L}\frac{p(\boldsymbol{\theta}, \boldsymbol{z}_l)}{q_2(\boldsymbol{z}_l, \boldsymbol{\theta})}.$$

We then run a Markov chain on an augmented state space $\boldsymbol{\Theta} \times \mathcal{Z}^L$, with Metropolis–Hastings acceptance ratio of

$$\frac{\widehat{p}(\boldsymbol{\theta}^\star; \boldsymbol{z}^\star_{1,t}, \boldsymbol{z}^\star_{L,t})}{\widehat{p}(\boldsymbol{\theta}_t; \boldsymbol{z}_{1,t-1}, \ldots, \boldsymbol{z}_{L,t-1})}\frac{q_1(\boldsymbol{\theta}_{t-1} \mid \boldsymbol{\theta}^\star_t)}{q_1(\boldsymbol{\theta}^\star_t \mid \boldsymbol{\theta}_{t-1})}.$$

Note that the terms involving $\prod_{l=1}^{L} q_2(\boldsymbol{z}_l; \boldsymbol{\theta})$ do not appear because they cancel out, as they are also part of the augmented state space likelihood.

The remarkable feature of the pseudo marginal approach is that even if our average approximation $\widehat{p}$ to the marginal is noisy, the marginal posterior of this Markov chain is the same as the original target.

Compared to regular data augmentation, we must store the full vector $\boldsymbol{z}_1^\star, \ldots, \boldsymbol{z}_L^\star$ and perform L evaluations of the augmented likelihood. The values of $\boldsymbol{z}$, if accepted, are stored for the next evaluation of the ratio.

The idea of pseudo-marginal extends beyond the user case presented above, as as long as we have an unbiased non-negative estimator of the likelihood $\mathsf{E}\{\widehat{p}(\boldsymbol{\theta})\} = p(\boldsymbol{\theta})$, even when the likelihood itself is intractable. This is useful for models where we can approximate the likelihood by simulation, like for particle filters. Pseudo marginal MCMC algorithms are notorious for yielding sticky chains.

Proposition 7.4 (Blocking)**.** *When parameters of the vector $\boldsymbol{\theta}$ that we wish to sample are strongly correlated, it is advisable when possible to simulate them jointly. Because the unnormalized posterior is evaluated at each step conditional on all values, the Markov chain will be making incremental moves and mix slowly if we sample them one step at a time.*

Before showcasing the effect of blocking and joint updates, we present another example of data augmentation using Example 6.2.

Example 7.3 (Tokyo rainfall)**.** We consider data from Kitagawa (1987) that provide a binomial time series giving the number of days in years 1983 and 1984 (a leap year) in which there was more than 1mm of rain in Tokyo. These data and the model we consider are discussed in in section 4.3.4 of H. Rue and Held (2005). We thus have $T = 366$ days and $n_t \in \{1, 2\}$ $(t = 1, \ldots, T)$ the number of observations in day t and $y_t = \{0, \ldots, n_t\}$ the number of days with rain. The objective is to obtain a smoothed probability of rain. The underlying probit model considered takes $Y_t \mid n_t, p_t \sim \mathsf{binom}(n_t, p_t)$ and $p_t = \Phi(\beta_t)$.

We specify the random effects $\boldsymbol{\beta} \sim \mathsf{Gauss}_T(\mathbf{0}, \tau^{-1}\mathbf{Q}^{-1})$, where $\mathbf{Q}$ is a $T \times T$ precision matrix that encodes the local dependence. A circular random walk structure of order 2 is used to model the smooth curves by smoothing over neighbors, and enforces small second derivative. This is a suitable prior because it enforces no constraint on the mean structure. This amounts to specifying the process with

$$\begin{aligned}\Delta^2\beta_t &= (\beta_{t+1} - \beta_t) - (\beta_t - \beta_{t-1}) \\ &= -\beta_{t-1} + 2\beta_t - \beta_{t+1} \sim \mathsf{Gauss}(0, \tau^{-1}), \qquad t \in \mathbb{N} \mod 366.\end{aligned}$$

This yields an intrinsic Gaussian Markov random field with a circulant precision matrix $\tau\mathbf{Q} = \tau\mathbf{G}\mathbf{G}^\top$ of rank $T-1$, where

$$\mathbf{G} = \begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & -1 \\ -1 & 2 & -1 & 0 & \ddots & 0 \\ 0 & -1 & 2 & -1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ -1 & 0 & 0 & 0 & \cdots & 2 \end{pmatrix},$$

$$\mathbf{Q} = \begin{pmatrix} 6 & -4 & 1 & 0 & \cdots & 1 & -4 \\ -4 & 6 & -4 & 1 & \ddots & 0 & 1 \\ 1 & -4 & 6 & -4 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ -4 & 1 & 0 & 0 & \cdots & -4 & 6 \end{pmatrix}.$$

Because of the linear dependency, the determinant of $\mathbf{Q}$ is zero. The contribution from the latent mean parameters is multivariate Gaussian and we exploit for computations the sparsity of the precision matrix $\mathbf{Q}$. Figure 7.6 shows five draws from the prior model, which loops back between December 31st and January 1st, and is rather smooth.

We can perform data augmentation by imputing Gaussian variables, say $\{z_{t,i}\}$ following Example 6.2 from truncated Gaussian, where $z_{t,i} = \beta_t + \varepsilon_{t,i}$ and $\varepsilon_{t,i} \sim \mathsf{Gauss}(0,1)$ are independent standard Gaussian and

$$z_{t,i} \mid y_{t,i}, \beta_t \sim \begin{cases} \mathsf{trunc.Gauss}(\beta_t, 1, -\infty, 0) & y_{t,i} = 0 \\ \mathsf{trunc.Gauss}(\beta_t, 1, 0, \infty) & y_{t,i} = 1 \end{cases}$$

The posterior is proportional to

$$p(\boldsymbol{\beta} \mid \tau)p(\tau) \prod_{t=1}^{T} \prod_{i=1}^{n_t} p(y_{t,i} \mid z_{t,i}) p(z_{t,i} \mid \beta_t)$$

and once we have imputed the Gaussian latent vectors, we can work directly with the values of $z_t = \sum_{i=1}^{n_t} z_{i,t}$. The posterior then becomes

$$\begin{aligned} p(\boldsymbol{\beta}, \tau) \propto \tau^{(n-1)/2} \exp\left(-\frac{\tau}{2}\boldsymbol{\beta}^\top \mathbf{Q}\boldsymbol{\beta}\right) \tau^{a-1} \exp(-\tau b) \\ \times \exp\left\{-\frac{1}{2}(\boldsymbol{z}/\boldsymbol{n} - \boldsymbol{\beta})^\top \mathrm{diag}(\boldsymbol{n})(\boldsymbol{z}/\boldsymbol{n} - \boldsymbol{\beta})\right\} \end{aligned}$$

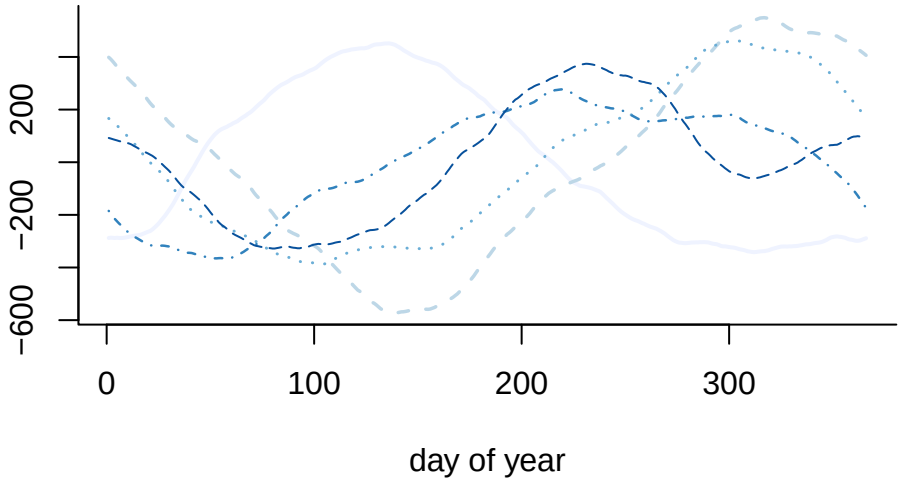

Figure 7.6: Five realizations from the cyclical random walk Gaussian prior of order 2.

where $\boldsymbol{z} = (z_1, \ldots, z_T)$. Completing the quadratic form shows that

$$\begin{aligned}\boldsymbol{\beta} \mid \boldsymbol{z}, \tau &\sim \mathsf{Gauss}_T\left[\{\tau\mathbf{Q} + \mathrm{diag}(\boldsymbol{n})\}^{-1}\boldsymbol{z}, \{\tau\mathbf{Q} + \mathrm{diag}(\boldsymbol{n})\}^{-1}\right] \\ \tau \mid \boldsymbol{\beta} &\sim \mathsf{gamma}\left(\frac{n-1}{2} + a, \frac{\boldsymbol{\beta}^\top\mathbf{Q}\boldsymbol{\beta}}{2} + b\right)\end{aligned}$$

```
library(Matrix)
library(TruncatedNormal)
data(tokyorain, package = "hecbayes")
# Aggregate data
tokyo <- tokyorain |>
   dplyr::group_by(day) |>
   dplyr::summarize(y = sum(y), n = dplyr::n())
nt <- 366L
# Circulant random walk of order two precision matrix
Q <- hecbayes::crw_Q(d = nt, type = "rw2", sparse = TRUE)
# Sparse Cholesky root
```

```
cholQ <- Matrix::chol(Q)
N <- Matrix::Diagonal(n = nt, x = tokyo$n)
# Create containers
B <- 1e4L # number of draws
beta_s <- matrix(nrow = B, ncol = nt)
x_s <- matrix(nrow = B, ncol = nt)
tau_s <- numeric(B)
# Initial values
beta <- rep(0, nt)
tau <- 1000
# Hyperprior parameter values
tau_a <- 1
tau_b <- 0.0001
# Gibbs sampling
for(b in seq_len(B)){
  # Step 1: data augmentation
  x <- TruncatedNormal::rtnorm(
    n = 1,  mu = beta[tokyorain$day], sd = 1,
    lb = ifelse(tokyorain$y == 0, -Inf, 0),
    ub = ifelse(tokyorain$y == 0, 0, Inf))
  tx <- aggregate(x = x, by = list(tokyorain$day), FUN = sum)$x
  x_s[b,] <- tx
  # Step 2: Simulate random effects in block
  beta <- beta_s[b,] <- c(hecbayes::rGaussQ(
    n = 1,
    b = tx,
    Q = tau * Q + N))
# Simulate precision
  tau <- tau_s[b] <- rgamma(
    n = 1,
    shape = (nt-1)/2 + tau_a,
    rate = 0.5*as.numeric(crossprod(cholQ %*% beta)) + tau_b)
  # if beta is VERY smooth, then precision is large
}
```

Example 7.4 (Blocking)**.** We revisit Example 7.3 with two modifications: imputing one parameter β_t at a time using random scan Gibbs step, which leads to slower mixing but univariate updates, and a joint update that first draws $\tau^\star$ from some proposal distribution,

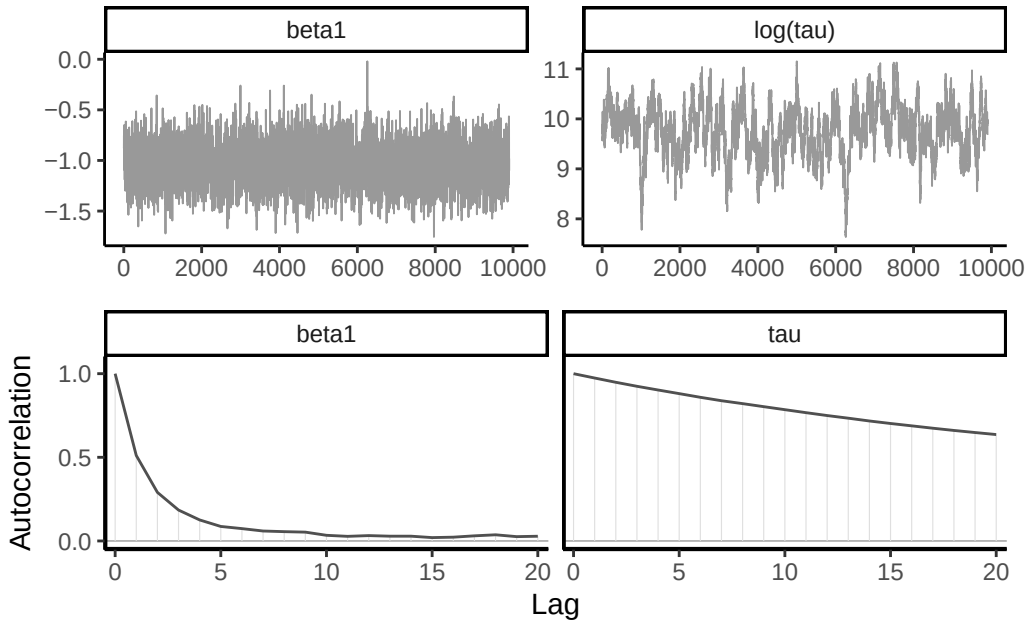

Figure 7.7: Trace plots (top) and correlograms (bottom) for two parameters of the Gibbs sampler for the Tokyo rainfall data, with block updates.

then sample conditional on that value generates the β vector and proposes acceptance using a Metropolis step.

A different (less efficient) strategy would be to simulate the β_t terms one at a time using a random scan Gibbs, i.e., picking $t_0 \in \{1, \ldots, 366\}$ and looping over indices. This yields higher autocorrelation between components than sampling by block.

```
# Compute mean vector for betas
mbeta <- Matrix::solve(a = tau*Q + N, b = tx)
# weights of precision for neighbours
nw <- c(1, -4, -4, 1)
# Sample an index at random
st <- sample.int(nt, 1)
for(i in (st + seq_len(nt)) %% nt + 1L){
 # Indices of the non-zero entries for row Q[i,]
nh <- c(i-3, i-2, i, i+1) %% 366 + 1
prec <- tau * 6 + tokyo$n[i]
```

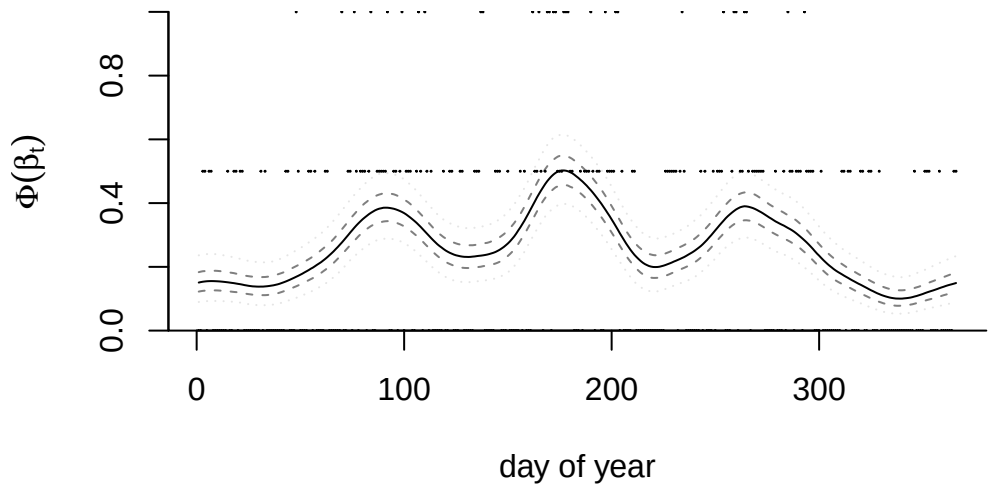

Figure 7.8: Tokyo rainfall fitted median probability with 50% and 89% pointwise credible intervals as a function of time of the year, with the proportion of days of rain (points).

```
condmean <- mbeta[i] - sum(nw*(beta[nh] - mbeta[nh])) * tau / prec
  beta[i] <- rnorm(n = 1, mean = condmean, sd = 1/sqrt(prec))
}
beta_s[b,] <- beta
```

Instead of making things worst, we can try to improve upon our initial sampler by simulating first a proposal $\tau^\star$ using a random walk Metropolis (on the log scale) or some other proposal $q(\tau^\star;\tau)$, then drawing from the full conditional $\boldsymbol{\beta} \mid \boldsymbol{z}, \tau^\star$ and accepting/rejecting the whole move. In doing this, all terms that depend on $\boldsymbol{\beta}$ cancel out, and the term $p(\tau^\star, \boldsymbol{\beta}^\star \mid \tau)/\{q(\tau^\star;\tau)p(\boldsymbol{\beta}^\star \mid \tau^\star)\}$ in the acceptance ratio becomes

$$\frac{\tau_t^{\star(n-1)/2} p(\tau_t^\star) \exp\left\{-\frac{1}{2}\boldsymbol{z}^\top(\boldsymbol{z}/\boldsymbol{n})\right\}}{q(\tau^\star;\tau)\,|\tau^\star\mathbf{Q} + \operatorname{diag}(\boldsymbol{n})| \exp\left[-\frac{1}{2}\boldsymbol{z}^\top\{\tau^\star\mathbf{Q} + \operatorname{diag}(\boldsymbol{n})\}^{-1}\boldsymbol{z}\right]}$$

