1 Appendix: Common distributions

This Appendix provides details for common univariate and multivariate distributions, including definitions, moments, and simulation. Deyroye (1986) provides a complete treatment of random number generation, although care must be taken as many distributions can be parameterized in different ways.

Uniform

- A random variable X has a uniform distribution on the interval $[\alpha, \beta]$, denoted $\mathcal{U}(\alpha, \beta)$, if the probability density function (pdf) is

$$p(x|\alpha, \beta) = \frac{1}{\beta - \alpha}$$

  for $x \in [\alpha, \beta]$ and 0 otherwise. The mean and variance of a uniform random variable are $E(X) = \frac{\alpha+\beta}{2}$ and $var(X) = \frac{(\beta-\alpha)^2}{12}$, respectively.

- The uniform distribution plays a foundational role in random number generation. In particular, uniform random numbers are required for the inverse transform simulation method, accept-reject algorithms, and the Metropolis algorithm. Fast and accurate pre-programmed algorithms are available in most statistical software packages and programming languages.

Bernoulli

- A random variable X has a Bernoulli distribution with parameter θ, denoted $X \sim \mathcal{B}er(\theta)$ if the probability mass function (pmf) is

$$\text{Prob}(X = x|\theta) = \theta^x (1 - \theta)^{1-x}.$$

  for $x \in \{0, 1\}$. The mean and variance of a Bernoulli random variable are $E(X) = \theta$ and $var(X) = \theta(1 - \theta)$, respectively.

- To simulate $X \sim \mathcal{B}er(\theta)$,

  1. Draw $U \sim \mathcal{U}(0,1)$
  2. Set $X = 1$ if $U < \theta$.

Binomial

- A random variable $X \in \{1, ..., n\}$ has a Binomial distribution with parameters n and θ, denoted, $X \sim \mathcal{B}in(n, \theta)$, if the pmf is

$$\text{Prob}(X = x|n, \theta) = \frac{n!}{x!(n-x)!}\theta^x (1-\theta)^{n-x},$$

  where $n! = n(n-1)! = n(n-1)\cdots 2 \cdot 1$. The mean and variance of a Binomial random variable are $E(X) = n\theta$ and $var(X) = n\theta(1-\theta)$, respectively. The Binomial distribution arises as the distribution of a sum of n independent Bernoulli trials. The Binomial is closely related to a number of other distributions. If $W_1, ..., W_n$ are i.i.d. $\mathcal{B}er(p)$, then $\sum_{i=1}^n W_i \sim \mathcal{B}in(n, p)$. As $n \to \infty$ with $np = \lambda$, $X \sim \mathcal{B}in(n, p)$ converges in distribution to a Poisson distribution with parameter λ.

- To simulate $X \sim \mathcal{B}in(n, \theta)$,

  1. Draw $X_1, ..., X_n$ independently $X_i \sim \mathcal{B}er(\theta)$
  2. Set $X = count(X_i = 1)$.

Multinomial

- A vector of random variables $X = (X_1, ..., X_k)$ has a Multinomial distribution, denoted $X \sim Mult(n, p_1, ..., p_k)$, if

$$p(X = x|p_1, ..., p_k) = \frac{n!}{x_1! \cdots x_k!} \prod_{i=1}^k p_i^{x_i}$$

  where $\sum_{i=1}^k x_i = n$. The Multinomial distribution is a natural extension of the Bernoulli and Binomial distributions. The Bernoulli distribution gives a single trial resulting in success or failure. The Binomial distribution is an extension that involves

n independently repeated Bernoulli trials. The Multinomial allows for multiple outcomes, instead of the two outcomes in the Binomial distribution. There are still n total trials, but now the outcome of each trial is assigned into one of k categories, and x_i counts the number of outcomes in category i. The probability of category i is p_i. The mean, variance, and covariances of the Multinomial distribution are given by

$$E(X_i) = np_i,\ var(X_i) = np_i(1-p_i), \text{ and } cov(X_i, X_j) = -np_ip_j$$

Multinomial distributions are often used in modeling finite mixture distributions where the Multinomial random variables represent the various mixture components.

Dirichlet

- A vector of random variables $X = (X_1, .., X_k)$ has a Dirichlet distribution, denoted $X \sim \mathcal{D}(\alpha_1, ..., \alpha_k)$, if $\sum_{i=1}^{k} X_i = 1$

$$p(x|\alpha_1, ..., \alpha_k) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}$$

The Dirichlet distribution is used as a prior for mixture probabilities in mixture models. The mean, variance, and covariances of the Multinomial distribution are

$$\begin{aligned} E(X_i) &= \frac{\alpha_i}{\sum_{i=1}^{k} \alpha_i} \\ var(X_i) &= \frac{\alpha_i \sum_{j \neq i} \alpha_j}{\left(\sum_{i=1}^{k} \alpha_i\right)^2 \left(\sum_{i=1}^{k} \alpha_i + 1\right)} \\ cov(X_i, X_j) &= \frac{-\alpha_i \alpha_j}{\left(\sum_{i=1}^{k} \alpha_i\right)^2 \left(\sum_{i=1}^{k} \alpha_i + 1\right)} \end{aligned}$$

- To simulate a Dirichlet $X = (X_1, \ldots, X_k) \sim \mathcal{D}(\alpha_1, ..., \alpha_k)$ use the two step procedure

Step 1: Draw k independent Gammas $Y_i \sim \Gamma(\alpha_i, 1)$

Step 2: Set $X_i = Y_i / \sum_{i=1}^{k} Y_i$

Poisson

- A random variable $X \in \mathbb{N}_+$ (the non-negative integers) has a Poisson distribution with parameter λ, denoted $X \sim \text{Poi}(\lambda)$, if the pmf is

$$\text{Prob}(X = x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}.$$

  The mean and variance of a Poisson random variable are $E(X) = \lambda$ and $var(X) = \lambda$, respectively.

- To simulate $X \sim \text{Poi}(\lambda)$,

1. Draw $Z_1, \ldots, Z_n$ independently, $Z_i \sim \exp(1)$
2. Set $X = \inf\left\{n > 0 : \sum_{i=1}^{n} Z_i > \lambda\right\}$,

Exponential

- A random variable $X \in \mathbb{R}^+$ has an exponential distribution with parameter μ, denoted, $X \sim \exp(\mu)$, if the pdf is

$$p(x|\mu) = \frac{1}{\mu}\exp\left(-\frac{x}{\mu}\right).$$

  The mean and variance of an exponential random variable are $E(X) = \mu$ and $var(X) = \mu^2$, respectively.

- The inverse transform method is the easiest way to simulate exponential random variables, since the cumulative distribution function (cdf) is $F(x) = 1 - e^{-\frac{x}{\mu}}$.

- To simulate $X \sim \exp(\mu)$,

1. Draw $U \sim \mathcal{U}[0, 1]$
2. Set $X = -\mu \ln(1 - U)$.

Gamma

- A random variable $X \in \mathbb{R}^+$ has a Gamma distribution with parameters α and β, denoted $X \sim \mathcal{G}(\alpha, \beta)$, if the pdf is

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x),$$

The mean and variance of a Gamma random variable are $E(X) = \alpha\beta^{-1}$ and $var(X) = \alpha\beta^{-2}$, respectively. It is important to realise that there are different parameterizations of the Gamma distribution; e.g. MATLAB parameterizes the Gamma density as

$$p(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp(-x/\beta).$$

If $X = Y/\beta$ and $Y \sim \mathcal{G}(\alpha, 1)$ then $X \sim \mathcal{G}(\alpha, \beta)$. To see this, use the inverse transform $Y = \beta X$ with $dY/dX = \beta$, which implies that

$$p(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)} (x\beta)^{\alpha-1} \exp(-\beta x)\beta = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x),$$

the density of a $\mathcal{G}(\alpha, \beta)$ random variable. The exponential distribution is a special case of the Gamma distribution when $\alpha = 1$: $X \sim \mathcal{G}(1, \mu^{-1})$ implies that $X \sim \exp(\mu)$.

- Gamma random variable simulation is standard, with built-in generators in most software packages. These algorithms typically use accept/reject algorithms that are customized to the specific values of α and β. To simulate $X \sim \mathcal{G}(\alpha, \beta)$, when α is integer-valued,

1. Draw $X_1, ..., X_\alpha$ independently $X_i \sim exp(1)$
2. Set $X = \beta \sum_{i=1}^{\alpha} X_i$.

For non-integer α, accept-reject methods provide fast and accurate algorithms for Gamma simulation.

Beta

- A random variable $X \in [0, 1]$ has a Beta distribution with parameters α and β,

denoted $X \sim \mathcal{B}(\alpha, \beta)$, if the pdf is

$$p(x|\alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)},$$

where $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ is the Beta function. As $\int p(x|\alpha, \beta)\, dx = 1$ we have $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx$.

The mean and variance of a Beta random variable are

$$E(X) = \frac{\alpha}{\alpha+\beta} \text{ and } var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},$$

respectively. If $\alpha = \beta = 1$, then $X \sim \mathcal{U}(0, 1)$.

- If α and β are integers, to simulate $X \sim \mathcal{B}(\alpha, \beta)$,

  1. Draw $X_1 \sim \mathcal{G}(\alpha, 1)$ and $X_2 \sim \mathcal{G}(\beta, 1)$
  2. Set $X = \frac{X_1}{X_1 + X_2}$.

  When α is integer valued and β is non-integer valued (a common case in Bayesian inference), exponential random variables can be used to simulate beta random variables via the transformation method.

Chi-squared

- A random variable $X \in \mathbb{R}_+$ has a Chi-squared distribution with parameter ν, denoted $X \sim \mathcal{X}^2_\nu$ if the pdf is

$$p(x|\nu) = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} x^{\frac{\nu}{2}-1} \exp\left(-\frac{x}{2}\right).$$

  The mean and variance of X are $E(X) = \nu$ and $var(X) = 2\nu$, respectively. The $\mathcal{X}^2_\nu$-distribution is a special case of the Gamma distribution: $\mathcal{X}^2_\nu = \mathcal{G}\left(\frac{\nu}{2}, \frac{1}{2}\right)$.

- Simulating chi-squared random variables typically uses the transformation method. For integer values of ν, the following two-step procedure simulates a $\mathcal{X}^2_\nu$ random

variable:

$$\text{Step 1: Draw } Z_1, ..., Z_\nu \text{ independently } Z_i \sim \mathcal{N}(0,1)$$
$$\text{Step 2: Set } X = \sum_{i=1}^{\nu} Z_i^2.$$

When ν is large, simulating using normal random variables is computationally costly and alternative more computationally efficient algorithms use Gamma random variable generation.

Inverse Gamma

- A random variable $X \in \mathbb{R}_+$ has a inverse Gamma distribution, denoted by $X \sim \mathcal{IG}(\alpha, \beta)$, if the pdf is

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\left(-\frac{\beta}{x}\right).$$

The mean and variance of the inverse Gamma distribution for $\alpha > 2$ are

$$E(X) = \frac{\beta}{\alpha - 1} \text{ and } var(X) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$$

If $Y \sim \mathcal{G}(\alpha, \beta)$, then $X = Y^{-1} \sim \mathcal{IG}(\alpha, \beta)$. To see this, write

$$\begin{aligned} 1 &= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp(-\beta y)\, dy = \int_\infty^0 \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{x}\right)^{\alpha-1} \exp\left(-\frac{\beta}{x}\right) \left(\frac{-1}{x^2}\right) dx \\ &= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{x^{\alpha+1}} \exp\left(-\frac{\beta}{x}\right) dx. \end{aligned}$$

- The following two-steps simulate an $\mathcal{IG}(\alpha, \beta)$

$$\text{Step 1: Draw } Y \sim \mathcal{G}(\alpha, 1)$$
$$\text{Step 2: Set } X = \frac{\beta}{Y}.$$

Again, as in the case of the Gamma distribution, some authors use a different parameterization for this distribution, so it is important to be careful to make sure you are

drawing using the correct parameters. In the case of prior distributions over the scale, σ^2, it is additional complicated because some authors (Zellner, 1971) parameterize σ instead of σ^2.

Pareto

- A random variable $X \in \mathbb{R}^+$ has a Pareto distribution, denoted by $Par(\alpha, \beta)$, if the pdf for $x > \beta$ and $\alpha > 0$ is

$$p(x|\alpha, \beta) = \frac{\alpha\beta^\alpha}{x^{\alpha+1}}$$

  The mean and variance are $E(X) = \frac{\alpha\beta}{\alpha-1}$ if $\alpha > 1$ and $var(X) = \frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)}$ if $\alpha > 2$.

- The following two-steps simulate a $Par(\alpha, \beta)$ distribution

$$\begin{aligned} &\text{Step 1: Draw } Y \sim \exp(\alpha) \\ &\text{Step 2: Set } X = \beta e^Y. \end{aligned}$$

Normal

- A random variable $X \in \mathbb{R}$ has a normal distribution with parameters μ and σ^2, denoted $X \sim \mathcal{N}(\mu, \sigma^2)$, if the pdf is

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

  we will also use $\phi(x|\mu, \sigma^2)$ to denote the pdf and $\Phi(x|\mu, \sigma^2)$ the cdf. The mean and variance are $E(X) = \mu$ and $var(X) = \sigma^2$.

- Given the importance of normal random variables, all software packages have functions to draw normal random variables. The algorithms typically use transformation methods drawing uniform and exponential random variables or look-up tables. The classic algorithm is the Box-Muller approach based on simulating uniforms.

Log-Normal

- A random variable $X \in \mathbb{R}_+$ has a lognormal distribution with parameters μ and σ^2, denoted by $X \sim \mathcal{LN}(\mu, \sigma^2)$ if the pdf is

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

  The mean and variance of the log-normal distribution are $E\left(X\right) = e^{\mu+\frac{1}{2}\sigma^2}$ and similarly $var\left(X\right) = \exp\left(2\mu+\sigma^2\right)\left(\exp\left(\sigma^2\right)-1\right)$. It is related to a normal distribution via the transformation $X = e^{\mu+\sigma Z}$. Although finite moments of the lognormal exist, the distribution does not admit a moment-generating function.

- Simulating lognormal random variables via the transformation method is straightforward since $X = e^{\mu+\sigma Z}$ where $Z \sim \mathcal{N}(0,1)$ is $\mathcal{LN}(\mu, \sigma^2)$.

Truncated Normal

- A random variable X has a truncated normal distribution, denoted by $\mathcal{TN}(\mu, \sigma^2)$, with parameters μ, σ^2 and truncation region (a, b) if the pdf is

$$p\left(x|a<x<b\right) = \frac{\phi\left(x|\mu,\sigma^2\right)}{\Phi\left(b|\mu,\sigma^2\right)-\Phi\left(a|\mu,\sigma^2\right)},$$

  where it is clear that $\int_{-\infty}^{b}\phi\left(x|\mu,\sigma^2\right)dx = \Phi\left(b|\mu,\sigma^2\right)$. The mean of a truncated normal distribution is

$$E(X|a<X<b) = \mu-\sigma\frac{\phi_a-\phi_b}{\Phi_b-\Phi_a},$$

  where ϕ_x is the standard normal density evaluated at $\left(x-\mu\right)/\sigma$ and Φ_x is the standard normal cdf evaluated at $\left(x-\mu\right)/\sigma$.