A second alternative is to ditch altogether the data augmentation step and write the unnor-

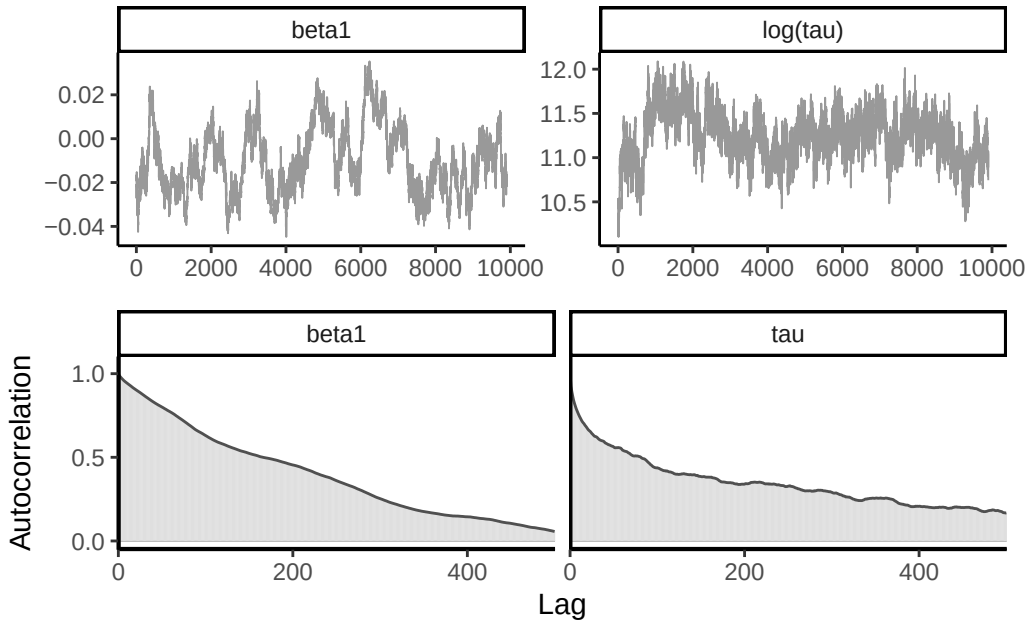

Figure 7.9: Trace plots (top) and correlograms (bottom) for two parameters of the Gibbs sampler for the Tokyo rainfall data, with individual updates for β_t.

malized log posterior for $\boldsymbol{\beta}$ as

$$\log p(\boldsymbol{\beta} \mid \boldsymbol{y}) \overset{\boldsymbol{\beta}}{\propto} -\frac{\tau}{2}\boldsymbol{\beta}^\top \mathbf{Q}\boldsymbol{\beta} + \sum_{t=1}^{366} y_t \log \Phi(\beta_t) + (n_t - y_t)\log\{1-\Phi(\beta_t)\}$$

and do a quadratic approximation to the posterior by doing a Taylor expansion of the terms $\log p(y_t \mid \beta_t)$ around the current value of the draw for $\boldsymbol{\beta}$. Given that observations are conditionally independent, we have a sum of independent terms $\ell(\boldsymbol{y}; \boldsymbol{\beta}) = \sum_{t=1}^{366} \log p(y_t \mid \beta_t)$ and this yields, expanding around $\boldsymbol{\beta}^0$, the Gaussian Markov field proposal

$$q(\boldsymbol{\beta} \mid \tau, \boldsymbol{\beta}^0) \sim \mathsf{Gauss}_{366}\left[\ell'(\boldsymbol{\beta}^0), \tau\mathbf{Q} + \mathrm{diag}\{\ell''(\boldsymbol{\beta}^0)\}\right].$$

Indeed, because of conditional independence, the jth element of ℓ' and ℓ'' are

$$\ell'(\boldsymbol{\beta}^0)_j = \left.\frac{\partial \ell(y_j;\beta_j)}{\partial \beta_j}\right|_{\beta_j=\beta_j^0}, \quad \ell''(\boldsymbol{\beta}^0)_j = \left.\frac{\partial^2 \ell(y_j;\beta_j)}{\partial \beta_j^2}\right|_{\beta_j=\beta_j^0}.$$

We can then simulate τ using an random walk step, then propose $\boldsymbol{\beta}$ conditional on this value using the Gaussian approximation above and accept/reject the pair $(\tau, \boldsymbol{\beta})$ using a

Metropolis step. As for the Metropolis-adjusted Langevin algorithm, we need to compute the backward move for the acceptance ratio. We refer to Section 4.4.1 of H. Rue and Held (2005) for more details.

8 Regression models

This chapter is dedicated to the study of regression models from a Bayesian standpoint. Starting with Gaussian data, we investigate the link between frequentist approaches to regularization and shrinkage priors. We also look at hierarchical models with mixed effects and variable selection using reversible jump MCMC and conditional Bayes factor.

Throughout, we consider regression models with model (or design) matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ with centered inputs, so $\mathbf{1}_n^\top \mathbf{X} = \mathbf{0}_p$. We are interested in the associated vector of regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ which describe the mean and act as weights for each covariate vector. In the ordinary linear regression model

$$\boldsymbol{Y} \mid \mathbf{X}, \boldsymbol{\beta}, \omega \sim \mathsf{Gauss}_n(\beta_0 \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}, \omega^{-1}\mathbf{I}_n),$$

so that observations are independent and homoscedastic. Inference is performed conditional on the observed covariate vectors $\mathbf{X}_i$; we omit this dependence hereafter, but note that this can be relaxed. The intercept β_0, which is added to capture the mean response and make it mean-zero, receives special treatment and is typically assigned an improper prior. We largely follow the exposition of Villani (2023).

Before we proceed with analysis of the Gaussian linear model, we derive a useful formula for completion of quadratic forms that arise in Gaussian models.

Proposition 8.1 (Decomposition of quadratic forms)**.** *For quadratic forms (in $\boldsymbol{x}$) with*

$$\begin{aligned}
&(\boldsymbol{x}-\boldsymbol{a})^\top \mathbf{A}(\boldsymbol{x}-\boldsymbol{a}) + (\boldsymbol{x}-\boldsymbol{b})^\top \mathbf{B}(\boldsymbol{x}-\boldsymbol{b}) \\
&= (\boldsymbol{x}-\boldsymbol{c})^\top \mathbf{C}(\boldsymbol{x}-\boldsymbol{c}) + (\boldsymbol{c}-\boldsymbol{a})^\top \mathbf{A}(\boldsymbol{c}-\boldsymbol{a}) + (\boldsymbol{c}-\boldsymbol{b})^\top \mathbf{B}(\boldsymbol{c}-\boldsymbol{b}) \\
&\overset{\boldsymbol{x}}{\propto} (\boldsymbol{x}-\boldsymbol{c})^\top \mathbf{C}(\boldsymbol{x}-\boldsymbol{c})
\end{aligned}$$

where $\mathbf{C} = \mathbf{A} + \mathbf{B}$ and $\boldsymbol{c} = \mathbf{C}^{-1}(\mathbf{A}\boldsymbol{a} + \mathbf{B}\boldsymbol{b})$.

Proposition 8.2 (Gaussian ordinary linear regression with conjugate priors)**.** *The conjugate prior for the Gaussian regression model for the mean and precision parameters β and ω, respectively, is a Gaussian-gamma and is defined hierarchically as*

$$\begin{aligned}
\boldsymbol{\beta} \mid \omega &\sim \mathsf{Gauss}\left\{\boldsymbol{\mu}_0, (\omega\boldsymbol{\Omega}_0)^{-1}\right\} \\
\omega &\sim \mathsf{gamma}(\nu_0/2, \tau_0/2).
\end{aligned}$$

Using properties of the Gaussian distribution, the sampling distribution of the ordinary least squares estimator is $\widehat{\boldsymbol{\beta}} \sim \mathsf{Gauss}_p\{\boldsymbol{\beta}, (\omega \mathbf{X}^\top \mathbf{X})^{-1}\}$.

The conditional and marginal posterior distributions for the mean coefficients $\boldsymbol{\beta}$ *and for the precision* ω *are*

$$\begin{aligned}\boldsymbol{\beta} \mid \omega, \boldsymbol{y} &\sim \mathsf{Gauss}_p\left\{\boldsymbol{\mu}_n, (\omega\boldsymbol{\Omega}_n)^{-1}\right\} \\ \omega \mid \boldsymbol{y} &\sim \mathsf{gamma}\left\{(\nu_0+n)/2, \tau_n^2/2\right\}, \\ \boldsymbol{\beta} \mid \boldsymbol{y} &\sim \mathsf{Student}_p(\boldsymbol{\mu}_n, \tau_n/(\nu_0+n) \times \boldsymbol{\Omega}_n^{-1}, \nu_0+n)\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{\Omega}_n &= \mathbf{X}^\top\mathbf{X} + \boldsymbol{\Omega}_0 \\ \boldsymbol{\mu}_n &= \boldsymbol{\Omega}_n^{-1}(\mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{\Omega}_0\boldsymbol{\mu}_0) = \boldsymbol{\Omega}_n^{-1}(\mathbf{X}^\top\boldsymbol{y} + \boldsymbol{\Omega}_0\boldsymbol{\mu}_0) \\ \tau_n &= \tau_0 + (\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})^\top(\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) + (\boldsymbol{\mu}_n-\widehat{\boldsymbol{\beta}})^\top\mathbf{X}^\top\mathbf{X}(\boldsymbol{\mu}_n-\widehat{\boldsymbol{\beta}}) \\ &\quad + (\boldsymbol{\mu}_n-\boldsymbol{\mu}_0)^\top\boldsymbol{\Omega}_0(\boldsymbol{\mu}_n-\boldsymbol{\mu}_0)\end{aligned}$$

Proof. Write the posterior as

$$\begin{aligned}p(\boldsymbol{\beta}, \omega \mid \boldsymbol{y}) &\propto p(\boldsymbol{y} \mid \boldsymbol{\beta}, \omega)p(\omega) \\ &\propto \omega^{n/2}\exp\left\{-\frac{\omega}{2}(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})\right\} \\ &\quad \times |\omega\boldsymbol{\Omega}_0|^{1/2}\exp\left\{-\frac{\omega}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Omega}_0(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\right\} \\ &\quad \times \omega^{\nu_0/2-1}\exp\left(-\tau_0\omega/2\right).\end{aligned}$$

We rewrite the first quadratic form in $\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}$ using the orthogonal decomposition

$$(\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) + (\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta})$$

since $(\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})^\top(\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}) = 0$. We can pull terms together and separate the conditional posterior $p(\boldsymbol{\beta} \mid \boldsymbol{y}, \omega)$ and $p(\omega \mid \boldsymbol{y})$ as

$$\begin{aligned}p(\boldsymbol{\beta}, \omega \mid \boldsymbol{y}) &\propto \omega^{(n+p+\nu_0)/2-1}\exp\left[-\frac{\omega}{2}\left\{\tau_0 + (\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})^\top(\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})\right\}\right] \\ &\quad \times \exp\left[-\frac{\omega}{2}\left\{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\mathbf{X}^\top\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) + (\boldsymbol{\beta}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Omega}_0(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\right\}\right]\end{aligned}$$

and using Proposition 8.1 for the terms in the exponent with $\boldsymbol{a} = \widehat{\beta}$, $\mathbf{A} = \mathbf{X}^\top\mathbf{X}$, $\boldsymbol{b} = \boldsymbol{\mu}_0$ and $\mathbf{B} = \boldsymbol{\Omega}_0$, we find

$$\begin{aligned} p(\boldsymbol{\beta}, \omega \mid \boldsymbol{y}) &\propto \omega^{(n+\nu_0)/2-1} \exp\left(-\frac{\omega\tau_n}{2}\right) \\ &\quad \times \omega^{p/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_n)^\top(\omega\boldsymbol{\Omega}_n)(\boldsymbol{\beta}-\boldsymbol{\mu}_n)\right\} \end{aligned}$$

whence the decomposition of the posterior as a Gaussian conditional on the precision, and a gamma for the latter. The marginal of $\boldsymbol{\beta}$ is obtained by regrouping all terms that depend on ω and integrating over the latter, recognizing the integral as an unnormalized gamma density, and thus

$$\begin{aligned} p(\boldsymbol{\beta}, \mid \boldsymbol{y}) &\overset{\boldsymbol{\beta}}{\propto} \int_0^\infty \omega^{(\nu_0+n+p)/2-1} \exp\left\{-\omega\frac{\tau_n + (\boldsymbol{\beta}-\boldsymbol{\mu}_n)^\top\boldsymbol{\Omega}_n(\boldsymbol{\beta}-\boldsymbol{\mu}_n)}{2}\right\} \mathrm{d}\omega \\ &\overset{\boldsymbol{\beta}}{\propto} \left\{\frac{\tau_n + (\boldsymbol{\beta}-\boldsymbol{\mu}_n)^\top\boldsymbol{\Omega}_n(\boldsymbol{\beta}-\boldsymbol{\mu}_n)}{2}\right\}^{-(\nu_0+n+p)/2} \\ &\overset{\boldsymbol{\beta}}{\propto} \left\{1 + \frac{(\boldsymbol{\beta}-\boldsymbol{\mu}_n)^\top\frac{\nu_0+n}{\tau_n}\boldsymbol{\Omega}_n(\boldsymbol{\beta}-\boldsymbol{\mu}_n)}{\nu_0+n}\right\}^{-(\nu_0+n+p)/2} \end{aligned}$$

so must be a Student-t distribution with location $\boldsymbol{\mu}_n$, scale matrix $\tau_n/(\nu_0+n) \times \boldsymbol{\Omega}_n^{-1}$ and $\nu_0 + n$ degrees of freedom.

The improper prior $p(\beta_0, \boldsymbol{\beta}, \omega) \propto \omega^{-1}$ can be viewed as a special case of the conjugate prior when the variance of the Gaussian is infinite and the $\mathsf{gamma}(a, b)$ when both $a, b \to 0$. □

The choice of prior precision $\boldsymbol{\Omega}_0$ is left to the user, but typically the components of the vector $\boldsymbol{\beta}$ are left apriori independent, with $\boldsymbol{\Omega}_0 \propto \lambda\mathbf{I}_n$.

Example 8.1. Study 4 of Lin et al. (2024) focuses on cultural appropriation using a fictional scenario focusing on soul food recipe cookbook from Chef Dax. The chef was identified as being African-American or not, and the authors manipulated the description of the way he obtained the recipes (either by peeking without permission in kitchens, by asking permission or without them mentioning altogether for the control category). Authors postulated that the perception of appropriation would vary by political ideology (liberal or conservative). The study results in a 3 by 2 by 2 three-way between-subject ANOVA.

We use the multivariate regression model to draw samples from $\boldsymbol{\beta}$ and use these to reconstruct the subgroup sample means for each of the 12 categories. Then, we consider the difference in perception of cultural appropriation for participants when Chef Dax is black (no cultural appropriation) versus when he is not, separately for liberals and conservatives,

by pooling data across all different descriptions. Figure 8.1 shows the posterior distribution of those contrasts, which indicate that on average liberal perceive cultural appropriation more strongly (with nearly 2 points more), than conservatives (0.7 points on average).

```
data(LKUK24_S4, package = "hecedsm")
options(contrasts = c("contr.sum", "contr.poly"))
model <- lm(appropriation ~ politideo * chefdax * brandaction,
            data = LKUK24_S4)
# Model matrix, response and dimensions
y <- LKUK24_S4$appropriation
X <- model.matrix(model)
n <- nrow(X)
p <- ncol(X)
# Priors
# We set zero for contrasts and a small value
# for the global mean (since response is on [1,7] Likert scale)
# with lower values indicating more cultural appropriation)
mu_0 <- c(2.5, rep(0, p-1))
Omega_0 <- diag(p)
# Prior precision of 0.25 (variance of 4)
nu_0 <- 0.25
tau_0 <- 1
Omega_n_inv <- solve(crossprod(X) + Omega_0)
mu_n <- Omega_n_inv %*% (crossprod(X, y) + crossprod(Omega_0, mu_0))
tau_n <- tau_0 + sum(resid(model)^2) + c(crossprod(X %*% (mu_n-coef(model)))) + sum((mu_n-mu
# Posterior draws from the model
omega <- rgamma(n = 1e3L,
                shape = (nu_0 + n)/2,
                tau_n^2/2)
beta <- matrix(rnorm(n = 1e3L*p), ncol = p) %*%
  chol(Omega_n_inv)
beta <- sweep(beta, 1, sqrt(omega), "/")
beta <- sweep(beta, 2, mu_n, "+")
# Posterior quartiles for beta
beta_qu <- t(apply(beta, 2, quantile, probs = c(0.25,0.5,0.75)))
# Standard dev. for beta (from Student-t)
beta_se <- sqrt((nu_0 + n)/(nu_0 + n - 2) *diag(tau_n/(nu_0 + n) * Omega_n_inv))

beta <- TruncatedNormal::rtmvt(
  n = 1e3L,
```

```
  mu = mu_n,
  sigma = tau_n/(nu_0 + n) * Omega_n_inv,
  df = nu_0 + n)
dfg <- expand.grid(politideo =c("conservative", "liberal"),
            chefdax = c("not black", "black"),
            brandaction = c("peeking","permission", "control"))
mm <- model.matrix( ~ politideo * chefdax * brandaction,
              data = dfg)
# Subgroup means for each of the 12 categories
mu <- tcrossprod(beta, mm)
# Contrast weights, averaging over brandaction
w1 <- rep(c(1, 0, -1, 0), length.out = nrow(dfg))/3
w2 <- rep(c(0, 1, 0, -1), length.out = nrow(dfg))/3
# Posterior distribution of contrasts
tc <- mu %*% cbind(w1, w2)
```

The coefficients and standard errors from the linear regression are very nearly similar to the posterior mean and standard deviations for β from the marginal Student-t, owing to the large sample size and uninformative priors.