- The inversion method can be used to simulate this distribution. A two-step algorithm that provides a draw from a truncated standard normal is

$$\begin{aligned}&\text{Step 1: } U \sim U\left[0,1\right]\\&\text{Step 2: } X = \Phi^{-1}\left[\Phi\left(a\right)+U\left(\Phi\left(b\right)-\Phi\left(a\right)\right)\right],\end{aligned}$$

where $\Phi(a) = \int_{-\infty}^{a} (2\pi)^{-1/2} \exp(-x^2/2)\, dx$. For the general truncated normal,

$$\begin{aligned}&\text{Step 1: Draw } U \sim U[0,1]\\ &\text{Step 2: } X = \mu + \sigma\Phi^{-1}\left[\Phi\left(\frac{a-\mu}{\sigma}\right) + U\left(\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\right)\right],\end{aligned}$$

where Φ^{-1} is the inverse of the cdf.

Double exponential

- A random variable $X \in \mathbb{R}$ has a double exponential (or Laplace) distribution with parameters μ and σ^2, denoted $X \sim \mathcal{DE}(\mu, \sigma)$, if the pdf is

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

  The mean and variance are $E(X) = \mu$ and $var(X) = 2\sigma^2$.

- The composition method can be used to simulate a $\mathcal{DE}(\mu, \sigma)$ random variable:

$$\begin{aligned}&\text{Step 1: Draw } \lambda \sim \exp(2) \text{ and } Z \sim \mathcal{N}(0,1)\\ &\text{Step 2: Set } X = \mu + \sigma\sqrt{\lambda}Z.\end{aligned}$$

Check exponential

- A random variable $X \in \mathbb{R}$ has a check (or asymmetric) exponential distribution with parameters μ and σ^2, denoted $X \sim \mathcal{CE}(\tau, \mu, \sigma)$, if the pdf is

$$p(x|\mu,\sigma) = \frac{1}{\sigma\mu_\tau} \exp\left(-\frac{1}{\sigma}\rho_\tau(x-\mu)\right).$$

  where $\rho_\tau(x) = |x| - (2\tau-1)x$ and $\mu_\tau^{-1} = 2\tau(1-\tau)$. The double exponential is a special case when $\tau = \frac{1}{2}$.

- The composition method simulates a $\mathcal{CE}(\mu, \sigma)$ random variable:

$$\begin{aligned}&\text{Step 1: Draw } \lambda \sim \exp(\mu_\tau) \text{ and } Z \sim \mathcal{N}(0,1)\\ &\text{Step 2: Set } X = \mu + (2\tau-1)\sigma\lambda + \sigma\sqrt{\lambda}Z.\end{aligned}$$

T

- A random variable $X \in \mathbb{R}^+$ has a t-distribution with parameters ν,μ, and σ^2, denoted $X \sim t_\nu(\mu, \sigma^2)$, if the pdf is

$$p\left(x|\nu,\mu,\sigma^2\right) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi\sigma^2}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}$$

  When $\mu = 0$ and $\sigma = 1$, the distribution is denoted merely as t_ν. The mean and variance of the t-distribution are $E(X) = \mu$ and $var(X) = \sigma^2\frac{\nu}{\nu-2}$ for $\nu > 2$. The Cauchy distribution is the special case where $\nu = 1$.

- The composition method simulates a $t_\nu(\mu, \sigma^2)$ random variable,

  Step 1. Draw $\lambda \sim \mathcal{IG}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$ and $Z \sim \mathcal{N}(0,1)$
  Step 2. Set $X = \mu + \sigma\sqrt{\lambda}Z$

Z

- The class of Z-distributions for $a, b > 0$ have density

$$f_Z(z|a,b,\mu,\sigma) = \frac{1}{\sigma B(a,b)}\frac{e^{a(z-\mu)/\sigma}}{(1+e^{(z-\mu)/\sigma})^{a+b}} \equiv Z(z;a,b,\sigma,\mu).$$

  A typical parameterisation is $a = \delta + \theta, b = \delta - \theta$ or $\delta = \frac{1}{2}(a+b), \theta = \frac{1}{2}(a-b)$. This is a variance-mean mixture of normals where

$$Z(z;a,b,\sigma,\mu) = \int_0^\infty \frac{1}{\sqrt{2\pi\lambda\sigma^2}}\exp\left\{-\frac{1}{2\lambda\sigma^2}\left(z-\mu-\frac{1}{2}(a-b)\lambda\sigma\right)^2\right\}p_{a,b}(\lambda)\,d\lambda$$

  where $p_{a,b}(\lambda)$ is a Polya distribution which is an infinite mixture of exponentials that can be easily sampled as

$$\lambda \stackrel{\mathrm{D}}{=} \sum_{k=0}^{\infty} 2\psi_k^{-1}Z_k, \quad \text{where} \quad \psi_k = (a+k)(b+k) \quad \text{and} \quad Z_k \sim \exp(1).$$

- Z-Distribution simulation is then given by

$$X = \mu + \frac{1}{2}(a-b)\lambda\sigma + \sigma\sqrt{\lambda}Z, \text{ where } Z \sim \mathcal{N}(0,1)$$

Exponential power

- A random variable has an exponential power distribution. denoted $X \sim \mathcal{EP}(\mu, \sigma, \gamma)$ if the pdf is

$$p\left(x|\mu,\sigma,\gamma\right) = \frac{1}{2\sigma\Gamma(1+\gamma^{-1})}\exp\left(-\left|\frac{x-\mu}{\sigma}\right|^{\gamma}\right).$$

  The mean and variance are $E\left(X\right) = \mu$ and $var\left(X\right) = \frac{\Gamma(3/\gamma)}{\Gamma(1/\gamma)}\sigma^2$. Normal and double exponential are then special cases.

- Exponential power simulation relies on the composition method: if $X \sim \mathcal{EP}\left(\mu,\sigma\right)$, then $X = \sigma\sqrt{\lambda}Z$, where the scale parameter is related to a positive stable random variable $\lambda \sim \lambda^{-\frac{3}{2}} St^{+}_{\alpha/2}\left(\lambda^{-1}\right)$ and $\varepsilon \sim \mathcal{N}\left(0,1\right)$.

Inverse Gaussian

- A random variable $X \in \mathbb{R}_{+}$ has an inverse Gaussian distribution with parameters μ and α, denoted $X \sim \mathcal{IN}\left(\mu,\alpha\right)$, if the pdf is

$$p\left(x|\mu,\alpha\right) = \sqrt{\frac{\alpha}{2\pi x^3}}\exp\left(-\frac{\alpha\left(x-\mu\right)^2}{2\mu^2 x}\right).$$

  The mean and variance of an inverse Gaussian random variable are $E\left(X\right) = \mu$ and $var\left(X\right) = \mu^3/\alpha$, respectively.

- To simulate an inverse Gaussian $\mathcal{IN}(\mu,\alpha)$,

$$\begin{aligned}
&\text{Step 0: Draw } U \sim U(0,1), V \sim \chi^2_1\\
&\text{Step 1: Draw } W = \xi + \frac{\xi^2}{2\mu}V - \frac{\xi}{2\sqrt{\mu}}\sqrt{4\xi\mu V + \xi^2 V^2}\\
&\text{Step 2: Set } X = W 1_{(U \geq \frac{\xi}{\xi+W})} + \frac{\xi^2}{W} 1_{(U \leq \frac{\xi}{\xi+W})}
\end{aligned}$$

  where $\xi = \sqrt{\mu/\alpha}$.

Generalized inverse Gaussian

- A random variable $X \in \mathbb{R}_+$ has an generalized inverse Gaussian distribution with parameters a, b, and p, $X \sim \mathcal{GIG}(a, b, p)$, if the pdf is

$$p(x|a, b, p) = \left(\frac{a}{b}\right)^{\frac{p}{2}} \frac{x^{p-1}}{2K_p\left(\sqrt{ab}\right)} \exp\left(-\frac{1}{2}\left[ax + \frac{b}{x}\right]\right),$$

  where K_p is the modified Bessel function of the third kind. The mean and variance are known, but are complicated expressions of the Bessel functions. The Gamma distribution is the special case with $\beta = b/2$ and $a = 0$, the inverse Gamma is the special case with $a = 0$.

- Simulating GIG random variables is typically done using resampling methods.

Multivariate normal

- A $k \times 1$ random vector $X \in \mathbb{R}_+^k$ has a multivariate normal distribution with parameters μ and Σ, denoted $X \sim \mathcal{N}_k(\mu, \Sigma)$, if the pdf is

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{k}{2}}} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x - \mu)\Sigma^{-1}(x - \mu)'\right)$$

  where $|\Sigma|$ is the determinant of the positive definite symmetric matrix Σ. The mean and covariance matrix of a multivariate normal are $E(X) = \mu$ and $cov(X) = \Sigma$, respectively.

Multivariate T

- A $1 \times k$ random vector $X \in \mathbb{R}_+^k$ has a multivariate t-distribution with parameters ν, μ, and Σ, denoted $X \sim t_\nu(\mu, \Sigma)$, if the pdf is given by

$$p(x|\nu, \mu, \Sigma) = \frac{\Gamma\left(\frac{\nu+k}{2}\right)\Gamma\left(\frac{\nu}{2}\right)}{(\nu\pi)^{1/2}|\Sigma|^{1/2}} \left[1 + \frac{(x - \mu)\Sigma^{-1}(x - \mu)'}{\nu}\right]^{-\frac{\nu+k}{2}}.$$

  The mean and covariance matrix of a multivariate t random variable are $E(X) = \mu$ and $cov(X) = \Sigma$, respectively.

- The following two steps provide a draw from a multivariate t-distribution:

$$\text{Step 1. Simulate } Y \sim \mathcal{N}_k(\mu, \Sigma) \text{ and } Z \sim \mathcal{X}^2_\nu$$
$$\text{Step 2. Set } X = \mu + Y\left(\frac{Z}{\nu}\right)^{-\frac{1}{2}}$$

Wishart

- A random $m \times m$ matrix Σ has a Wishart distribution, $\Sigma \sim \mathcal{W}_m(v, V)$, if the density function is given by

$$p(\Sigma|v, V) = \frac{|\Sigma|^{\frac{(v-m-1)}{2}}}{2^{\frac{vm}{2}}|V|^{\frac{v}{2}}\Gamma_m\left(\frac{v}{2}\right)}\exp\left(-\frac{1}{2}tr\left(V^{-1}\Sigma\right)\right),$$

  for $v > m$, where

$$\Gamma_m\left(\frac{v}{2}\right) = \prod_{k=1}^{m}\pi^{\frac{m(m-1)}{4}}\Gamma_m\left(\frac{v-k+1}{2}\right)$$

  is the multivariate Gamma function. If $v < m$, then S does not have a density although its distribution is well defined. The Wishart distribution arises naturally in multivariate settings with normally distributed random variables as the distribution of quadratic forms of multivariate normal random variables.

- The Wishart distribution can be viewed as a multivariate generalization of the $\mathcal{X}^2_\nu$ distribution. From this, it is clear how to sample from a Wishart distribution:

$$\text{Step 1: Draw } X_j \sim \mathcal{N}(0, V) \text{ for } j = 1, ..., v$$
$$\text{Step 2: Set } S = \sum_{j=1}^{v} X_j X_j'.$$

Inverted Wishart

- A random $m \times m$ matrix Σ has an inverted Wishart distribution, denoted $\Sigma \sim \mathcal{IW}_m(v, V)$ if the density function is

$$p(\Sigma|v, V) = \frac{|\Sigma|^{-\frac{(v+m+1)}{2}}}{2^{\frac{vm}{2}}|V|^{-m/2}\Gamma_m\left(\frac{v}{2}\right)}\exp\left(-\frac{1}{2}tr\left(V\Sigma^{-1}\right)\right).$$

This also implies that Σ^{-1} has a Wishart distribution, $\Sigma^{-1} \sim \mathcal{W}_m(v, V^{-1})$. The Jacobian of the transformation

$$\left|\frac{\partial \Sigma^{-1}}{\partial \Sigma}\right| = |\Sigma|^{-(m+1)}.$$

- To generate $\Sigma \sim \mathcal{IW}_m(w, W)$, follow the two step procedure:

  Step 1: Draw $X_i \sim \mathcal{N}\left(0, W^{-1}\right)$
  Step 2: Set $\Sigma = \sum_{i=1}^{v} X_i X_i'$.

  In cases where m is extremely large, there are more efficient algorithms for drawing inverted Wishart random variables that involves factoring W and sampling from univariate $\mathcal{X}^2$ distributions.

2 Likelihoods, Priors, and Posteriors

This appendix provides combinations of likelihoods and priors for the following types of observed data: Bernoulli, Poisson, exponential, normal, normal regression, and multivariate normal. For each specification, proper conjugate and Jeffreys' priors are given. The overriding Bayesian paradigm takes the form of Bayes rule

$$p(\text{parameters}|\text{data}) = \frac{p(\text{data}|\text{parameters})p(\text{parameters})}{p(\text{data})}$$

where the types of data and parameters are problem specific.

Bernoulli observations

- If the data $(y_t|\theta) \sim \mathcal{B}er(\theta)$ with $\theta \in [0,1]$, then the likelihood is

$$p(y|\theta) = \prod_{t=1}^{T} p(y_t|\theta) = \prod_{t=1}^{T} \theta^{y_t} (1-\theta)^{1-y_t} = \theta^{\sum_{t=1}^{T} y_t} (1-\theta)^{T-\sum_{t=1}^{T} y_t},$$

- A conjugate prior distribution is the beta family, $\theta \sim \mathcal{B}(a, A)$, where

$$p(\theta) = \frac{\Gamma(a+A)}{\Gamma(a)\Gamma(A)} \theta^{a-1} (1-\theta)^{A-1}.$$

  By Bayes rule, the posterior distribution is also Beta

$$p(\theta|y) \propto p(y|\theta)\, p(\theta) = \lambda^{a+\sum_{t=1}^{T} y_t - 1} (1-\theta)^{A+T-\sum_{t=1}^{T} y_t - 1} \sim \mathcal{B}(a_T, A_T),$$

  where $a_T = a + \sum_{t=1}^{T} y_t$ and $A_T = A + T - \sum_{t=1}^{T} y_t$.

- Fisher's information for Bernoulli observations is

$$I(\theta) = -E_\theta \left[\frac{\partial^2 \ln p(y_t|\theta)}{\partial \theta^2} \right] = \frac{1}{\theta(1-\theta)},$$

  where E_θ denotes the expectation under a $Ber(\theta)$. Jeffreys' prior is

$$p(\theta) = I(\theta)^{\frac{1}{2}} = \theta^{-\frac{1}{2}} (1-\theta)^{-\frac{1}{2}} \sim \mathcal{B}\left(\frac{1}{2}, \frac{1}{2}\right).$$

Multinomial observations

- If $(y|\theta)$ is Multinomial data from k categories where $(y|\theta) \sim Multi(\theta_1, \ldots, \theta_k)$, then the likelihood for T trials is given by

$$p(y_1, \ldots, y_k|\theta_1, \ldots, \theta_k) = \frac{T!}{y_1! \ldots y_k!} \theta_1^{y_1} \ldots \theta_k^{y_k} \text{ where } \sum_{i=1}^{k} y_i = T$$

- A conjugate prior is a Dirichlet distribution, $\theta \sim Dir(\alpha)$, with density

$$p(\theta_1, \ldots, \theta_k|\alpha) = \frac{\Gamma(\sum \alpha_i)}{\prod_i \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \ldots \theta_k^{\alpha_k - 1}$$

  The posterior is then

$$\begin{aligned} p(\theta_1, \ldots, \theta_k|\alpha, y_1, \ldots, y_k) &\propto p(y_1, \ldots, y_k|\theta_1, \ldots, \theta_k) p(\theta_1, \ldots, \theta_k|\alpha) \\ &\propto \theta_1^{y_1} \ldots \theta_k^{y_k} \theta_1^{\alpha_1 - 1} \ldots \theta_k^{\alpha_k - 1} \\ &= \theta_1^{\alpha_1 + y_1 - 1} \ldots \theta_k^{\alpha_k + y_k - 1} \\ &\sim Dir(\alpha + y) \end{aligned}$$

  which is again a Dirchlet with parameter $\alpha + y$.