In multivariate regression, it sometimes is useful to specify correlated coefficients (e.g., for random effects). This leads to the necessity to set a prior on a covariance matrix.

Proposition 8.3 (Wishart distribution)**.** *Let $\boldsymbol{Q}$ by a random $p \times p$ symmetric positive definite matrix with Wishart distribution, denoted $\mathbf{Q} \sim \mathsf{Wishart}_p(\nu, \mathbf{S})$ for $\nu > 0$ degrees of freedom and scale $\mathbf{S}$. It's density is proportional to*

$$f(\boldsymbol{Q}) \overset{\boldsymbol{Q}}{\propto} |\boldsymbol{Q}|^{(\nu-p-1)/2} \exp\{-\mathrm{tr}(\mathbf{S}^{-1}\boldsymbol{Q})/2\}, \qquad \nu > p-1.$$

where $|\cdot|$ denotes the determinant of the matrix and tr the trace operator. The Wishart also arises from considering $n \geq p$ independent and identically distributed mean zero Gaussian vectors $\boldsymbol{Y}_i \sim \mathsf{Gauss}_p(\mathbf{0}_p, \mathbf{S})$, where

$$\sum_{i=1}^{\nu} \boldsymbol{Y}_i \boldsymbol{Y}_i^\top \sim \mathsf{Wishart}_p(\nu, \mathbf{S}).$$

For prior elicitation, ν is thus a prior sample size, whereas we can specify $\mathbf{S}$ using the fact that the mean of the Wishart is $\nu\mathbf{S}$; taking an identity matrix is standard. For more mathematical properties, consult Chapter 8 of Eaton (2007).

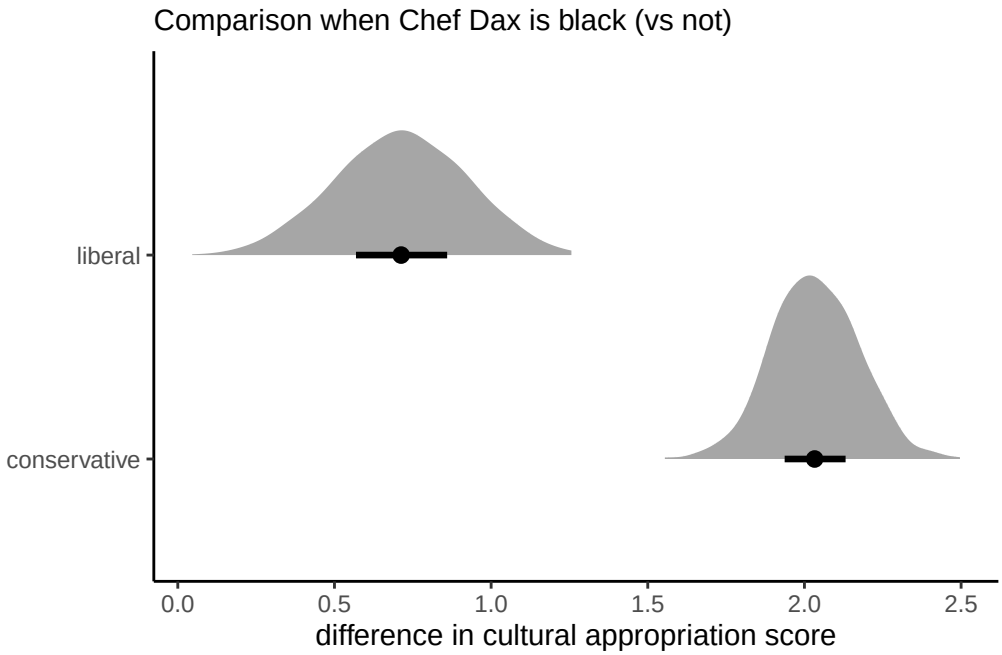

Figure 8.1: Difference in appropriation rating for black vs non-black Chef Dax, average accross different levels of brand action.

Definition 8.1 (Inverse Wishart)**.** Analogous to the relationship between gamma prior for the precision and inverse gamma for the variance, we can also similarly consider a prior for the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{Q}^{-1}$. Applying the change of variable formula, we get Jacobian $|\boldsymbol{\Sigma}|^{p+1}$, and so $\boldsymbol{\Sigma}$ is inverse Wishart $\mathsf{inv.Wishart}(\nu, \mathbf{S}^{-1})$, with density proportional to

$$p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu+p+1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{S}^{-1}\boldsymbol{\Sigma}^{-1}\right)\right\}$$

with expectation $\mathbf{S}^{-1}(\nu - p - 1)$ for $\nu > p + 1$.

Proposition 8.4 (Wishart as conjugate prior in Gaussian model)**.** *Consider* $\boldsymbol{\mu} \sim \mathsf{Gauss}_p(\boldsymbol{\mu}_0, \boldsymbol{Q}^{-1})$ *and* $\boldsymbol{Q} \sim \mathsf{Wishart}_p(\nu, \mathbf{S})$ *for* $\nu \geq p$*. Then, the conditional density of* $\boldsymbol{Q} \mid \boldsymbol{\mu}, \boldsymbol{\mu}_0$ *is proportional to*

$$|\boldsymbol{Q}|^{1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{Q}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right\} |\boldsymbol{Q}|^{(\nu-p-1)/2} \exp\{-\mathrm{tr}(\mathbf{S}^{-1}\boldsymbol{Q})/2\}$$

and thus $\mathsf{Wishart}_p\{\nu + 1/2, \boldsymbol{S} + (\boldsymbol{\mu} - \boldsymbol{\mu}_0)(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top\}$*. To see this, note that a* 1×1 *matrix is equal to it's trace, and the trace operator is invariant to cyclic of it's argument, meaning that*

$$(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{Q}(\boldsymbol{\mu} - \boldsymbol{\mu}_0) = \mathrm{tr}\left\{\boldsymbol{Q}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top\right\}.$$

We then combine elements to get the parameters. This extends naturally to the case of n independent observations, for example linear regression model.

Proposition 8.5 (Priors for variance matrices)**.** *The marginal precision for the Wishart variate are gamma distributed with the same degrees of freedom ν. The problem with using this prior is that it has a single parameter governing all scale parameters (i.e., the marginal variance) and the absolute value of the correlation and marginal variance parameters are negatively related (Gelman et al. 2013), as seen in Figure 8.2. Large variance thus correspond to small correlations shrunk towards zero when the degrees of freedom increase. There exists alternative distributions, such as the scaled-inverse Wishart, that add redundant scale parameters to decouple $\mathrm{diag}(\boldsymbol{\xi}) \times \boldsymbol{\Sigma} \times \mathrm{diag}(\boldsymbol{\xi})$, but we refrain from pursuing these here.*

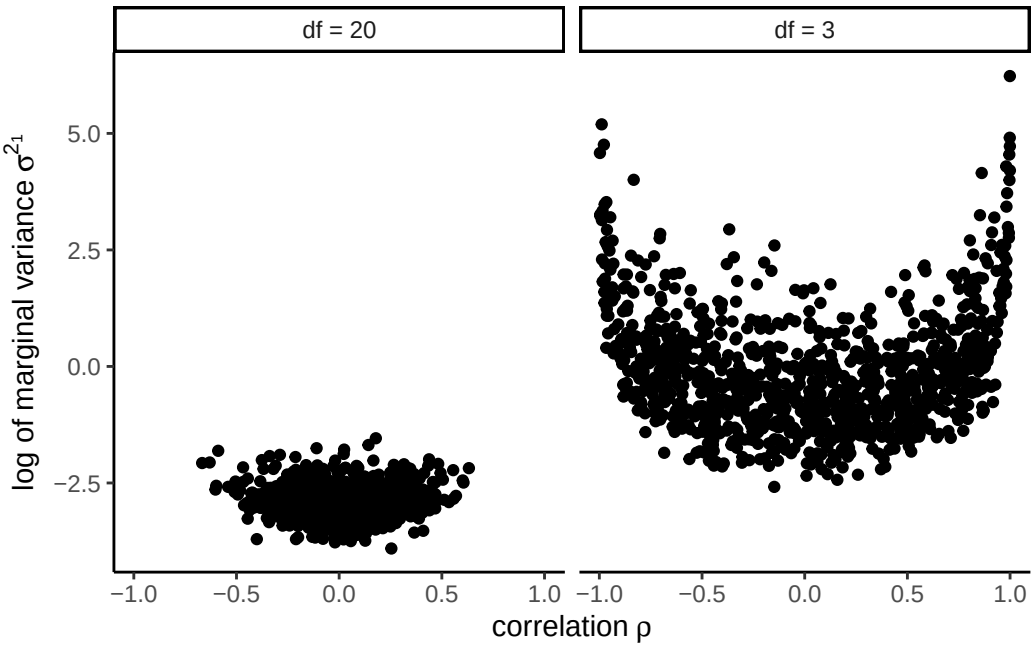

Figure 8.2: Prior draws from a bivariate inverse Wishart with identity scale matrix and $\nu \in \{3, 20\}$ degrees of freedom.

A better alternative is to specify different prior for each marginal scale σ_j and a prior on the correlation matrix $\mathbf{R}$. For the latter, the onion peel or LKJ prior, named after the authors of Lewandowski, Kurowicka, and Joe (2009), is $p(\mathbf{R}) \propto |\mathbf{R}|^{\eta-1}$ for a scale $\eta > 0$. The case $\eta = 1$ leads to uniform over the space of correlation matrices, and $\eta > 1$ favours the identity matrix.

8.1 Shrinkage priors

In contexts where the number of regressors p is considerable relative to the sample size n, it may be useful to constrain the parameter vector if we assume that the signal is sparse, with a large proportion of coefficients that should be zero. This is notably important when the ratio $p/n \to c$ for $c > 0$, meaning that the number of coefficients and covariates increases proportional to the sample size. Shrinkage priors can regularize and typically consist of distributions that have a mode at zero, and another that allows for larger signals.

In the Bayesian paradigm, regularization is achieved via priors that have mass at or towards zero, pushing coefficients of the regression model towards zero unless there is strong evidence from the likelihood against this. We however want to allow non-zero coefficients, typically by setting coefficient-specific parameters with a heavy tailed distribution to prevent overshrinking. Most if not all parameter can be viewed as scale mixtures of Gaussian.

We make a distinction between **global** shrinkage priors those that consider a common shrinkage parameter for all regression coefficients, to be compared with **local** scale mixtures that have coefficient-specific parameters.

The scale parameters and hyperparameters can be estimated jointly with the model, and uncertainty diagnostics follow naturally. We assume that coefficients β_j are independent apriori, although it is possible to specify group structures (e.g., for handling coefficients associated to a common categorical covariate with K levels, represented by $K-1$ columns of dummy group indicators).

Proposition 8.6 (Spike-and-slab prior)**.** *The spike-and-slab prior is a two-component mixture with that assigns a positive probability to zero via a point mass δ_0 or a vary narrow distribution centered at the origin (the spike) and the balance to the slab, a diffuse distribution.*

The spike-and-slab prior was originally proposed by Mitchell and Beauchamp (1988) with a uniform on a large interval and a point mass at zero. The term was also used in George and McCulloch (1993), which replaced the prior by a mixture of Gaussians, one of which diffuse and the other with near infinite precision and centered at the origin. The latter is known under the vocable stochastic search variable selection prior. Letting $\gamma_j \in [0,1]$ denote the probability of the slab or inclusion of the variable, the independent priors for the regression coefficients are

$$\beta_j \mid \gamma_j, \sigma_j^2, \phi_j^2 \sim (1-\gamma_j)\mathsf{Gauss}(0, \sigma_j^2\phi_j^2) + \gamma_j \mathsf{Gauss}(0, \phi^2)$$

where ϕ_j^2 is very nearly zero, e.g., $\phi_j^2 = 0.001$. The construction allows for variable augmentation with mixture indicators and Gibbs sampling, although mixing tends to be poor.

Proposition 8.7 (Horseshoe prior)**.** *The horseshoe prior of Carvalho, Polson, and Scott (2010) is a hierarchical prior of the form*

$$\beta_j \mid \sigma_j^2 \sim \mathsf{Gauss}(0, \sigma_j^2), \quad \sigma_j^2 \mid \lambda \sim \mathsf{Student}_+(0, \lambda, 1), \quad \lambda \sim \mathsf{Student}_+(0, \omega, 1)$$

where $\mathsf{Student}_+(0, a, 1)$ *denotes a half-Cauchy distribution with scale* $a > 0$*, truncated on* $\mathbb{R}_+$*.*

This prior has no explicit density, but is continuous and can be simulated. It is useful to consider the behaviour of the random variance σ_j^2 *term, which leads to an unconditional scale mixture of Gaussian for* β_j*. More useful to understanding is looking at* $\kappa = 1 - 1/(1+\sigma^2)$*, which gives a weight in* $[0, 1]$*. We can see what happens to the shrunk components close to zero by looking at the density of* $\kappa \to 0$*, and similarly at the large signals when* $\kappa \to 1$*. The Cauchy prior does not shrink towards zero, and lets large signals pass, whereas the Bayesian LASSO double exponential shrinkage leads to attenuation of strong signals. The horseshoe prior name comes from the shape of the prior, which leads to a shrinkage analog to* $\mathsf{beta}(1/2, 1/2)$ *and thus penalizes, forcing components to be either large or small.*

Figure 8.3 shows the weighting implied by the mixture density for a Cauchy prior on the variance, the double exponential of the Laplace from Bayesian LASSO and the horseshoe.

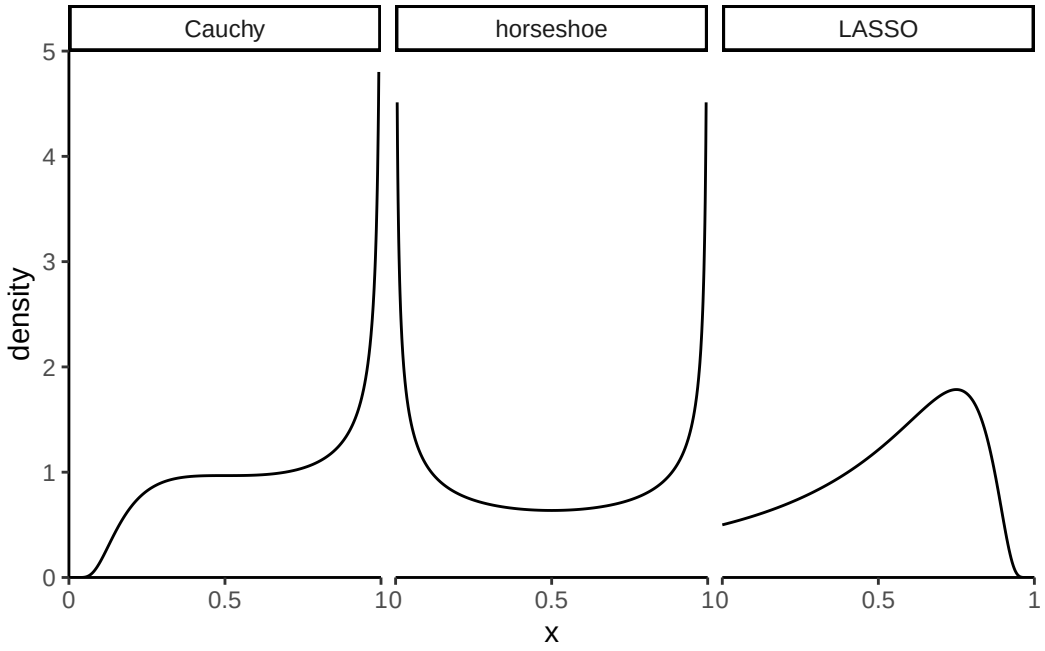

Figure 8.3: Density of penalization weights κ of spike (near zero) and slab (near one) for three shrinkage priors.