Poisson observations

- If the data $(y_t|\lambda) \sim \text{Poi}(\lambda)$, then the likelihood is

$$p(y|\lambda) = \prod_{t=1}^{T} \frac{e^{-\lambda} \lambda^{y_t}}{y_t!} \propto e^{-\lambda T} \lambda^{\sum_{t=1}^{T} y_t},$$

- A conjugate prior for λ is a Gamma distribution, $\lambda \sim \mathcal{G}(a, A)$, with density

$$p(\lambda) = \frac{A^a}{\Gamma(a)} \lambda^{a-1} \exp(-\lambda A).$$

  The posterior distribution is Gamma

$$p(\lambda|y) \propto p(y|\lambda) p(\lambda) = e^{-\lambda(A+T)} \lambda^{a + \sum_{t=1}^{T} y_t - 1} \sim \mathcal{G}(a_T, A_T)$$

where $a_T = a + \sum_{t=1}^{T} y_t$ and $A_T = A + T$.

- Fisher's information for Poisson observations is

$$I(\lambda) = -E_\lambda \left[\frac{\partial^2 \ln p(y_t|\lambda)}{\partial \lambda^2} \right] = \frac{1}{\lambda}.$$

  Jeffreys' prior is then $p(\lambda) \equiv I(\lambda)^{\frac{1}{2}} = \lambda^{-\frac{1}{2}}$. This can be viewed as a special case of the Gamma prior with $a = \frac{1}{2}$ and $A = 0$.

Exponential observations

- If the data $(y_t|\mu) \sim \exp(\mu)$, then the likelihood is

$$p(y|\mu) = \prod_{t=1}^{T} \mu \exp(-\mu y_t) \propto \mu^T \exp\left(-\mu \sum_{t=1}^{T} y_t \right),$$

- A conjugate prior for μ is a Gamma distribution, $\mu \sim \mathcal{G}(a, A)$. The posterior

$$p(\mu|y) \propto p(y|\mu)\, p(\mu) \propto \mu^{a+T-1} e^{-\mu\left(A + \sum_{t=1}^{T} y_t\right)} \sim \mathcal{G}(a_T, A_T)$$

  where $a_T = a + T$ and $A_T = A + \sum_{t=1}^{T} y_t$.

- Fishers' information for exponential observations is

$$I(\mu) = -E_\mu \left[\frac{\partial^2 \ln p(y_t|\mu)}{\partial \mu^2} \right] = \left(\frac{1}{\mu} \right)^2.$$

  Jeffreys prior for exponential observations is $p(\mu) \equiv \mu^{-1}$. This is a special case of the Gamma prior with $a = 1$ and $A = 0$.

Normal observations with known variance

- If the data $(y_t|\mu, \sigma^2) \sim \mathcal{N}(\mu, \sigma^2)$, the likelihood is

$$p\left(y|\mu, \sigma^2\right) = \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{T}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{t=1}^{T} (y_t - \mu)^2 \right).$$

We can factorize

$$\sum_{t=1}^{T}(y_t-\mu)^2=\sum_{t=1}^{T}\left[(y_t-\overline{y})^2+(\overline{y}-\mu)^2\right].$$

Thus, the likelihood, as a function of μ, is proportional to

$$\exp\left(-\frac{T(\mu-\overline{y})^2}{2\sigma^2}\right),$$

with the other terms in $p(y|\mu,\sigma^2)$ being absorbed into the constant of integration.

- A conjugate prior distribution for μ that is independent of σ^2 is given by $\mu\sim\mathcal{N}(a,A)$. The posterior distribution is

$$p\left(\mu|y,\sigma^2\right)\propto p\left(y|\mu,\sigma^2\right)p(\mu)\propto\exp\left(-\frac{1}{2}\left[\frac{(\mu-\overline{y})^2}{\sigma^2/T}-\frac{(\mu-a)^2}{A}\right]\right).$$

Completing the square yields

$$\frac{(\mu-\overline{y})^2}{\sigma^2/T}-\frac{(\mu-a)^2}{A}=\frac{(\mu-a_T)^2}{A_T}+\frac{(\overline{y}-a)^2}{\sigma^2/T+A},$$

with parameters

$$\frac{a_T}{A_T}=\frac{a}{A}+\frac{\overline{y}}{\sigma^2/T}\text{ and }\frac{1}{A_T}=\frac{1}{\sigma^2/T}+\frac{1}{A}.$$

The posterior distribution is

$$p\left(\mu|y,\sigma^2\right)\propto\exp\left(-\frac{1}{2}\left[\frac{(\mu-a_T)^2}{A_T}+\frac{(\overline{y}-a)^2}{\sigma^2/T+A}\right]\right)\sim\mathcal{N}(a_T,A_T).$$

- A conjugate prior distribution for μ conditional on σ^2 is $\mu\sim\mathcal{N}(a,\sigma^2A)$. The posterior distribution is $p(\mu|y,\sigma^2)\propto p(y|\mu,\sigma^2)p(\mu)$ which gives

$$p\left(\mu|y,\sigma^2\right)\propto\exp\left\{-\frac{1}{2\sigma^2}\left(\frac{(\mu-\overline{y})^2}{T^{-1}}+\frac{(\mu-a)^2}{A}\right)\right\}$$

Completing the square for the quadratic term in the exponential,

$$\frac{(\mu-\overline{y})^2}{T^{-1}}+\frac{(\mu-a)^2}{A}=\frac{(\mu-a_T)^2}{A_T}+\frac{(\overline{y}-a)^2}{T^{-1}+A}$$

where

$$\frac{a_T}{A_T}=\frac{a}{A}+\frac{\overline{y}}{T^{-1}} \text{ and } \frac{1}{A_T}=\frac{1}{T^{-1}}+\frac{1}{A}.$$

The posterior distribution is

$$p\left(\mu|y,\sigma^2\right)\propto\exp\left(-\frac{(\mu-a_T)^2}{2\sigma^2A_T}\right)\sim\mathcal{N}\left(a_T,A_T\sigma^2\right).$$

Notice the slight differences between this example and the previous one, in terms of the hyper-parameters and the form of the posterior distribution.

- Fisher's information for normal observations with σ^2 known is

$$I\left(\mu\right)=-E_\lambda\left[\frac{\partial^2\ln p\left(y_t|\mu,\sigma^2\right)}{\partial\mu^2}\right]\equiv 1.$$

Jeffreys prior for normal observations (with known variance) is a constant, $p\left(\mu\right)\equiv 1$ which is improper. However, the posterior is proper and can be viewed as a limiting of the normal conjugate prior with $a=0$ and $A\to\infty$.

Normal Variance with known Mean

- Given μ, if $\left(y_t|\mu,\sigma^2\right)\sim\mathcal{N}\left(\mu,\sigma^2\right)$ the likelihood for σ^2, is

$$\left(\frac{1}{\sigma^2}\right)^{\frac{T}{2}}\exp\left(-\frac{\sum_{t=1}^{T}\left(y_t-\mu\right)^2}{2\sigma^2}\right).$$

- A conjugate inverse Gamma prior $\sigma^2\sim\mathcal{IG}\left(\frac{b}{2},\frac{B}{2}\right)$ has pdf

$$p\left(\sigma^2\right)=\frac{\left(\frac{B}{2}\right)^{\frac{b}{2}}}{\Gamma\left(\frac{b}{2}\right)}\left(\sigma^2\right)^{-\frac{b}{2}-1}\exp\left(-\frac{B}{2\sigma^2}\right)$$

By Bayes rule,

$$\begin{aligned} p\left(\sigma^2|\mu, y\right) &\propto p\left(y|\mu, \sigma^2\right) p\left(\sigma^2\right) \\ &\propto \left(\frac{1}{\sigma^2}\right)^{\frac{b+T}{2}+1} \exp\left(-\frac{B+\sum_{t=1}^{T}\left(y_t-\mu\right)^2}{2\sigma^2}\right) \\ &\sim \mathcal{IG}\left(\frac{b_T}{2}, \frac{B_T}{2}\right), \end{aligned}$$

where $b_T = b + T$ and $B_T = B + \sum_{t=1}^{T}\left(y_t - \mu\right)^2$.

The parameterization of the inverse Gamma, $\sigma^2 \sim \mathcal{IG}\left(\frac{b}{2}, \frac{B}{2}\right)$, is used as opposed to $\sigma^2 \sim \mathcal{IG}\left(b, B\right)$ because the hyperparameters do not have any $1/2$ terms. This is chosen for notational simplicity. It is also common in the literature to assume $p\left(\sigma\right) \sim \mathcal{IG}\left(\frac{b}{2}, \frac{B}{2}\right)$, which only changes the first term in the expression for b_T.

- Fisher's information for normal observations (with known mean) is

$$I\left(\sigma^2\right) = -E_{\sigma^2}\left[\frac{\partial^2 \ln p\left(y_t|\mu, \sigma^2\right)}{\partial\left(\sigma^2\right)^2}\right] \equiv \frac{1}{\sigma^2}.$$

Jeffreys' prior is $p\left(\sigma^2\right) \propto \left(\sigma^2\right)^{-1}$, which is improper distribution. However, the resulting posterior is proper and can be viewed as a limiting of the inverse Gamma conjugate prior with $B = 0$ and $b = 0$. A flat or constant prior for σ^2 also leads to a proper posterior. Assuming $p\left(\sigma^2\right) \equiv 1$ yields a conditional posterior

$$p\left(\sigma^2|\mu, y\right) \sim \mathcal{IG}\left(\frac{T}{2}, \frac{\sum_{t=1}^{T}\left(y_t-\mu\right)^2}{2}\right).$$

Unknown mean and variance: dependent priors

- If the data $\left(y_t|\mu, \sigma^2\right) \sim \mathcal{N}\left(\mu, \sigma^2\right)$ and assuming that both μ and σ^2 are unknown, then the likelihood as a function of μ and σ^2 is

$$\left(\frac{1}{\sigma^2}\right)^{\frac{T}{2}} \exp\left(-\frac{\sum_{t=1}^{T}\left(y_t-\mu\right)^2}{2\sigma^2}\right).$$

- A conjugate prior for (μ, σ^2) is $p(\mu, \sigma^2) = p(\mu|\sigma^2)\, p(\sigma^2)$ where

$$p\left(\mu|\sigma^2\right) \sim \mathcal{N}\left(a, A\sigma^2\right) \text{ and } p\left(\sigma^2\right) \sim \mathcal{IG}\left(\frac{b}{2}, \frac{B}{2}\right).$$

This distribution is often expressed as $\mathcal{N}(a, A\sigma^2)\,\mathcal{IG}\left(\frac{b}{2}, \frac{B}{2}\right)$. This prior assumes that μ and σ^2 are dependent. Bayes rule and a few lines of algebra yields a posterior

$$\begin{aligned} p\left(\mu, \sigma^2|y\right) &\propto p\left(y|\mu, \sigma^2\right) p\left(\mu|\sigma^2\right) p\left(\sigma^2\right) \\ &\propto \left(\frac{1}{\sigma^2}\right)^{\frac{T+b}{2}+\frac{1}{2}+1} \exp\left(-\frac{1}{2\sigma^2}\left[\frac{(\mu-\overline{y})^2}{1/T} + \frac{(\mu-a)^2}{A} + \sum_{t=1}^{T}(y_t-\overline{y})^2 + B\right]\right). \end{aligned}$$

Combining the quadratic terms that depend on μ by completing the square

$$\frac{(\mu-\overline{y})^2}{1/T} + \frac{(\mu-a)^2}{A} = \frac{(\mu-a_T)^2}{A_T} + \frac{(\overline{y}-a)^2}{1/T+A},$$

where the hyper-parameters are

$$\frac{a_T}{A_T} = \frac{a}{A} + \frac{\overline{y}}{T^{-1}} \text{ and } \frac{1}{A_T} = \frac{1}{A} + \frac{1}{T^{-1}}.$$

Inserting this into the likelihood, gives a posterior

$$p\left(\mu, \sigma^2|y\right) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{T+b}{2}+\frac{1}{2}+1} \exp\left(-\frac{1}{2\sigma^2}\left[\frac{(\mu-a_T)^2}{A_T} + B + S\right]\right)$$

where we have

$$S = \frac{(\overline{y}-a)^2}{1/T+A} + \sum_{t=1}^{T}(y_t-\overline{y})^2.$$

Given the conjugate prior structure, the posterior $p(\mu, \sigma^2|y) \propto p(\mu|\sigma^2, y)\, p(\sigma^2|y)$ A few lines of algebra shows that

$$p\left(\mu, \sigma^2|y\right) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{(\mu-a_T)^2}{\sigma^2 A_T}\right) \times \left(\frac{1}{\sigma^2}\right)^{\frac{T+b}{2}+1} \exp\left(-\frac{B+S}{2\sigma^2}\right)$$

given $p(\mu|\sigma^2, y) \sim \mathcal{N}(a_T, \sigma^2 A_T)$ and $p(\sigma^2|y) \sim \mathcal{IG}\left(\frac{b_T}{2}, \frac{B_T}{2}\right)$ with $b_T = b + T, B_T =$

$B + S$.

Marginal parameter distributions. In this specification, the marginal parameter distributions, $p\left(\sigma^2|y\right)$ and $p\left(\mu|y\right)$, are both known analytically. $p\left(\sigma^2|y\right)$ is inverse Gamma. The marginal $p\left(\mu|y\right)$ is

$$p\left(\mu|y\right) = \int_0^\infty p\left(\mu, \sigma^2|y\right) d\sigma^2 = \int_0^\infty p\left(\mu|\sigma^2, y\right) p\left(\sigma^2|y\right) d\sigma^2.$$

Both $p\left(\mu|\sigma^2, y\right)$ and $p\left(\sigma^2|y\right)$ are known, and the integral can be computed analytically. Ignoring integration constants,

$$\begin{aligned} p\left(\mu|y\right) &\propto \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{b_T+1}{2}+1} \exp\left(-\frac{1}{\sigma^2}\left[\frac{(\mu - a_T)}{2A_T} + \frac{B_T}{2}\right]\right) d\sigma^2 \\ &\propto \left[1 + \frac{(\mu - a_T)^2}{A_T B_T}\right]^{-\frac{b_T+1}{2}} \end{aligned}$$

using our integration results. This is the kernel of a t-distribution, thus the marginal posterior is $p(\mu|y) \sim t_{b_T}\left(a_T, A_T B_T\right)$.