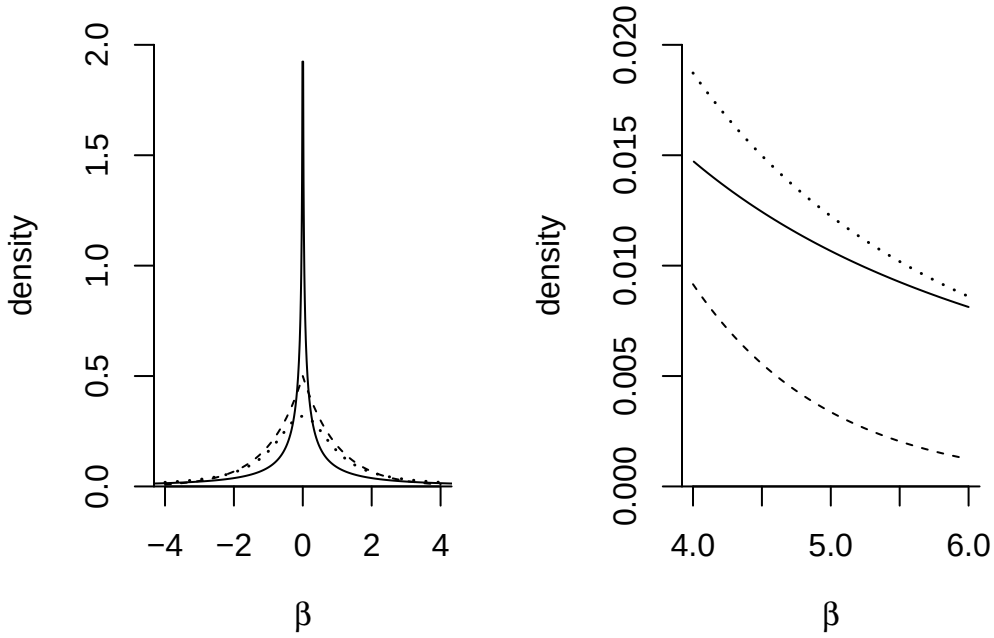

Figure 8.4: Marginal density for a regression coefficient β with horseshoe prior (full), Laplace (dashed) and a Student-t (thick dotted). The plot on the right shows the tail behaviour. The density of the horseshoe is unbounded at the origin. Inspired from Figure 1 of Carvalho, Polson, and Scott (2010).

While the horseshoe prior guarantees that large coefficients are not regularized, this feature of the shrinkage prior is harmful in certain instances, for example separation of variables for logistic regression. Markov chain Monte Carlo simulations are hampered by these parameters whose posterior mean does not exist, leading to poor mixing. Some very weak regularization for these big components can thus help. Piironen and Vehtari (2017) proposed the regularized horseshoe, nicknamed Finnish horseshoe, where

$$\begin{aligned}\beta_j \mid \lambda, \tau_j, c^2 &\sim \mathsf{Gauss}\left(0, \lambda\frac{c^2\tau_j^2}{c^2+\tau_j^2\lambda^2}\right),\\ \tau_j &\sim \mathsf{Student}_+(0,1,1)\\ c^2 \mid s^2, \nu &\sim \mathsf{inv.gamma}(\nu/2, \nu s^2/2).\end{aligned}$$

When $\tau^2\lambda_j^2$ is much greater than c^2, this amounts to having a Student slab with ν degrees of freedom for large coefficients; taking a small value of ν allows for large, but not extreme components, and the authors use $s^2 = 2, \nu = 4$. The above specification does not specify the prior for the global scale λ, for which Piironen and Vehtari (2017) recommend using an

empirical Bayes prior, with

$$\lambda \sim \text{Student}_+ \left\{0, \frac{p_0}{(p-p_0)}\frac{\sigma}{n^{1/2}}, 1\right\},$$

where p_0 is a prior guess for the number of non-zero components out of p, n is the sample size and σ is some level of the noise.

Example 8.2 (Comparison of shrinkage priors)**.** We revisit the `diabetes` data from the **R** package `lars`, which was used in Park and Casella (2008) to illustrate the Bayesian LASSO. We consider three methods: the default Gaussian prior, which gives a ridge penalty, the Bayesian LASSO and finally the horseshoe. Models are fitted using the `bayesreg` package.

Figure 8.5 shows the ordered coefficients for each method. We can see that the ridge has the widest intervals of all methods, providing some shrinkage only for large values of β. The horseshoe has typically narrower intervals, with more mass in a neighborhood of zero for smaller coefficients, and asymmetric intervals.

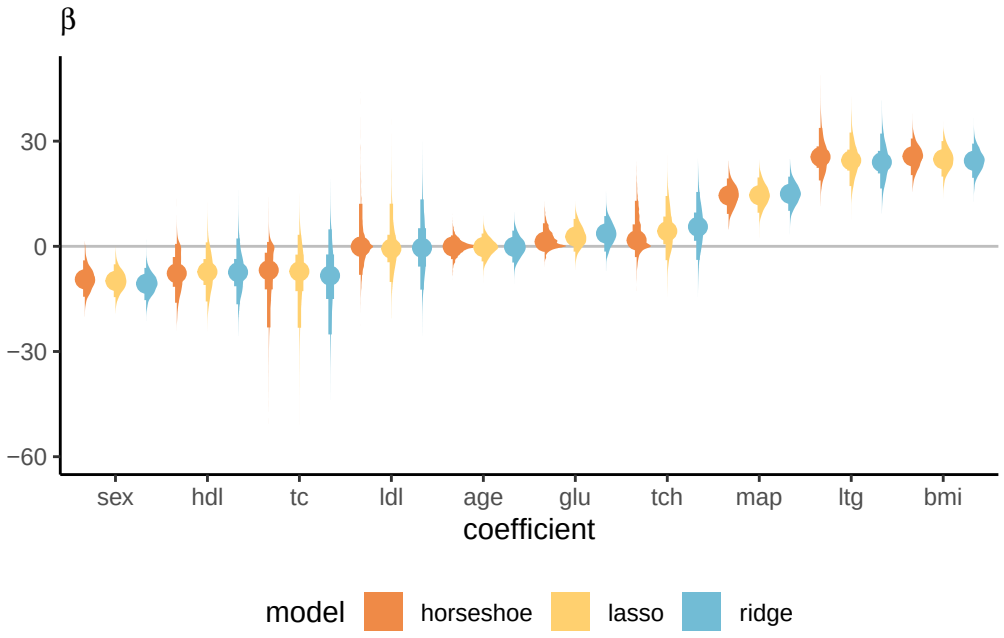

Figure 8.5: Density estimates for regression coefficients with Gaussian (ridge), double exponential (Laplace) and horseshoe priors for the `diabetes` data.

One aspect worth mentioning is that the horseshoe prior impacts strongly the geometry and leads to slower mixing than the conjugate or Student-t priors: the effective sample size

fraction relative to the number of samples ranges from 15% to 100%, compared to 50% to 96% for the Bayesian LASSO and near-independent draws with the conjugate ridge.

8.2 Bayesian model averaging via reversible jump

Variable selection refers to selection of covariates from a set of p candidates from a design matrix which is orthogonal to the intercept; these could include higher order terms or interactions. From the Bayesian perspective, all of these $2^p - 1$ submodels must be assigned a prior. We can also rule out models without lower order interactions through the prior, but exploring or fitting all possible models is too costly. The goal of the analysis is to account for the uncertainty associated with variable selection. The target of inference is often the weight of the combinations, and the probability that a predictor is included aposteriori.

There are different methods for doing this: one is through spike and slab priors. Software for Hamiltonian Monte Carlo like Stan cannot handle discrete variables. Other methods for Bayesian model averaging includes use of reversible jump Markov chain Monte Carlo (Green 1995), an extension that allows for arbitrary measures and through this varying dimensions, which occurs not only with variable selection, but also changepoint analysis and mixture models with varying number of components (Green 2001). Reversible jump requires a form of data augmentation to have dimension-balancing and defining different types of moves for jumping between dimensions. These are integrated in the Metropolis–Hastings step through a Jacobian term added to R. with different types of moves giving termed padding andone to pad the dimensions to match and define. In regression models, we will consider moves that adds or removes one parameter/regressor at a time.

Consider the setup of Proposition 8.2, where we consider models $M_1, \ldots, M_m$ with for simplicity $p(M_i) = 1$ for all models that include an intercept. We define $\mathbf{X}^{(m)}$ and $\boldsymbol{\beta}^{(m)}$ as the model matrix and the associated vector of non-zero coefficients associated with it. Let the response vector

$$\boldsymbol{Y} \mid M_m, \boldsymbol{\beta}, \sim \mathsf{Gauss}(\mathbf{X}^{(m)}\boldsymbol{\beta}^{(m)}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\beta}^{(m)} \mid M_m$, is assigned a Gaussian prior, and likewise the covariance parameters are given suitable priors. We write $|M|$ for the cardinality of the set of non-zero coefficients $\boldsymbol{\beta}$ in model M.

Write $\boldsymbol{\theta}$ for all parameters other than the response, model and vector of coefficients. We can consider a joint update of the regression parameters $\boldsymbol{\beta}, M$ by sampling from their joint distribution via $p(\boldsymbol{\beta} \mid M, \boldsymbol{\theta})p(M \mid \boldsymbol{\theta})$. The update for the coefficients $p(\boldsymbol{\beta} \mid M, \boldsymbol{\theta})$ is as usual,

while for $p(M \mid \boldsymbol{\theta})$, we get the conditional Bayes factor,

$$\begin{aligned} p(M \mid \boldsymbol{Y}, \boldsymbol{\theta}) &\overset{M}{\propto} p(M) p(\boldsymbol{Y} \mid M, \boldsymbol{\theta}) \\ &= p(M) \int_{\mathbb{R}^{|\mathbb{M}|}} p(\boldsymbol{Y} \mid M, \boldsymbol{\beta}, \boldsymbol{\theta}) p(\boldsymbol{\beta} \mid M, \boldsymbol{\theta}) d\boldsymbol{\beta} \end{aligned}$$

We can marginalize out $\boldsymbol{\beta}$ and show (demonstration omitted) that

$$p(\boldsymbol{M} \mid \boldsymbol{Y}, \boldsymbol{\theta}) \propto p(M) |\boldsymbol{Q}_\beta|^{-1/2} \exp\left(\frac{1}{2} \boldsymbol{\mu}_\beta^\top \boldsymbol{Q}_\beta \boldsymbol{\mu}_\beta\right)$$

where $\boldsymbol{\mu}_\beta$ and $\boldsymbol{Q}_\beta$ are the mean and precision of $p(\beta \mid \boldsymbol{Y}, M, \boldsymbol{\theta})$.

We then consider different types of move for the $k_{\max}$ potential covariates (including interactions, etc.) (Holmes, Denison, and Mallick 2002)

- expansion: adding an unused covariate chosen at random from the remaining ones
- shrinkage: removing one covariate at random from the current matrix
- swapping an active covariate for an unused one.

Only the last type of move preserves the dimension, whereas the shrinkage and expansion moves lead to a decrease or increase of the model matrix dimension. The probability of rejection for Metropolis becomes

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

where for most moves $J = 1$ in this case, except in four cases where the dimension $|M| \in \{1, 2, k_{\max} - 1, k_{\max}\}$ and

- $J = 2/3$ if $|M| = 1$ and we try to add a covariate
- $J = 2/3$ if $|M| = k_{\max}$ and we try to remove a covariate
- $J = 3/2$ if $|M| = 2$ and we try to remove a covariate
- $J = 3/2$ if $|M| = k_{\max} - 1$ and we try to add the last covariate.

The move for M is accepted as usual if the drawn uniform $U \sim \mathsf{unif}(0, 1)$ satisfies $U < R$.

9 Deterministic approximations

So far, we have focused on stochastic approximations of integral. In very large models, Markov chain Monte Carlo suffer from the curse of dimensionality and it is sometimes useful to resort to cheaper approximations. We begin this review by looking at the asymptotic Gaussian limiting distribution of the maximum aposteriori, the Laplace approximations for integrals (Tierney and Kadane 1986), and their applications for model comparison (Raftery 1995) and evaluation of the marginal likelihood. We also discuss integrated nested Laplace approximations (Håvard Rue, Martino, and Chopin 2009; Wood 2019), used in hierarchical models with Gaussian components to obtain approximations to the marginal distribution. This material also borrows from Section 8.2 and appendix C.2.2 of Held and Bové (2020).

We make use of Landau's notation to describe the growth rate of some functions: we write $x = \mathrm{O}(n)$ (big-O) to indicate that the ratio $x/n \to c \in \mathbb{R}$ and $x = \mathrm{o}(n)$ when $x/n \to 0$, both when $n \to \infty$.

9.1 Laplace approximation and it's applications

Proposition 9.1 (Laplace approximation for integrals)**.** *The Laplace approximation uses a Gaussian approximation to evaluate integrals of the form*

$$I_n = \int_a^b g(x)\mathrm{d}x = \int_a^b \exp\{nh(x)\}\mathrm{d}x.$$

Assume that $g(x)$ and thus $h(x)$, is concave and and twice differentiable, with a maximum at $x_0 \in [a,b]$. We can Taylor expand $h(x)$ to get,

$$h(x) = h(x_0) + h'(x_0)(x-x_0) + h''(x_0)(x-x_0)^2/2 + R$$

where the remainder $R = \mathrm{O}\{(x-x_0)^3\}$. If x_0 is a maximizer and solves $h'(x_0) = 0$, then letting $\tau = -nh''(x_0)$, we can write ignoring the remainder term the approximation

$$\begin{aligned} I_n &\approx \exp\{nh(x_0)\} \int_a^b \exp\left\{-\frac{1}{2}(x-x_0)^2\right\} \\ &= \exp\{nh(x_0)\} \left(\frac{2\pi}{\tau}\right)^{1/2} \left[\Phi\left\{\tau(b-x_0)\right\} - \Phi\left\{\tau(a-x_0)\right\}\right] \end{aligned}$$

upon recovering the unnormalized kernel of a Gaussian random variable centered at x_0 *with precision* τ*. The approximation error is* $\mathrm{O}(n^{-1})$.

The multivariate analog is similar, where now for an integral of the form $\exp\{nh(\boldsymbol{x})\}$ *over a subset of* $\mathbb{R}^d$*, we consider the Taylor series expansion*

$$h(\boldsymbol{x}) = h(\boldsymbol{x}_0) + (\boldsymbol{x}-\boldsymbol{x}_0)^\top h'(\boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^\top h''(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0) + R.$$

We obtain the Laplace approximation at the mode $\boldsymbol{x}_0$ *satisfying* $h'(\boldsymbol{x}_0) = \mathbf{0}_d$,

$$I_n \approx \left(\frac{2\pi}{n}\right)^{p/2} |\mathbf{H}(\boldsymbol{x}_0)|^{-1/2} \exp\{nh(\boldsymbol{x}_0)\},$$

where $|\mathbf{H}(\boldsymbol{x}_0)|$ *is the determinant of the Hessian matrix of* $-h(\boldsymbol{x})$ *evaluated at the mode* $\boldsymbol{x}_0$.

Laplace approximation uses a Taylor series approximation to approximate the density, but since the latter must be non-negative, it performs the approximation on the log scale and back-transform the result. It is important to understand that we can replace $nh(\boldsymbol{x})$ by any $\mathrm{O}(n)$ term.

Corollary 9.1 (Laplace approximation for marginal likelihood)**.** *Consider a simple random sample* $\boldsymbol{Y}$ *of size* n *from a distribution with parameter vector* $\boldsymbol{\theta} \in \mathbb{R}^p$*. We are interested in approximating the marginal likelihood for a parametric model with* $\boldsymbol{\theta} \in \mathbb{R}^d$*. Write (Raftery 1995)*

$$p(\boldsymbol{y}) = \int_{\mathbb{R}^d} p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$$

and take

$$nh(\boldsymbol{\theta}) = \log p(\boldsymbol{y} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

in Proposition 9.1. Then, evaluating at the maximum a posteriori $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}$*, we get*

$$p(\boldsymbol{y}) = p(\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}) p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}) (2\pi)^{d/2} |\mathbf{H}(\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}})|^{-1/2} + \mathrm{O}(n^{-1})$$

where $-\mathbf{H}$ *is the Hessian matrix of second partial derivatives of the unnormalized log posterior. We get the same relationship on the log scale, whence (Tierney and Kadane 1986)*

$$\log p(\boldsymbol{y}) = \log p(\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}) + \log p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}) + \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{H}(\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}})| + \mathrm{O}(n^{-1})$$

If $p(\boldsymbol{\theta}) = \mathrm{O}(1)$ *and* $p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \mathrm{O}(n)$ *and provided the prior does not impose unnecessary support constraints, we get the same limiting approximation if we replace the maximum a*

posteriori point estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ by the maximum likelihood estimator, and $-\mathbf{H}(\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}})$ by $n\boldsymbol{\imath}$, where $\boldsymbol{\imath}$ denotes the Fisher information matrix for a sample of size one. We can write the determinant of the n-sample Fisher information as $n^d|\boldsymbol{\imath}|$.