The marginal likelihood, $p\left(y\right) = \int p\left(y|\mu, \sigma^2\right) p\left(\mu, \sigma^2\right) d\mu d\sigma^2$ can be computed

$$\begin{aligned} p\left(y|\mu, \sigma^2\right) =& K_y \left(\frac{1}{\sigma^2}\right)^{\frac{T}{2}} \exp\left(-\frac{T\left(\mu - \overline{y}\right)^2 + S}{2\sigma^2}\right) \\ p\left(\mu|\sigma^2\right) =& K_\mu \left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}\frac{(\mu - a)^2}{A}\right) \\ p\left(\sigma^2\right) =& K_\sigma \left(\frac{1}{\sigma^2}\right)^{\frac{b}{2}+1} \exp\left(-\frac{B}{2\sigma^2}\right), \end{aligned}$$

with constants $K_y = (2\pi)^{-T/2}$, $K_\sigma = (B/2)^{\acute{b}/2} / \Gamma\left(b/2\right)$ and $K_\mu = (2\pi A)^{-\frac{1}{2}}$. Substi-

tuting these expressions, the marginal likelihood is

$$p(y) = KK_\mu K_\sigma \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{b+T+1}{2}+1} \exp\left(-\frac{S+B}{2\sigma^2}\right) d\sigma^2 \cdot \int_{-\infty}^\infty \exp\left(-\frac{1}{2\sigma^2}\left[\frac{(\mu-\overline{y})^2}{T^{-1}} + \frac{(\mu-a)^2}{A}\right]\right) d\mu$$

Completing the square inside the integrand gives

$$\frac{(\mu-\overline{y})^2}{T^{-1}} + \frac{(\mu-a)^2}{A} = \frac{(\mu-a_T^\mu)^2}{A^\mu} + \frac{(\overline{y}-a)^2}{T^{-1}+A},$$

with hyper-parameters

$$\frac{a_T^\mu}{A_T^\mu} = \frac{\overline{y}}{T^{-1}} + \frac{a}{A} \text{ and } \frac{1}{A_T^\mu} = \frac{1}{T^{-1}} + \frac{1}{A}.$$

The integrals can be expressed as

$$\int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{b+T+1}{2}+1} \exp\left(-\frac{S+B+\frac{(\overline{y}-a)^2}{T^{-1}+A}}{2\sigma^2}\right) d\sigma^2 \int_{-\infty}^\infty \exp\left(-\frac{1}{2\sigma^2}\frac{(\mu-a_T^\mu)^2}{A^\mu}\right) d\mu$$

The second integral is $\int_{-\infty}^\infty \exp\left(-\frac{1}{2}\frac{\left(\mu-a_T^\mu\right)^2}{\sigma^2 A^\mu}\right) d\mu = \sqrt{2\pi A^\mu}\left(\sigma^2\right)^{\frac{1}{2}}$. Using this, σ^2 can be integrated out yielding the expression for the marginal likelihood:

$$\begin{aligned} p(y) &= \sqrt{2\pi A^\mu} KK_\mu K_\sigma \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{b+T}{2}+1} \exp\left(-\frac{S+B+\frac{(\overline{y}-a)^2}{T^{-1}+A}}{2\sigma^2}\right) d\sigma^2 \\ &= \left(\frac{1}{2\pi}\right)^{\frac{T}{2}} \left(\frac{A^\mu}{A}\right)^{\frac{1}{2}} \left(\frac{B}{2}\right)^{\frac{\acute{b}}{2}} \frac{\Gamma\left(\frac{b+T}{2}\right)}{\Gamma\left(\frac{b}{2}\right)} \left[S+B+\frac{(\overline{y}-a)^2}{T^{-1}+A}\right]^{-\frac{b+T}{2}}. \end{aligned}$$

Finally, the predictive distribution can also be computed analytically. Since

$$p\left(y_{T+1}|y^T\right) = \int p\left(y_{T+1}|\mu,\sigma^2\right) p\left(\mu,\sigma^2|y^T\right) d\mu d\sigma^2.$$

To simplify, first compute the integral against μ by substituting from the posterior.

Since $p(\mu|\sigma^2, y) \sim \mathcal{N}(a_T, \sigma^2 A_T)$, we have that $\mu = a_T + \sigma\sqrt{A_T}Z$ where Z is an independent normal. Substituting in $y_{T+1} = \mu + \sigma\varepsilon_{T+1}$, gives

$$y_{T+1} = a_T + \sigma\sqrt{A_T}Z + \sigma\varepsilon_{T+1} = a_T + \sigma\eta_{T+1}$$

where $\eta_{T+1} \sim \mathcal{N}(0, A_T + 1)$. The predictive is therefore

$$\begin{aligned}
p\left(y_{T+1}|y^T\right) &\propto \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}\frac{(y_{T+1}-a_T)^2}{A_T+1}\right)\left(\frac{1}{\sigma^2}\right)^{\frac{b_T}{2}+1}\exp\left(-\frac{B_T}{2\sigma^2}\right)d\sigma^2 \\
&\propto \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{B_T+1}{2}+1}\exp\left(-\frac{1}{2\sigma^2}\left[B_T + \frac{(y_{T+1}-a_T)^2}{A_T+1}\right]\right)d\sigma^2 \\
&\propto \left[1 + \frac{(y_{T+1}-a_T)^2}{B_T(A_T+1)}\right]^{\frac{B_T+1}{2}} \\
&\sim t_{b_T}\left(a_T, B_T(A_T+1)\right).
\end{aligned}$$

- Jeffreys' prior for normal observations with unknown mean and variance is a bivariate distribution. Fishers' information is

$$I\left(\mu,\sigma^2\right) = -E_{\mu,\sigma^2}\left[\begin{array}{cc} \frac{\partial^2 \ln p\left(y_t|\mu,\sigma^2\right)}{\partial\mu^2} & \frac{\partial^2 \ln p\left(y_t|\mu,\sigma^2\right)}{\partial\mu\partial\sigma^2} \\ \frac{\partial^2 \ln p\left(y_t|\mu,\sigma^2\right)}{\partial\mu\partial\sigma^2} & \frac{\partial^2 \ln p\left(y_t|\mu,\sigma^2\right)}{\partial(\sigma^2)^2}\end{array}\right] = \left[\begin{array}{cc} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2}\end{array}\right],$$

  generating a prior distribution $p(\mu, \sigma^2) \equiv \det\left(I(\mu, \sigma^2)\right)^{\frac{1}{2}} = \sigma^{-2}$. This prior is improper, but leads to a proper posterior distribution that can be viewed as a limiting case of the usual conjugate posterior

$$p\left(\mu, \sigma^2|y\right) \sim \mathcal{N}\left(a_T, A_T\sigma^2\right)\mathcal{IG}\left(\frac{b_T}{2}, \frac{B_T}{2}\right),$$

  where $a_T = 0$, $A_T \to \infty$, $b = 0$ and $B = 0$.

Regression

- Consider a regression model specification, $y_t|x_t, \beta, \sigma^2 \sim \mathcal{N}(x_t\beta, \sigma^2)$, where x_t is a vector of observed covariates, β is a $k \times 1$ vector of regression coefficients and $\varepsilon_t \sim$

$\mathcal{N}(0, \sigma^2)$. To express the likelihood, it is useful to stacking the data into matrices: $y = X\beta + \varepsilon$, where Y is a $T \times 1$ vector of dependent variables, X is a $T \times k$ matrix of regressor variables and ε, where Y is a $T \times 1$ vector of errors.

The usual OLS regression estimator is $\widehat{\beta} = (X'X)^{-1} X'y$ and the residual sum of squares is $S = (y - X\hat{\beta})'(y - X\hat{\beta})$. Completing the square,

$$\begin{aligned}(y - X\beta)'(y - X\beta) &= (\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta}) + (y - X\hat{\beta})'(y - X\hat{\beta}) \\ &= (\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta}) + S,\end{aligned}$$

which implies that

$$\begin{aligned}p(y|\beta, \sigma^2) &= \left(\frac{1}{\sigma^2}\right)^{T/2} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right) \\ &= \left(\frac{1}{\sigma^2}\right)^{T/2} \exp\left(-\frac{1}{2\sigma^2}(\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta}) - \frac{S}{2\sigma^2}\right)\end{aligned}$$

Hence $\left(\hat{\beta}, S\right)$ are sufficient statistics for (β, σ^2).