If we use this approximation instead, we get

$$\log p(\boldsymbol{y}) = \log p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}) - \frac{d}{2}\log n + \\ \log p(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}) - \frac{1}{2}\log|\boldsymbol{\imath}| + \frac{d}{2}\log(2\pi) + \mathrm{O}(n^{-1/2})$$

where the error is now $\mathrm{O}(n^{-1/2})$ due to replacing the true information by the evaluation at the MLE. The likelihood is $\mathrm{O}(n)$, the second is $\mathrm{O}(\log n)$ and the other three are $\mathrm{O}(1)$. If we take the prior to be a multivariate Gaussian with mean $\boldsymbol{\theta}_{\mathrm{MLE}}$ and with variance $\boldsymbol{\imath}$, then the approximation error is $\mathrm{O}(n^{-1/2})$, whereas the marginal likelihood has error $\mathrm{O}(1)$ if we only keep the first two terms. This gives the approximation

$$-2\log p(\boldsymbol{y}) \approx \mathsf{BIC} = -2\log p(\boldsymbol{y} \mid \boldsymbol{\theta}) + p\log n$$

If the likelihood contribution dominates the posterior, the BIC *approximation will improve with increasing sample size, so $\exp(-\mathsf{BIC}/2)$ is an approximation fo the marginal likelihood sometimes used for model comparison in Bayes factor, although this derivation shows that the latter neglects the impact of the prior.*

Example 9.1 (Bayesian model averaging approximation)**.** Consider the `diabetes` model from Park and Casella (2008). We fit various linear regression models, considering all best models of a certain type with at most the 10 predictors plus the intercept. In practice, we typically restrict attention to models within some distance of the lowest BIC value, as the weights otherwise will be negligible.

Most of the weight is on a handful of complex models, where the best fitting model only has around 30% of the posterior mass.

Remark 9.1 (Parametrization for Laplace). Compare to sampling-based methods, the Laplace approximation requires optimization to find the maximum of the function. The Laplace approximation is not invariant to reparametrization: in practice, it is best to perform it on a scale where the likelihood is as close to quadratic as possible in $g(\boldsymbol{\theta})$ and back-transform using a change of variable.

We can also use Laplace approximation to obtain a crude second-order approximation to the posterior. We suppose that the prior is proper.

We can Taylor expand the log prior and log density around their respective mode, say $\widehat{\boldsymbol{\theta}}_0$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}$, with $\jmath_0(\widehat{\boldsymbol{\theta}}_0)$ and $\jmath(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}})$ denoting negative of the corresponding Hessian matrices

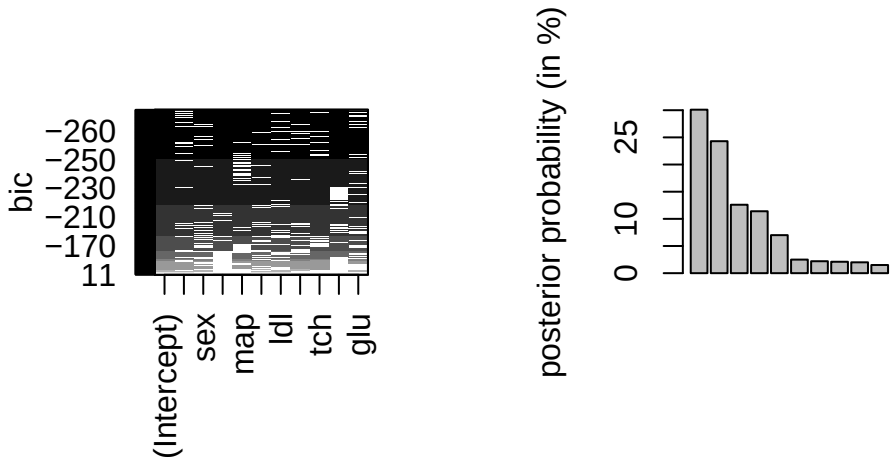

Figure 9.1: BIC as a function of the linear model covariates (left) and Bayesian model averaging approximate weights (in percentage) for the 10 models with the highest posterior weights according to the BIC approximation.

evaluated at their mode, meaning the observed information matrix for the likelihood component. Together, these yield

$$\log p(\boldsymbol{\theta}) \approx \log p(\widehat{\boldsymbol{\theta}}_0) - \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_0)^\top \jmath_0(\widehat{\boldsymbol{\theta}}_0)(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_0)$$
$$\log p(\boldsymbol{y} \mid \boldsymbol{\theta}) \approx \log p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}) - \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathrm{MLE}})^\top \jmath(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathrm{MLE}})$$

In the case of flat prior, the curvature is zero and the prior contribution vanishes altogether. If we apply now Proposition 8.1 to this unnormalized kernel, we get that the approximate posterior must be Gaussian with precision $\jmath_n^{-1}$ and mean μ_n, where

$$\jmath_n = \jmath_0(\widehat{\boldsymbol{\theta}}_0) + \jmath(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}})$$
$$\widehat{\boldsymbol{\theta}}_n = \jmath_n^{-1}\left\{\jmath_0(\widehat{\boldsymbol{\theta}}_0)\widehat{\boldsymbol{\theta}}_0 + \jmath(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}})\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}\right\}$$

and note that $\jmath_0(\widehat{\boldsymbol{\theta}}_0) = \mathrm{O}(1)$, whereas $\jmath_n$ is $\mathrm{O}(n)$.

Theorem 9.1 (Bernstein-von Mises theorem)**.** *Consider any estimator asymptotically equivalent to the maximum likelihood estimator and suppose that the prior is continuous and positive in a neighborhood of the maximum. Assume further that the regularity conditions for maximum likelihood estimator holds. Then, in the limit as* $n \to \infty$

$$\boldsymbol{\theta} \mid \boldsymbol{y} \dot{\sim} \mathsf{Gauss}\{\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}, \jmath^{-1}(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}})\}$$

The conclusions from this result is that, in large samples, the inference obtained from using likelihood-based inference and Bayesian methods will be equivalent: credible intervals will also have guaranteed frequentist coverage.

We can use the statement by replacing the maximum likelihood estimator and the observed information matrix with variants thereof ($\boldsymbol{\theta}_n$ and $\jmath_n$, or the Fisher information, or any Monte Carlo estimate of the posterior mean and covariance). The differences will be noticeable for small samples, but will vanish as n grows.

Example 9.2 (Gaussian approximations to the posterior)**.** To assess the performance of Laplace approximation, we consider an exponential likelihood $Y_i \mid \lambda \sim \mathsf{expo}(\lambda)$ with conjugate gamma prior $\lambda \sim \mathsf{gamma}(a, b)$. The exponential model has information $i(\lambda) = n/\lambda^2$ and the mode of the posterior is

$$\widehat{\lambda}_{\mathrm{MAP}} = \frac{n + a - 1}{\sum_{i=1}^n y_i + b}.$$

Let us now use Laplace approximation to obtain an estimate of the marginal likelihood: the latter is

$$p(\boldsymbol{y}) = \frac{\Gamma(n+a)}{\Gamma(a)} \frac{b^a}{(b + \sum_{i=1}^n y_i)^{n+a}}.$$

For the sample of size 62, the exponential model marginal likelihood is -276.5, whereas the Laplace approximation gives -281.9.

Proposition 9.2 (Posterior expectation using Laplace method)**.** *If we are interested in computing the posterior expectation of a positive real-valued functional* $g(\boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R}_+$*, we may write*

$$\mathsf{E}_{\boldsymbol{\Theta} \mid \boldsymbol{Y}}(g(\boldsymbol{\theta}) \mid \boldsymbol{y}) = \frac{\int g(\boldsymbol{\theta}) p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}{\int p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}$$

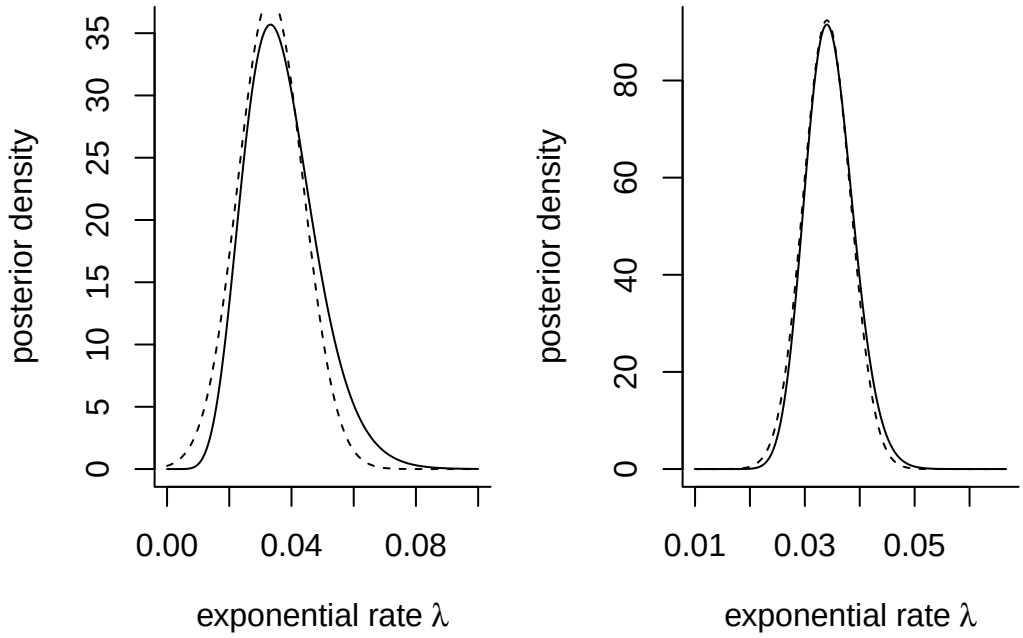

Figure 9.2: Gaussian approximation (dashed) to the posterior density (full line) of the exponential rate λ for the `waiting` dataset with an exponential likelihood and a gamma prior with $a = 0.01$ and $b = 0.01$. The plots are based on the first 10 observations (left) and the whole sample of size $n = 62$ (right).

We can apply Laplace's method to both numerator and denominator. Let $\widehat{\boldsymbol{\theta}}_g$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ of the integrand of the numerator and denominator, respectively, and the negative of the Hessian matrix of the log integrands

$$\begin{aligned} \jmath_g &= -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\left\{\log g(\boldsymbol{\theta}) + \log p(\boldsymbol{y} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\right\}, \\ \jmath &= -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\left\{\log p(\boldsymbol{y} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\right\}. \end{aligned}$$

Putting these together

$$\mathsf{E}_{\boldsymbol{\Theta}\mid\boldsymbol{Y}}(g(\boldsymbol{\theta}) \mid \boldsymbol{y}) = \frac{|\jmath(\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}})|^{1/2}}{|\jmath_g(\widehat{\boldsymbol{\theta}}_g)|^{1/2}} \frac{g(\widehat{\boldsymbol{\theta}}_g)p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}}_g)p(\widehat{\boldsymbol{\theta}}_g)}{p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}}_{\mathrm{MAP}})p(\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}})} + \mathrm{O}(n^{-2})$$

While the Laplace method has an error $\mathrm{O}(n^{-1})$, the leading order term of the expansion cancel out from the ratio.

Example 9.3 (Posterior mean for the exponential likelihood)**.** Consider the posterior mean $\mathsf{E}_{\Lambda|\boldsymbol{Y}}(\lambda)$ for the model of Example 9.2. Let $s = \sum_{i=1}^{n} y_i$. Then,

$$\widehat{\lambda}_g = \frac{(n+a)}{s+b}$$
$$|\jmath_g(\widehat{\lambda}_g)|^{1/2} = \left(\frac{n+a}{\widehat{\lambda}_g^2}\right)^{1/2} = \frac{s+b}{(n+a)^{1/2}}$$

Simplification gives the approximation

$$\widehat{\mathsf{E}}_{\Lambda|\boldsymbol{Y}}(\Lambda) \approx \frac{\exp(-1)}{s+b}\frac{(n+a)^{n+a+1/2}}{(n+a-1)^{n+a-1/2}}$$

which gives 0.03457, whereas the true posterior mean is $(n+a)/(s+b) = 0.03457$. The Laplace approximation is equal to the true value up to five significant digits.

9.2 Integrated nested Laplace approximation

In many high dimensional models, use of MCMC is prohibitively expensive and fast, yet accurate calculations are important. One class of models whose special structure is particularly amenable to deterministic approximations.

Consider a model with response $\boldsymbol{y}$ which depends on covariates $\mathbf{x}$ through a latent Gaussian process; typically the priors on the coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$. In applications with splines, or space time processes, the prior precision matrix for $\boldsymbol{\beta}$ will be sparse with a Gaussian Markov random field structure. The dimension p can be substantial (several thousands) with a comparably low-dimensional hyperparameter vector $\boldsymbol{\theta} \in \mathbb{R}^m$. Interest typically then lies in marginal parameters

$$p(\beta_i \mid \boldsymbol{y}) = \int p(\beta_i \mid \boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}$$
$$p(\theta_i \mid \boldsymbol{y}) = \int p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}_{-i}$$

where $\boldsymbol{\theta}_{-i}$ denotes the vector of hyperparameters excluding the ith element θ_i. The INLA method builds Laplace approximations to the integrands $p(\beta_i \mid \boldsymbol{\theta}, \boldsymbol{y})$ and $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, and evaluates the integral using quadrature rules over a coarse grid of values of $\boldsymbol{\theta}$.

The marginal posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ is approximated by writing $p(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \boldsymbol{y}) \propto p(\boldsymbol{\beta} \mid \boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta} \mid \boldsymbol{y})$ and performing a Laplace approximation for fixed value of $\boldsymbol{\theta}$ for the term $p(\boldsymbol{\beta} \mid \boldsymbol{\theta}, \boldsymbol{y})$, whose

mode we denote by $\widehat{\boldsymbol{\beta}}$. This yields

$$\widetilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \frac{p(\widehat{\boldsymbol{\beta}}, \boldsymbol{\theta} \mid \boldsymbol{y})}{p_G(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{y}, \boldsymbol{\theta})} = \frac{p(\widehat{\boldsymbol{\beta}}, \boldsymbol{\theta} \mid \boldsymbol{y})}{|\mathbf{H}(\widehat{\boldsymbol{\beta}})|^{1/2}}$$

and the Laplace approximation has kernel

$$p_G(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{\theta}) \propto |\mathbf{H}(\widehat{\boldsymbol{\beta}})|^{1/2} \exp\{-(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top \mathbf{H}(\widehat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})/2\};$$

since it is evaluated at $\widehat{\boldsymbol{\beta}}$, we retrieve only the determinant of the negative Hessian of $p(\boldsymbol{\beta} \mid \boldsymbol{\theta}, \boldsymbol{y})$, namely $\mathbf{H}(\widehat{\boldsymbol{\beta}})$. Note that the latter is a function of $\boldsymbol{\theta}$.

To obtain $p(\theta_i \mid \boldsymbol{y})$, we then proceed with

1. finding the mode of $\widetilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y})$ using a Newton's method, approximating the gradient and Hessian via finite differences.
2. Compute the negative Hessian at the mode to get an approximation to the covariance of $\boldsymbol{\theta}$. Use an eigendecomposition to get the principal directions $\boldsymbol{z}$.
3. In each direction of $\boldsymbol{z}$, consider drops in $\widetilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y})$ as we move away from the mode and define a coarse grid based on these, keeping points where the difference in $\widetilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y})$ relative to the mode is less than some numerical tolerance δ.
4. Retrieve the marginal by numerical integration using the central composition design outline above. We can also use directly avoid the integration and use the approximation at the posterior mode of $\widetilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y})$.

To approximate $p(\beta_i \mid \boldsymbol{y})$, Håvard Rue, Martino, and Chopin (2009) proceed instead by building an approximation of it based on maximizing $\boldsymbol{\beta}_{-i} \mid \beta_i, \boldsymbol{\theta}, \boldsymbol{y}$ to yield $\widehat{\boldsymbol{\beta}}_{(i)}$ whose ith element is β_i, yielding

$$\widetilde{p}(\beta_i \mid \boldsymbol{\theta}, \boldsymbol{y}) \propto \frac{p(\widehat{\boldsymbol{\beta}}_{(i)}, \boldsymbol{\theta} \mid \boldsymbol{y})}{\widetilde{p}(\widehat{\boldsymbol{\beta}}_{(i),-i} \mid \beta_i, \boldsymbol{\theta}, \boldsymbol{y})},$$

with a suitable renormalization of $\widetilde{p}(\widehat{\boldsymbol{\beta}}_{(i),-i} \mid \beta_i, \boldsymbol{\theta}, \boldsymbol{y})$. Such approximations are reminiscent of profile likelihood.

While we could use the Laplace approximation $p_G(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{y}, \boldsymbol{\theta})$ and marginalize the latter directly, this leads to evaluation of the Laplace approximation to the density far from the mode, which is often inaccurate. One challenge is that p is often very large, so calculation of the Hessian $\mathbf{H}$ is costly to evaluate. Having to evaluate it repeatedly for each marginal β_i for $i = 1, \ldots, p$ is prohibitive since it involves factorizations of $p \times p$ matrices.

To reduce the computational costs, Håvard Rue, Martino, and Chopin (2009) propose to use the approximate mean to avoid optimizing and consider the conditional based on the

conditional of the Gaussian approximation with mean $\widehat{\boldsymbol{\beta}}$ and covariance $\boldsymbol{\Sigma} = \mathbf{H}^{-1}(\widehat{\boldsymbol{\beta}})$,

$$\boldsymbol{\beta}_{-i} \mid \beta_i, \boldsymbol{\theta}, \boldsymbol{y} \approx \mathsf{Gauss}_{p-1}\left\{\widetilde{\boldsymbol{\beta}}_{(i)} = \widehat{\boldsymbol{\beta}}_{-i} + \boldsymbol{\Sigma}_{i,i}^{-1}\boldsymbol{\Sigma}_{i,-i}(\beta_i - \widehat{\beta}_i, \mathbf{M}_{-i,-i}^{-1}\right\};$$

cf. Proposition 1.6. This only requires a rank-one update. Wood (2019) suggest to use a Newton step to correct $\widetilde{\boldsymbol{\beta}}_{(i)}$, starting from the conditional mean. The second step is to exploit the local dependence on $\boldsymbol{\beta}$ using the Markov structure to build an improvement to the Hessian. Further improvements are proposed in Håvard Rue, Martino, and Chopin (2009), who used a simplified Laplace approximation to correct the Gaussian approximation for location and skewness, a necessary step when the likelihood itself is not Gaussian. This leads to a Taylor series approximation to correct the log determinant of the Hessian matrix. Wood (2019) consider a BFGS update to $\mathbf{M}_{-i,-i}^{-1}$ directly, which works less well than the Taylor expansion near $\widehat{\beta}_i$, but improves upon when we move far from this value. Nowadays, the INLA software uses a low-rank variational correction to Laplace method, proposed in van Niekerk and Rue (2024).

The `INLA` **R** package provides an interface to fit models with Gaussian latent random effects. While the software is particularly popular for spatio-temporal applications using the SPDE approach, we revisit two examples in the sequel where we can exploit the Markov structure.

Example 9.4 (Stochastic volatility model with INLA)**.** Financial returns Y_t typically exhibit time-varying variability. The **stochastic volatility** model is a parameter-driven model that specifies

$$\begin{aligned} Y_t &= \exp(h_t/2)Z_t \\ h_t &= \gamma + \phi(h_{t-1} - \gamma) + \sigma U_t \end{aligned}$$

where $U_t \overset{\text{iid}}{\sim} \mathsf{Gauss}(0,1)$ and $Z_t \overset{\text{iid}}{\sim} \mathsf{Gauss}(0,1)$. The `INLA` documentation provides information about which default prior and hyperparameters are specified. We use a $\mathsf{gamma}(1, 0.001)$ prior for the precision.

```
library(INLA)
# Stochastic volatility model
data(exchangerate, package = "hecbayes")
# Compute response from raw spot exchange rates at noon
y <- 100*diff(log(exchangerate$rate))
# 'y' is now a series of percentage of log daily differences
time <- seq_along(y)
data <- data.frame(y = y, time = time)
```

```
# Stochastic volatility model
# https://inla.r-inla-download.org/r-inla.org/doc/likelihood/stochvolgaussian.pdf
# The model uses a log link, and a (log)-gamma prior for the precision
f_stochvol <- formula(y ~ f(time, model = "ar1", param = list(prec = c(1, 0.001))))
mod_stochvol <- inla(f_stochvol, family = "stochvol", data = data)
# Obtain summary
summary <- summary(mod_stochvol)
# plot(mod_stochvol)
marg_prec <- mod_stochvol$marginals.hyperpar[[1]]
marg_phi <- mod_stochvol$marginals.hyperpar[[2]]
```

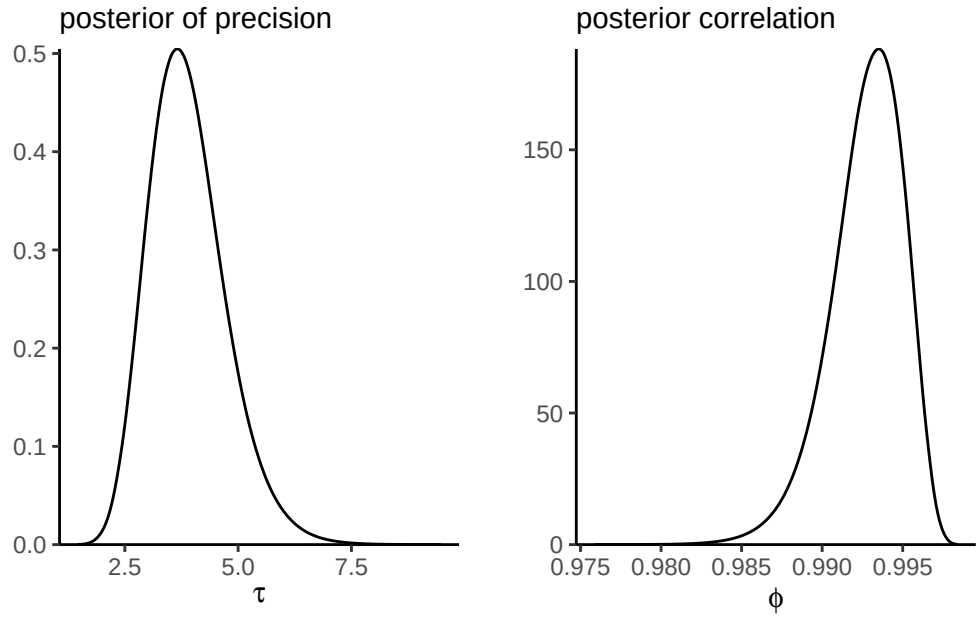

Figure 9.3: Marginal densities of precision and autocorrelation parameters from the Gaussian stochastic volatility model.

Figure 9.3 shows that the correlation ϕ is nearly one, leading to random walk behaviour and high persistence over time (this is also due to the frequency of observations). This strong serial dependence in the variance is in part responsible for the difficulty in fitting this model using MCMC.

We can use the marginal density approximations to obtain quantiles for summary of interest.

The software also includes utilities to transform the parameters using the change of variable formula.

```
# Compute density, quantiles, etc. via inla.*marginal
## approximate 95% credible interval and marginal post median
INLA::inla.qmarginal(marg_phi, p = c(0.025, 0.5, 0.975))
```

```
[1] 0.9874671 0.9929791 0.9963883
```

```
# Change of variable to get variance from precision
marg_var <- INLA::inla.tmarginal(
  fun = function(x) { 1 / x },
  marginal = marg_prec)
INLA::inla.qmarginal(marg_var, p = c(0.025, 0.975))
```

```
[1] 0.1727294 0.4049276
```

```
# Posterior marginal mean and variance of phi
mom1 <- INLA::inla.emarginal(
    fun = function(x){x},
    marginal = marg_phi)
mom2 <- INLA::inla.emarginal(
    fun = function(x){x^2},
    marginal = marg_phi)
c(mean = mom1, sd = sqrt(mom2 - mom1^2))
```

```
       mean          sd
0.992721400 0.002303183
```

Example 9.5 (Tokyo binomial time series)**.** We revisit Example 7.3, but this time fit the model with `INLA`. We specify the mean model without intercept and fit a logistic regression, with a second-order cyclic random walk prior for the coefficients, and the default priors for the other parameters.

```
data(Tokyo, package = "INLA")
# Formula (removing intercept)
formula <- y ~ f(time, model = "rw2", cyclic = TRUE) - 1
mod <- INLA::inla(
   formula = formula,
   family = "binomial",
   Ntrials = n,
   data = Tokyo)
```

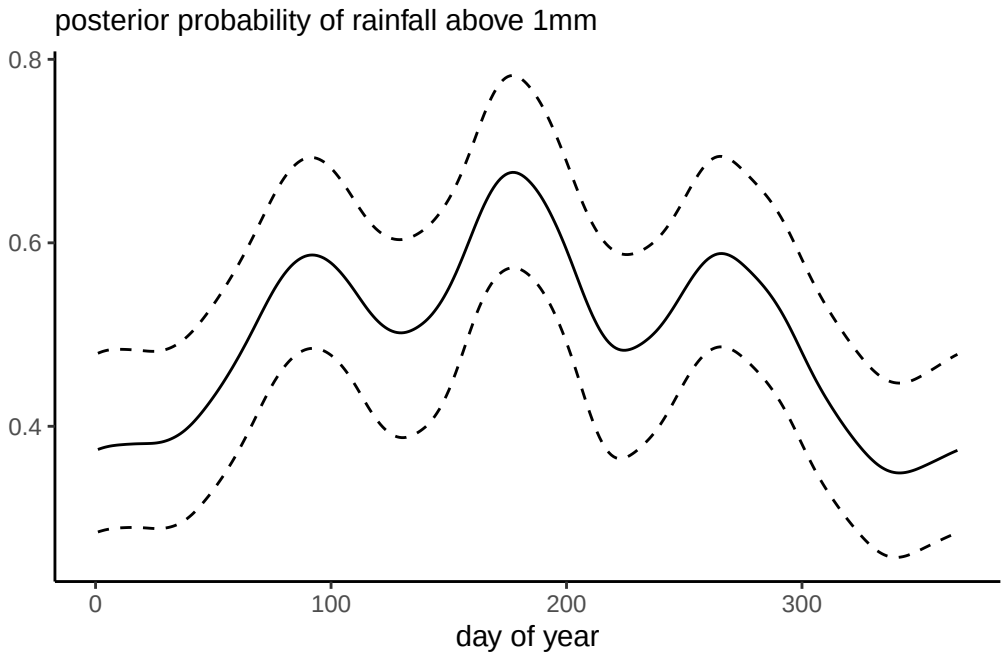

Figure 9.4: Posterior probability per day of the year with posterior median and 95% credible interval for the Tokyo rainfall binomial time series.

Figure 9.4 shows posterior summaries for the β, which align with the results for the probit model.

If we wanted to obtain predictions, we need to augment the model matrix and set missing values for the response variable. These then get imputed alongside with the other parameters.

10 Variational inference

The field of **variational inference**, which derives it's name from calculus of variation, uses approximations to a parametric distribution $p(\cdot)$ by a member from a family of distributions whose density or mass function is $g(\cdot; \boldsymbol{\psi})$ with parameters $\boldsymbol{\psi}$. The objective of inference is thus to find the parameters that minimize some metric that measure discrepancy between the true postulated posterior and the approximation: doing so leads to optimization problems. Variational inference is widespread in machine learning and in large problems where Markov chain Monte Carlo or other methods might not be feasible.

This chapter is organized as follows: we first review notions of model misspecification and the Kullback–Leibler divergence. We then consider approximation schemes and some examples involving mixtures and model selection where analytical derivations are possible: these show how variational inference differs from Laplace approximation and she light on some practical aspects. Good references include Chapter 10 of Bishop (2006); most modern application use automatic differentiation variational inference (ADVI, Kucukelbir et al. (2017)) or stochastic optimization via black-box variational inference.

10.1 Model misspecification

Suppose data were generated from a model with true density f_t and we consider a family of distributions $g(\cdot; \boldsymbol{\psi})$ not necessarily containing f_t as special case. Intuitively, if we were to estimate the model by maximum likelihood, we expect that the model returned will be the one closest to f_t among those considered in some sense.

Definition 10.1 (Kullback–Leibler divergence)**.** The Kullback–Leibler divergence between densities $f_t(\cdot)$ and $g(\cdot; \boldsymbol{\psi})$, is

$$\begin{aligned}
\mathsf{KL}(f_t \parallel g) &= \int \log\left(\frac{f_t(\boldsymbol{x})}{g(\boldsymbol{x}; \boldsymbol{\psi})}\right) f_t(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&= \int \log f_t(\boldsymbol{x}) f_t(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int \log g(\boldsymbol{x};; \boldsymbol{\psi}) f_t(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&= \mathsf{E}_{f_t}\{\log f_t(\boldsymbol{X})\} - \mathsf{E}_{f_t}\{\log g(\boldsymbol{X}; \boldsymbol{\psi})\}
\end{aligned}$$

where the subscript of the expectation indicates which distribution we integrate over. The first term $\mathsf{E}_{f_t}\{\log f_t(\boldsymbol{X})\}$ is called the entropy of the distribution. The divergence is strictly positive unless $g(\cdot;\boldsymbol{\psi}) \equiv f_t(\cdot)$. Note that, by construction, it is not symmetric.

The Kullback–Leibler divergence notion is central to study of model misspecification: if we fit $g(\cdot)$ when data arise from f_t, the maximum likelihood estimator of the parameters $\widehat{\boldsymbol{\psi}}$ will be the value of the parameter that minimizes the Kullback–Leibler divergence $\mathsf{KL}(f_t \parallel g)$; this value will be positive unless the model is correctly specified and $g(\cdot;\widehat{\boldsymbol{\psi}}) = f_t(\cdot)$. See Davison (2003), pp.122–125 for a discussion.

Definition 10.2 (Convex function)**.** A real-valued function $h : \mathbb{R} \to \mathbb{R}$ is convex if for any $x_1, x_2 \in \mathbb{R}$ if any linear combination of x_1 and x_2 satisfies

$$h(tx_1 + (1-t)x_2) \leq th(x_1) + (1-t)h(x_2), \qquad 0 \leq t \leq 1$$

The left panel of Figure 10.1 shows an illustration of the fact that the chord between any two points lies above the function. Examples include the exponential function, or a quadratic $ax^2 + bx + c$ with $a > 0$.

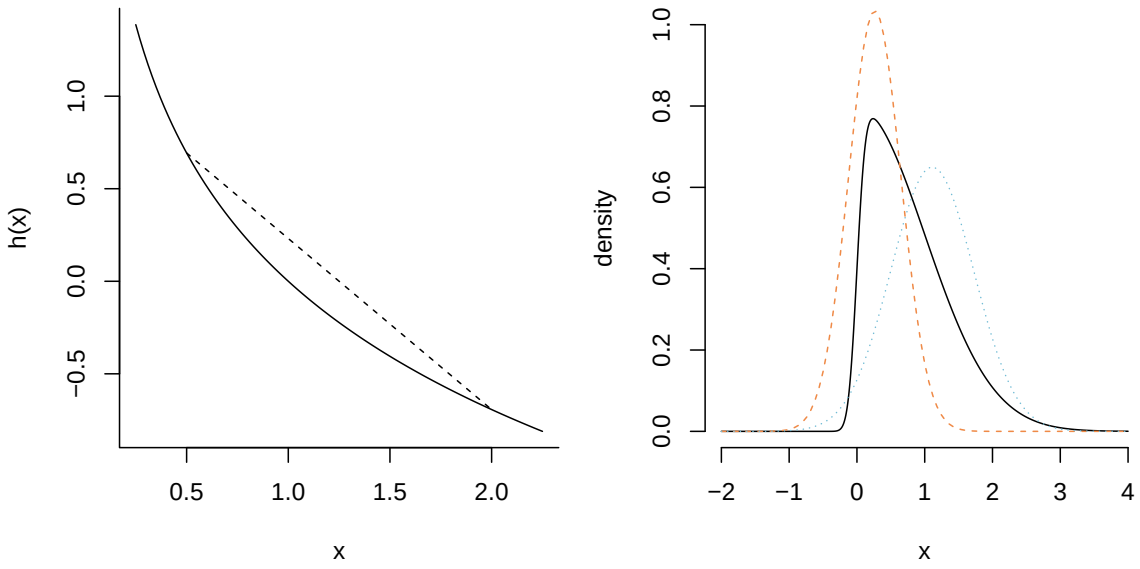

Figure 10.1: Left panel: the convex function $h(x) = -\log(x)$, with a straight line between any two points falls above the function. Right panel: a skewed density with the Laplace approximation (dashed orange) and variational Gaussian approximation (dotted blue).

Consider now the problem of approximating the marginal likelihood, sometimes called the evidence,

$$p(\boldsymbol{y}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{y}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}.$$

where we only have the joint $p(\boldsymbol{y}, \boldsymbol{\theta})$ is the product of the likelihood times the prior. The marginal likelihood is typically intractable, or very expensive to compute, but it is necessary to calculate probability and various expectations with respect to the posterior unless we draw samples from it. Consider $g(\boldsymbol{\theta}; \boldsymbol{\psi})$ with $\boldsymbol{\psi} \in \mathbb{R}^J$ an approximating density function whose integral is one over $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$ and whose support includes that of $p(\boldsymbol{y}, \boldsymbol{\theta})$ over $\boldsymbol{\Theta}$: we can then rewrite

$$p(\boldsymbol{y}) = \int_{\boldsymbol{\Theta}} \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{g(\boldsymbol{\theta}; \boldsymbol{\psi})} g(\boldsymbol{\theta}; \boldsymbol{\psi}) \mathrm{d}\boldsymbol{\theta}.$$

With convex functions, **Jensen's inequality** implies that $h\{\mathsf{E}(X)\} \leq \mathsf{E}\{h(X)\}$, and applying this with $h(x) = -\log(x)$, we get

$$-\log p(\boldsymbol{y}) \leq -\log \left(\frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{g(\boldsymbol{\theta}; \boldsymbol{\psi})} \right) g(\boldsymbol{\theta}; \boldsymbol{\psi}) \mathrm{d}\boldsymbol{\theta}.$$

We can thus consider the model that minimizes the reverse Kullback–Leibler divergence

$$g(\boldsymbol{\theta}; \widehat{\boldsymbol{\psi}}) = \mathrm{argmin}_{\boldsymbol{\psi}} \mathsf{KL}\{g(\boldsymbol{\theta}; \boldsymbol{\psi}) \parallel p(\boldsymbol{\theta}, \boldsymbol{y})\}.$$

We can get a slightly different take if we consider the reformulation

$$\mathsf{KL}\{g(\boldsymbol{\theta}; \boldsymbol{\psi}) \parallel p(\boldsymbol{\theta}, \boldsymbol{y}) = \mathsf{E}_g\{\log g(\boldsymbol{\theta})\} - \mathsf{E}_g\{\log p(\boldsymbol{y}, \boldsymbol{\theta})\} + \log p(\boldsymbol{y}).$$

Instead of minimizing the Kullback–Leibler divergence, we can equivalently maximize the so-called **evidence lower bound** (ELBO)

$$\mathsf{ELBO}(g) = \mathsf{E}_g\{\log p(\boldsymbol{y}, \boldsymbol{\theta})\} - \mathsf{E}_g\{\log g(\boldsymbol{\theta})\}$$

The ELBO as an objective function balances between two terms: the first term is the expected value of the joint posterior under the approximating density g, which will be maximized by taking a distribution placing all mass at the maximum of $p(\boldsymbol{y}, \boldsymbol{\theta})$, whereas the second term can be viewed as a penalty for the entropy of the approximating family, which rewards distributions which are diffuse. We thus try to maximize the evidence, subject to a regularization term.

The ELBO is a lower bound for the marginal likelihood because the Kullback–Leibler divergence is non-negative and

$$\log p(\boldsymbol{y}) = \mathsf{ELBO}(g) + \mathsf{KL}\{g(\boldsymbol{\theta}; \boldsymbol{\psi}) \parallel p(\boldsymbol{\theta}, \boldsymbol{y})\}.$$

If we could estimate the marginal likelihood of a (typically simpler) competing alternative and the lower bound on the evidence in favour of the more complex model was very much larger, then we could use this but generally there is no theoretical guarantee for model comparison if we compare two lower evidence lower bounds. The purpose of variational inference is that approximations to expectations, credible intervals, etc. are obtained from $g(\cdot;\boldsymbol{\psi})$ instead of $p(\cdot)$.

Remark 10.1 (Approximation of latent variables). While we have focused on exposition with only parameters $\boldsymbol{\theta}$, this can be generalized by including latent variables $\boldsymbol{U}$ as in Section 6.1 in addition to the model parameters $\boldsymbol{\theta}$ as part of the variational modelling.

Example 10.1 (Variational inference vs Laplace approximation)**.** The Laplace approximation differs from the Gaussian variational approximation; The right panel of Figure 10.1 shows a skew-Gaussian distribution with location zero, unit scale and a skewness parameter of $\alpha = 10$; it's density is $2\phi(x)\Phi(\alpha x)$.

The Laplace approximation is easily obtained by numerical maximization; the mode is the mean of the resulting approximation, with a std. deviation that matches the square root of the reciprocal Hessian of the negative log density.

Consider an approximation with g the density of $\mathsf{Gauss}(m, s^2)$; we obtain the parameters by minimizing the ELBO. The entropy term for a Gaussian approximating density is

$$-\mathsf{E}_g(\log g) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2s^2}\mathsf{E}_g\left\{(X-m)^2\right\} = \frac{1}{2}\left\{1 + \log(2\pi s^2)\right\}$$

given $\mathsf{E}_g\{(x-m)^2\} = s^2$ by definition of the variance. Ignoring constants terms that do not depend on the parameters of g, optimization of the ELBO amounts to maximization of

$$\begin{aligned}
&\mathrm{argmax}_{m,s^2}\left[-\frac{1}{2}\mathsf{E}_g\left\{\frac{(X-\mu)^2}{\sigma^2}\right\} + \mathsf{E}_g\left\{\log\Phi(\alpha X)\right\} + \log(s^2)\right] \\
&\quad = \mathrm{argmax}_{m,s^2}\left[-\frac{1}{2}\mathsf{E}_g\left\{\frac{(X-\mu)^2}{\sigma^2}\right\} + \mathsf{E}_g\left\{\log\Phi(\alpha X)\right\} + \log(s^2)\right] \\
&\quad = \mathrm{argmax}_{m,s^2}\left[-\frac{s^2+m^2-2\mu m}{2\sigma^2} + \log(s^2) + \mathsf{E}_Z\left\{\log\Phi(\alpha s X + m)\right\}\right]
\end{aligned}$$

where $Z \sim \mathsf{Gauss}(0,1)$. We can approximate the last term by Monte Carlo with a single sample (recycled at every iteration) and use this to find the optimal parameters. The right panel of Figure 10.1 shows that the resulting approximation aligns with the bulk. It of course fails to capture the asymmetry, since the approximating function is symmetric.

10.2 Optimization of the evidence lower bound

Variational inference in itself does not determine the choice of approximating density $g(\cdot;\boldsymbol{\psi})$; the quality of the approximation depends strongly on the latter. The user has ample choice to decide whether to use the fully correlated, factorized, or the mean-field approximation, along with the parametric family for each block. Note that the latter must be support dependent, as the Kullback–Leibler divergence will be infinite if the support of g does not include that of $p(\boldsymbol{\theta} \mid \boldsymbol{y})$.

There are two main approaches: the first is to start off with the model with a factorization of the density, and deduce the form of the most suitable parametric family for the approximation that will minimize the ELBO. This requires bespoke derivation of the form of the density and the conditionals for each model, and does not lead itself easily to generalizations. The second approach alternative is to rely on a generic family for the approximation, and an omnibus procedure for the optimization using a reformulation via stochastic optimization that approximates the integrals appearing in the ELBO formula.

10.2.1 Factorization

Factorizations of $g(;\boldsymbol{\psi})$ into blocks with parameters $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_M$, where

$$g(\boldsymbol{\theta};\boldsymbol{\psi}) = \prod_{j=1}^{M} q_j(\boldsymbol{\theta}_j;\boldsymbol{\psi}_j)$$

If we assume that each of the J parameters $\theta_1, \ldots, \theta_J$ are independent, then we obtain a **mean-field** approximation. The latter will be poor if parameters are strongly correlated, as we will demonstrate later.

We use a factorization of g, and denote the components $g_j(\cdot)$ for simplicity, omitting dependence on the parameters $\boldsymbol{\psi}$. We can write the ELBO as

$$\begin{aligned}
\mathsf{ELBO}(g) &= \int \log p(\boldsymbol{y},\boldsymbol{\theta}) \prod_{j=1}^{M} g_j(\boldsymbol{\theta}_j)\mathrm{d}\boldsymbol{\theta} - \sum_{j=1}^{M} \int \log\{g_j(\boldsymbol{\theta}_j)\} g_j(\boldsymbol{\theta}_j)\mathrm{d}\boldsymbol{\theta}_j \\
&= \int\int \left\{ \log p(\boldsymbol{y},\boldsymbol{\theta}) \prod_{\substack{i\neq j \\ j=1}}^{M} g_j(\boldsymbol{\theta}_j)\mathrm{d}\boldsymbol{\theta}_{-i} \right\} \mathrm{d}\boldsymbol{\theta}_{-i} g_i(\boldsymbol{\theta}_i)\mathrm{d}\boldsymbol{\theta}_i - \sum_{j=1}^{M} \int \log\{g_j(\boldsymbol{\theta}_j)\} g_j(\boldsymbol{\theta}_j)\mathrm{d}\boldsymbol{\theta}_j \\
&\overset{\boldsymbol{\theta}_i}{\propto} \int \mathsf{E}_{-i}\left\{\log p(\boldsymbol{y},\boldsymbol{\theta})\right\} g_i(\boldsymbol{\theta}_i)\mathrm{d}\boldsymbol{\theta}_i - \int \log\{g_i(\boldsymbol{\theta}_i)\} g_i(\boldsymbol{\theta}_i)\mathrm{d}\boldsymbol{\theta}_i
\end{aligned}$$

where the last line is a negative Kullback–Leibler divergence between q_j and

$$\mathsf{E}_{-i}\left\{\log p(\boldsymbol{y}, \boldsymbol{\theta})\right\} = \int \log p(\boldsymbol{y}, \boldsymbol{\theta}) \prod_{\substack{i \neq j \\ j=1}}^{M} g_j(\boldsymbol{\theta}_j) \mathrm{d}\boldsymbol{\theta}_{-i}$$

and the subscript $-i$ indicates that we consider all but the ith component of the J vector. This reveals that the form of approximating density g_i that maximizes the ELBO is of the form

$$g_j^\star(\boldsymbol{\theta}_j) \propto \exp\left[\mathsf{E}_{-i}\left\{\log p(\boldsymbol{y}, \boldsymbol{\theta})\right\}\right].$$

Often, we can look at the kernel of $g_j^\star$ and deduce the normalizing constant, which is defined as the integral of the above. The posterior approximation will have a closed form expression if we consider cases of conditionally conjugate distributions in the exponential family: we can see that the optimal $g_j^\star$ relates to the conditional since $p(\boldsymbol{\theta}, \boldsymbol{y}) \overset{\boldsymbol{\theta}_i}{\propto} p(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{-i}, \boldsymbol{y})$. This provides a connection to Gibbs sampling.

If we consider maximization of the ELBO for g_i, we can see from the law of iterated expectation that the latter is proportional to

$$\mathsf{ELBO}(g_i) \propto \mathsf{E}_i\left[\mathsf{E}_{-i}\{\log p(\boldsymbol{\theta}, \boldsymbol{y})\}\right] - \mathsf{E}_i\{\log g_i(\boldsymbol{\theta}_i)\}$$

Due to the nature of this conditional expectation, we can devide an algorithm to maximize the ELBO of the factorized approximation. Each parameter update depends on the other components, but the $\mathsf{ELBO}(q_i)$ is concave. We can maximize $g_j^\star$ in turn for each $j = 1, \ldots, M$ treating the other parameters as fixed, and iterate this scheme. The resulting approximation, termed coordinate ascent variational inference (CAVI), is guaranteed to monotonically increase the evidence lower bound until convergence to a local maximum; see Sections 3.1.5 and 3.2.4–3.2.5 of Boyd and Vandenberghe (2004). The scheme is a valid coordinate ascent algorithm. At each cycle, we compute the ELBO and stop the algorithm when the change is lower then some present numerical tolerance. Since the approximation may have multiple local optima, we can perform random initializations and keep the one with highest performance.

We consider the example from Section 2.2.2 of Ormerod and Wand (2010) (see also Example 10.1.3 from Bishop (2006)) for approximation of a Gaussian distribution with conjugate prior parametrized in terms of precision, with

$$\begin{aligned} Y_i &\sim \mathsf{Gauss}(\mu, \tau^{-1}), \qquad i = 1, \ldots, n; \\ \mu &\sim \mathsf{Gauss}(\mu_0, \tau_0^{-1}) \\ \tau &\sim \mathsf{gamma}(a_0, b_0). \end{aligned}$$

This is an example where the full posterior is available in closed-form, so we can compare our approximation with the truth. We assume a factorization of the variational approximation $g_\mu(\mu)g_\tau(\tau)$; the factor for q_μ is proportional to

$$\log g_\mu^\star(\mu) \propto -\frac{\mathsf{E}_\tau(\tau)}{2}\sum_{i=1}^n (y_i-\mu)^2 - \frac{\tau_0}{2}(\mu-\mu_0)^2$$

which is quadratic in μ and thus must be Gaussian with precision $\tau_n = \tau_0 + n\tau$ and mean $\tau_n^{-1}\{\tau_0\mu_0 + \mathsf{E}_\tau(\tau)n\overline{y}\}$ using Proposition 8.1, where $n\overline{y} = \sum_{i=1}^n y_i$. We could also note that this corresponds (up to expectation) to $p(\mu \mid \tau, \boldsymbol{y})$. As the sample size increase, the approximation converges to a Dirac delta (point mass at the sample mean. The optimal precision factor satisfies

$$\ln g_\tau^\star(\tau) \propto (a_0 - 1 + n/2)\log\tau - \tau\left[b_0 + \frac{1}{2}\mathsf{E}_\mu\left\{\sum_{i=1}^n (y_i-\mu)^2\right\}\right].$$

This is the form as $p(\tau \mid \mu, \boldsymbol{y})$, namely a gamma with shape $a_n = a_0 + n/2$ and rate b_n given by the term in the square brackets. It is helpful to rewrite the expected value as

$$\mathsf{E}_\mu\left\{\sum_{i=1}^n (y_i-\mu)^2\right\} = \sum_{i=1}^n \{y_i - \mathsf{E}_\mu(\mu)\}^2 + n\mathsf{Var}_\mu(\mu),$$

so that it depends on the parameters of the distribution of μ directly. We can then apply the coordinate ascent algorithm. Derivation of the ELBO, even in this toy setting, is tedious:

$$\begin{aligned}\mathsf{ELBO}(g) = {} & a_0\log(b_0) - \log\Gamma(a_0) - \frac{n+1}{2}\log(2\pi) + \frac{\log(\tau_0)}{2} \\ & + (a_n-1)\mathsf{E}_\tau(\log\tau) - \frac{\mathsf{E}_\tau(\tau)\left[b_0 + \mathsf{E}_\mu\left\{\sum_{i=1}^n (y_i-\mu)^2\right\}\right]}{2} - \frac{\tau_0}{2}\mathsf{E}_\mu\{(\mu-\mu_0)^2\} \\ & + \frac{1+\log(2\pi) - \log\tau_n}{2} - a_n\log b_n - \log\Gamma(a_n) - (a_n-1)\mathsf{E}_\tau(\log\tau) - b_n\mathsf{E}_\tau(\tau)\end{aligned}$$

The expected value of $\mathsf{E}_\tau(\tau) = a_n/b_n$ and the mean and variance of the Gaussian are given by it's parameters. The terms involving $\mathsf{E}_\tau(\log\tau)$ cancel out; the first line involves only normalizing constants for the hyperparameters, and a_n is constant. We can keep track only of

$$-\frac{\tau_0}{2}\mathsf{E}_\mu\{(\mu-\mu_0)^2\} - \frac{\log\tau_n}{2} - a_n\log b_n$$

for convergence, although other normalizing constants would be necessary if we wanted to approximate the marginal likelihood.

10.2.2 General derivation

We consider alternative numeric schemes which rely on stochastic optimization. The key idea behind these methods is that we can use gradient-based algorithms, and approximate the expectations with respect to g by drawing samples from the approximating densities. This gives rises to a general omnibus procedure for optimization, although some schemes capitalize on the structure of the approximating family. Hoffman et al. (2013) consider stochastic gradient for exponential families mean-field approximations, using natural gradients to device an algorithm. While efficient within this context, it is not a generic algorithm.

Ranganath, Gerrish, and Blei (2014) extend this to more general distributions by noting that the gradient of the ELBO, interchanging the derivative and the integral using the dominated convergence theorem, is

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\psi}}\mathsf{ELBO}(g) &= \int g(\boldsymbol{\theta};\boldsymbol{\psi})\frac{\partial}{\partial \boldsymbol{\psi}}\log\left(\frac{p(\boldsymbol{\theta},\boldsymbol{y})}{g(\boldsymbol{\theta};\boldsymbol{\psi})}\right)\mathrm{d}\boldsymbol{\theta} + \int \frac{\partial g(\boldsymbol{\theta};\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \times \log\left(\frac{p(\boldsymbol{\theta},\boldsymbol{y})}{g(\boldsymbol{\theta};\boldsymbol{\psi})}\right)\mathrm{d}\boldsymbol{\theta} \\ &= \int \frac{\partial \log g(\boldsymbol{\theta};\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \times \log\left(\frac{p(\boldsymbol{\theta},\boldsymbol{y})}{g(\boldsymbol{\theta};\boldsymbol{\psi})}\right) g(\boldsymbol{\theta};\boldsymbol{\psi})\mathrm{d}\boldsymbol{\theta}\end{aligned}$$

where $p(\boldsymbol{\theta},\boldsymbol{y})$ does not depend on $\boldsymbol{\psi}$, and

$$\begin{aligned}\int \frac{\partial \log g(\boldsymbol{\theta};\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} g(\boldsymbol{\theta};\boldsymbol{\psi})\mathrm{d}\boldsymbol{\theta} &= \int \frac{\partial g(\boldsymbol{\theta};\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\mathrm{d}\boldsymbol{\theta} \\ \frac{\partial}{\partial \boldsymbol{\psi}}\int g(\boldsymbol{\theta};\boldsymbol{\psi})\mathrm{d}\boldsymbol{\theta} &= 0.\end{aligned}$$

This term vanishes since the integral of a density is one regardless of the value of $\boldsymbol{\psi}$, so it's derivative vanishes. We are left with the expected value

$$\mathsf{E}_g\left\{\frac{\partial \log g(\boldsymbol{\theta};\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \times \log\left(\frac{p(\boldsymbol{\theta},\boldsymbol{y})}{g(\boldsymbol{\theta};\boldsymbol{\psi})}\right)\right\}$$

which we can approximate via Monte Carlo by drawing samples from g, which gives a stochastic gradient algorithm. Ranganath, Gerrish, and Blei (2014) provide two methods to reduce the variance of this expression using control variates and Rao–Blackwellization, as excessive variance hinders the convergence and requires larger Monte Carlo sample sizes to be reliable. In the context of large samples of independent observations, we can also resort to mini-batching, by randomly selecting a subset of observations.

Some families of distributions, notably location-scale (cf. Definition 1.12) and exponential families (Definition 1.13) are particularly convenient, because we can get expressions for the ELBO that are simpler. For exponential families approximating distributions, we have sufficient statistics $S_k \equiv t_k(\boldsymbol{\theta})$ and the gradient of $\log g$ becomes S_k under mean-field.

Kucukelbir et al. (2017) proposes a stochastic gradient algorithm, but with two main innovations. The first is the general use of Gaussian approximating densities for factorized density, with parameter transformations to map from the support of $T : \boldsymbol{\Theta} \mapsto \mathbb{R}^p$ via $T(\boldsymbol{\theta}) = \boldsymbol{\zeta}$. We then consider an approximation $g(\boldsymbol{\zeta}; \boldsymbol{\psi})$ where $\boldsymbol{\psi}$ consists of mean parameters $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, parametrized in terms of independent components via $\boldsymbol{\Sigma} = \mathsf{diag}\{\exp(\boldsymbol{\omega})\}$ or through a Cholesky decomposition of the covariance $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$, where $\mathbf{L}$ is a lower triangular matrix. The full approximation is of course more flexible when the transformed parameters $\boldsymbol{\zeta}$ are correlated, but is more expensive to compute than the mean-field approximation. The change of variable introduces a Jacobian term for the approximation to the density $p(\boldsymbol{\theta}, \boldsymbol{y})$. Another benefit is that the entropy of the multivariate Gaussian for g the density of $\mathsf{Gauss}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\begin{aligned}
-\mathsf{E}_g(\log g) &= \frac{D\log(2\pi) + \log|\boldsymbol{\Sigma}|}{2} - \frac{1}{2}\mathsf{E}_g\left\{(\boldsymbol{X}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right\} \\
&= \frac{D\log(2\pi) + \log|\boldsymbol{\Sigma}|}{2} - \frac{1}{2}\mathsf{E}_g\left[\mathsf{tr}\left\{(\boldsymbol{X}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right\}\right] \\
&= \frac{D\log(2\pi) + \log|\boldsymbol{\Sigma}|}{2} - \frac{1}{2}\mathsf{tr}\left[\boldsymbol{\Sigma}^{-1}\mathsf{E}_g\left\{(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\top\right\}\right] \\
&= \frac{D + D\log(2\pi) + \log|\boldsymbol{\Sigma}|}{2}.
\end{aligned}$$

This follows from taking the trace of a 1×1 matrix, and applying a cyclic permutation to which the trace is invariant. Since the trace is a linear operator, we can interchange the trace with the expected value. Finally, we have $\mathsf{E}_g\left\{(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\top\right\} = \boldsymbol{\Sigma}$, and the trace of a $D \times D$ identity matrix is simply D. The Gaussian entropy depends only on the covariance, so is not random. The transformation to $\mathbb{R}^p$ is not unique and different choices may yield to differences, but the choice of optimal transformation requires knowledge of the true posterior, which is thus intractable.

We focus on the case of full transformation; the derivation for independent components is analogous. Since the Gaussian is a location-scale family, we can rewrite the model in terms of a standardized Gaussian variable $\boldsymbol{Z} \sim \mathsf{Gauss}_p(\mathbf{0}_p, \mathbf{I}_p)$ where $\boldsymbol{\zeta} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{Z}$ The ELBO with the transformation is of the form

$$\mathsf{E}_{\boldsymbol{Z}}\left\{\log p(\boldsymbol{y}, T^{-1}(\boldsymbol{\mu} + \mathbf{L}\boldsymbol{Z}) + \log|\mathbf{J}_{T^{-1}}(\boldsymbol{\mu} + \mathbf{L}\boldsymbol{\eta})|\right\} + \frac{D + D\log(2\pi) + \log|\mathbf{L}\mathbf{L}^\top|}{2}.$$

where we have the absolute value of the determinant of the Jacobian of the transformation. If we apply the chain rule

$$\frac{\partial}{\partial\boldsymbol{\psi}}\log p(\boldsymbol{y}, T^{-1}(\boldsymbol{\mu} + \mathbf{L}\boldsymbol{Z}) = \frac{\partial\log p(\boldsymbol{y},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial T^{-1}(\boldsymbol{\zeta})}{\partial\boldsymbol{\zeta}}\frac{\partial\boldsymbol{\mu} + \mathbf{L}\boldsymbol{Z}}{\partial\boldsymbol{\psi}}$$

and the derivative with respect to and we retrieve the gradients of the ELBO with respect to the mean and variance, which are

$$\frac{\partial \mathsf{ELBO}(g)}{\partial \boldsymbol{\mu}} = \mathsf{E}_{\boldsymbol{Z}} \left\{ \frac{\partial \log p(\boldsymbol{y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial T^{-1}(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}} + \frac{\partial \log |\mathbf{J}_{T^{-1}}(\boldsymbol{\zeta})|}{\partial \boldsymbol{\zeta}} \right\}$$
$$\frac{\partial \mathsf{ELBO}(g)}{\partial \mathbf{L}} = \mathsf{E}_{\boldsymbol{Z}} \left[\left\{ \frac{\partial \log p(\boldsymbol{y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial T^{-1}(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}} + \frac{\partial \log |\mathbf{J}_{T^{-1}}(\boldsymbol{\zeta})|}{\partial \boldsymbol{\zeta}} \right\} \boldsymbol{Z}^\top \right] + \mathbf{L}^{-\top}.$$

Compared to the black-box variational inference algorithm, this requires calculating the gradient of the log posterior with respect to $\boldsymbol{\theta}$. This step can be done using automatic differentiation (hence the terminology ADVI), and moreover this gradient estimator is several orders less noisy than the black-box counterpart. The ELBO can be approximated via Monte Carlo integration.

We can thus build a stochastic gradient algorithm with a Robins–Munroe sequence of updates. Kucukelbir et al. (2017) use an adaptive step-size for convergence. The ADVI algorithm is implemented in Carpenter et al. (2017); see the manual for more details.

11 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 Gareth O. Roberts. 2009. "The Pseudo-Marginal Approach for Efficient Monte Carlo Computations." *The Annals of Statistics* 37 (2): 697–725. https://doi.org/10.1214/07-AOS574.

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.

Beaumont, Mark A. 2003. "Estimation of Population Growth or Decline in Genetically Monitored Populations." *Genetics* 164 (3): 1139–60. https://doi.org/10.1093/genetics/164.3.1139.

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. New York, NY: Springer. https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/.

Bolin, David, Alexandre B. Simas, and Zhen Xiong. 2023. "Wasserstein Complexity Penalization Priors: A New Class of Penalizing Complexity Priors." *arXiv e-Prints*, arXiv:2312.04481. https://doi.org/10.48550/arXiv.2312.04481.

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.

Box, G. E. P., and D. R. Cox. 1964. "An Analysis of Transformations." *Journal of the Royal Statistical Society: Series B (Methodological)* 26 (2): 211–43. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x.

Boyd, Stephen, and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511804441.

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.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1): 1–32. https://doi.org/10.18637/jss.v076.i01.

Carvalho, Carlos M., Nicholas G. Polson, and James G. Scott. 2010. "The Horseshoe Estimator for Sparse Signals." *Biometrika* 97 (2): 465–80. https://doi.org/10.1093/biomet/asq017.

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.

Davison, A. C. 2003. *Statistical Models*. Cambridge, UK: Cambridge University Press.

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

Duke, Kristen E., and On Amir. 2023. "The Importance of Selling Formats: When Integrating Purchase and Quantity Decisions Increases Sales." *Marketing Science* 42 (1): 87–109. https://doi.org/10.1287/mksc.2022.1364.

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.

Eaton, Morris L. 2007. *Multivariate Statistics: A Vector Space Approach*. Institute for Mathematical Statistics. https://doi.org/10.1214/lnms/1196285102.

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, and Donald B. Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science* 7 (4): 457–72. https://doi.org/10.1214/ss/1177011136.

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.

George, Edward I., and Robert E. McCulloch. 1993. “Variable Selection via Gibbs Sampling.” *Journal of the American Statistical Association* 88 (423): 881–89. https://doi.org/10.1080/01621459.1993.10476353.

Geweke, John. 1992. “Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments.” In *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting, Dedicated to the Memory of Morris h. DeGroot, 1931–1989*. Oxford University Press. https://doi.org/10.1093/oso/9780198522669.003.0010.

———. 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/b10905-3.

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. 1995. “Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination.” *Biometrika* 82 (4): 711–32. https://doi.org/10.1093/biomet/82.4.711.

———. 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.

Held, Leonhard, and Daniel Sabanés Bové. 2020. *Likelihood and Bayesian Inference: With Applications in Biology and Medicine*. 2nd ed. Heidelberg: Springer Berlin. https://doi.org/10.1007/978-3-662-60792-3.

Hobert, James. 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, 253–93. Boca Raton: CRC Press. https://doi.org/10.1201/b10905-11.

Hoffman, Matthew D., David M. Blei, Chong Wang, and John Paisley. 2013. “Stochastic Variational Inference.” *Journal of Machine Learning Research* 14 (40): 1303–47. http://jmlr.org/papers/v14/hoffman13a.html.

Holmes, C. C., D. G. T. Denison, and B. K. Mallick. 2002. “Accounting for Model Uncertainty in Seemingly Unrelated Regressions.” *Journal of Computational and Graphical Statistics* 11 (3): 533–51. http://www.jstor.org/stable/1391112.

Jasra, A., C. C. Holmes, and D. A. Stephens. 2005. “Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling.” *Statistical Science* 20 (1):

50–67. https://doi.org/10.1214/088342305000000016.
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: 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.
Kitagawa, Genshiro. 1987. "Non-Gaussian State—Space Modeling of Nonstationary Time Series." *Journal of the American Statistical Association* 82 (400): 1032–41. https://doi.org/10.1080/01621459.1987.10478534.
Kucukelbir, Alp, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. 2017. "Automatic Differentiation Variational Inference." *Journal of Machine Learning Research* 18 (14): 1–45. http://jmlr.org/papers/v18/16-107.html.
Lewandowski, Daniel, Dorota Kurowicka, and Harry Joe. 2009. "Generating Random Correlation Matrices Based on Vines and Extended Onion Method." *Journal of Multivariate Analysis* 100 (9): 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008.
Lin, Jason D, Nicole You Jeung Kim, Esther Uduehi, and Anat Keinan. 2024. "Culture for Sale: Unpacking Consumer Perceptions of Cultural Appropriation." *Journal of Consumer Research*. https://doi.org/10.1093/jcr/ucad076.
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.
Martins, Eduardo S., and Jery R. Stedinger. 2000. "Generalized Maximum-Likelihood Generalized Extreme-Value Quantile Estimators for Hydrologic Data." *Water Resources Research* 36 (3): 737–44. https://doi.org/10.1029/1999WR900330.
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.
Mitchell, T. J., and J. J. Beauchamp. 1988. "Bayesian Variable Selection in Linear Regression." *Journal of the American Statistical Association* 83 (404): 1023–32. https://doi.org/10.1080/01621459.1988.10478694.
Nadarajah, Saralees. 2008. "Marshall and Olkin's Distributions." *Acta Applicandae Mathe-*

maticae 103 (1): 87–100. https://doi.org/10.1007/s10440-008-9221-7.

Neal, Radford M. 2011. "MCMC Using Hamiltonian Dynamics." In *Handbook of Markov Chain Monte Carlo*, edited by S. Brooks, A. Gelman, G. Jones, and X. L. Meng, 113–62. Boca Raton: CRC Press. https://doi.org/10.1201/b10905-5.

Northrop, Paul J. 2024. *rust: Ratio-of-Uniforms Simulation with Transformation*. https://doi.org/10.32614/CRAN.package.rust.

Northrop, Paul J., and Nicolas Attalides. 2016. "Posterior Propriety in Bayesian Extreme Value Analyses Using Reference Priors." *Statistica Sinica* 26 (2): 721–43. https://doi.org/10.5705/ss.2014.034.

Nychka, Douglas, Soutir Bandyopadhyay, Dorit Hammerling, Finn Lindgren, and Stephan Sain. 2015. "A Multiresolution Gaussian Process Model for the Analysis of Large Spatial Datasets." *Journal of Computational and Graphical Statistics* 24 (2): 579–99.

Ormerod, J. T., and M. P. Wand. 2010. "Explaining Variational Approximations." *The American Statistician* 64 (2): 140–53. https://doi.org/10.1198/tast.2010.09058.

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.

Peskun, P. H. 1973. "Optimum Monte-Carlo Sampling Using Markov Chains." *Biometrika* 60 (3): 607–12. https://doi.org/10.1093/biomet/60.3.607.

Piironen, Juho, and Aki Vehtari. 2017. "Sparsity Information and Regularization in the Horseshoe and Other Shrinkage Priors." *Electronic Journal of Statistics* 11 (2): 5018–51. https://doi.org/10.1214/17-ejs1337si.

Plummer, Martyn, Nicky Best, Kate Cowles, and Karen Vines. 2006. "CODA: Convergence Diagnosis and Output Analysis for MCMC." *R News* 6 (1): 7–11. https://doi.org/10.32614/CRAN.package.coda.

Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." *Sociological Methodology* 25: 111–63. https://doi.org/10.2307/271063.

Ranganath, Rajesh, Sean Gerrish, and David Blei. 2014. "Black Box Variational Inference." In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, edited by Samuel Kaski and Jukka Corander, 33:814–22. Proceedings of Machine Learning Research. Reykjavik, Iceland: PMLR. https://proceedings.mlr.press/v33/ranganath14.html.

Robert, Christian P., and George Casella. 2004. *Monte Carlo Statistical Methods*. 2nd ed. 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.

Rue, Håvard, Sara Martino, and Nicolas Chopin. 2009. "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71 (2): 319–92. https://doi.org/10.1111/j.1467-9868.2008.00700.x.

Rue, H., and L. Held. 2005. *Gaussian Markov Random Fields: Theory and Applications*.

Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Boca Raton: CRC Press.

Säilynoja, Teemu, Paul-Christian Bürkner, and Aki Vehtari. 2022. "Graphical Test for Discrete Uniformity and Its Applications in Goodness-of-Fit Evaluation and Multiple Sample Comparison." *Statistics and Computing* 32 (2): 32. https://doi.org/10.1007/s11222-022-10090-6.

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.

Smith, Richard L. 1985. "Maximum Likelihood Estimation in a Class of Nonregular Cases." *Biometrika* 72 (1): 67–90. https://doi.org/10.1093/biomet/72.1.67.

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.

Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin, and Angelika Linde. 2014. "The Deviance Information Criterion: 12 Years On." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 76 (3): 485–93. https://doi.org/10.1111/rssb.12062.

Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde. 2002. "Bayesian Measures of Model Complexity and Fit." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64 (4): 583–639. https://doi.org/10.1111/1467-9868.00353.

Stephens, Matthew. 2002. "Dealing with Label Switching in Mixture Models." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 62 (4): 795–809. https://doi.org/10.1111/1467-9868.00265.

Talts, Sean, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. 2020. "Validating Bayesian Inference Algorithms with Simulation-Based Calibration." https://doi.org/10.48550/arXiv.1804.06788.

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.

Tierney, Luke, and Joseph B. Kadane. 1986. "Accurate Approximations for Posterior Moments and Marginal Densities." *Journal of the American Statistical Association* 81 (393): 82–86. https://doi.org/10.1080/01621459.1986.10478240.

van Niekerk, Janet, and Håavard Rue. 2024. "Low-Rank Variational Bayes Correction to the Laplace Method." *Journal of Machine Learning Research* 25 (62): 1–25. http://jmlr.org/papers/v25/21-1405.html.

Villani, Mattias. 2023. "Bayesian Learning: A Gentle Introduction." https://mattiasvillani.com/BayesianLearningBook/.

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.

Wood, Simon N. 2019. “Simplified Integrated Nested Laplace Approximation.” *Biometrika* 107 (1): 223–30. https://doi.org/10.1093/biomet/asz044.

Zellner, Arnold. 1971. *An Introduction to Bayesian Inference in Econometrics*. Wiley.

———. 1986. “On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions.” In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 233–43. North-Holland/Elsevier.