- A proper conjugate prior is $p(\beta|\sigma^2) \sim \mathcal{N}_k(a, \sigma^2 A)$ and $p(\sigma^2) \sim \mathcal{IG}\left(\frac{b}{2}, \frac{B}{2}\right)$ with

$$p\left(\beta|\sigma^2\right) = (2\pi)^{\frac{k}{2}} \left(\sigma^2\right)^{-\frac{k}{2}} |A|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(\beta - a)'(A)^{-1}(\beta - a)\right)$$

since $|\sigma^2 A| = (\sigma^2)^k |A|$. The posterior distribution is

$$p\left(\beta, \sigma^2|y\right) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{b+T}{2}+1} \left(\frac{1}{\sigma^2}\right)^{\frac{k}{2}} \exp\left(-\frac{1}{2\sigma^2} Q\left(\beta, \sigma^2\right)\right),$$

where the quadratic form is defined by

$$Q\left(\beta, \sigma^2\right) = (y - X\beta)'(y - X\beta) + (\beta - a)' A^{-1}(\beta - a) + B.$$

we can complete the square as follows: combine $(y - X\beta)'(y - X\beta) + (\beta - a)' A^{-1}(\beta - a)$

by stacking vectors

$$\tilde{y} = \begin{pmatrix} y \\ A^{-\frac{1}{2}} a \end{pmatrix} \text{ and } W = \begin{pmatrix} X \\ A^{-\frac{1}{2}} \end{pmatrix},$$

where $\tilde{y}$ is a $(T+k) \times 1$ vector and W is a $(T+k) \times k$ matrix. Then,

$$(y - X\beta)'(y - X\beta) + (\beta - a)' A^{-1} (\beta - a) = (\tilde{y} - W\beta)'(\tilde{y} - W\beta).$$

By analogy to the previous case, define $\overline{\beta} = (W'W)^{-1} W'\tilde{y}$, where

$$\begin{aligned} W'W &= X'X + A^{-1} \\ W'\tilde{y} &= X'y + A^{-1}a. \end{aligned}$$

Adding and subtracting $W\overline{\beta}$ to $\tilde{y} - W\beta$ and simplifying gives

$$\begin{aligned} (\tilde{y} - W\beta)'(\tilde{y} - W\beta) &= \left[\left(\tilde{y} - W\overline{\beta}\right) + \left(W\overline{\beta} - W\beta\right)\right]' \left[\left(\tilde{y} - W\overline{\beta}\right) + \left(W\overline{\beta} - W\beta\right)\right] \\ &= \left(\beta - \overline{\beta}\right)' W'W \left(\beta - \overline{\beta}\right) + \left(\tilde{y} - W\overline{\beta}\right)' \left(\tilde{y} - W\overline{\beta}\right) \\ &= (\beta - a_T)' A_T^{-1} (\beta - a_T) + (\tilde{y} - Wa_T)'(\tilde{y} - Wa_T), \end{aligned}$$

where $A_T^{-1} = W'W = X'X + A^{-1}$ and $a_T = A_T [X'y + A^{-1}a]$. Straightforward algebra shows that $\left(\tilde{y} - W\overline{\beta}\right)' \left(\tilde{y} - W\overline{\beta}\right) = y'y + a'A^{-1}a - a_T'A_T^{-1}a_T$. Putting the pieces together, the quadrtic form is

$$\begin{aligned} Q\left(\beta, \sigma^2\right) &= (\beta - a_T)' A_T^{-1} (\beta - a_T) + y'y + a'A^{-1}a - a_T'A_T^{-1}a_T + B \\ &= (\beta - a_T)' A_T^{-1} (\beta - a_T) + B_T, \end{aligned}$$

where $B_T = y'y + a'A^{-1}a - a_T'A_T^{-1}a_T + B$. This leads to the same posterior.

3 Direct sampling

3.1 Generating i.i.d. random variables from distributions

The Gibbs sampler and MH algorithms require simulating i.i.d. random variables from "recognizable" distributions. Appendix 1 provides a list of common "recognizable" dis-

tributions, along with methods for generating random variables from these distributions. This section briefly reviews that standard methods and approaches for simulating random variables from recognizable distributions.

Most of these algorithms first generate random variables from a relatively simple "building block" distribution, such as a uniform or normal distribution, and then transform these draws to obtain a sample from another distribution. This section describes a number of these approaches that are commonly encountered in practice. Most "random-number" generators actual use deterministic methods, along with transformations. In this regard, it is important to remember Von Neumann's famous quotation: *"Anyone who attempts to generate random numbers by deterministic means is, of course, living in a state of sin."*

Inverse CDF method The inverse distribution method uses samples of uniform random variables to generate draws from random variables with a continuous distribution function, F. Since $F(x)$ is uniformly distributed on $[0,1]$, draw a uniform random variable and invert the CDF to get a draw from F. Thus, to sample from F,

$$\begin{aligned} &\text{Step 1: Draw } U \sim U[0,1] \\ &\text{Step 2: Set } X = F^{-1}(U), \end{aligned}$$

where $F^{-1}(U) = \inf\{x : F(x) = U\}$.

This inversion method provides i.i.d. draws from F provided that $F^{-1}(U)$ can be exactly calculated. For example, the CDF of an exponential random variable with parameter μ is $F(x) = 1 - \exp(-\mu x)$, which can easily be inverted. When F^{-1} cannot be analytically calculated, approximate inversions can be used. For example, suppose that the density is a known analytical function. Then, $F(x)$ can be computed to an arbitrary degree of accuracy on a grid and inversions can be approximately calculated, generating an approximate draw from F. With all approximations, there is a natural trade-off between computational speed and accuracy. One example where efficient approximations are possible are inversions involving normal distributions, which is useful for generating truncated normal random variables. Outside of these limited cases, the inverse transform method does not provide a computationally attractive approach for drawing random variables from a given distribution function. In particular, it does not work well in multiple dimensions.

Functional Transformations The second main method uses functional transformations to express the distribution of a random variable that is a known function of another random variable. Suppose that $X \sim F$, admitting a density f, and that $y = h(x)$ is an increasing continuous function. Thus, we can define $x = h^{-1}(y)$ as the inverse of the function h. The distribution of y is given by

$$F(a) = \text{Prob}(Y \leq y) = \int_{-\infty}^{h^{-1}(y)} f(x)\, dx = F\left(X \leq h^{-1}(y)\right).$$

Differentiating with respect to y gives the density via Leibnitz's rule:

$$f_Y(y) = f\left(h^{-1}(y)\right) \left| \frac{dh^{-1}(y)}{dy} \right|,$$

where we make explicit that the density is over the random variable Y. This result is used widely. For example, if $X \sim \mathcal{N}(0,1)$, then $Y = \mu + \sigma X$. Since $x = h^{-1}(y) = \frac{y-\mu}{\sigma}$, the distribution function is $F\left(\frac{x-\mu}{\sigma}\right)$ and density

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

Transformations are widely used to simulate both univariate and multivariate random variables. As examples, if $Y \sim \mathcal{X}^2_\nu$ and ν is an integer, then $Y = \sum_{i=1}^{\nu} X_i^2$ where each X_i is independent standard normal. Exponential random variables can be used to simulate $\mathcal{X}^2$, Gamma, beta, and Poisson random variables. The famous Box-Muller algorithm simulates normals from uniform and exponential random variables. In the multivariate setting, Wishart (and inverse Wishart) random variables can be via sums of squared vectors of standard normal random variables.

Mixture distributions In the multidimensional case, a special case of the transformation generates continuous mixture distributions. The density of a continuous mixture distribution is given by

$$p(x) = \int p(x|\lambda)\, p(\lambda)\, d\lambda,$$

where $p(x|\lambda)$ is viewed as density conditional on the parameter λ. One example of this is the class of scale mixtures of normal distributions, where

$$p(x|\lambda) = \frac{1}{\sqrt{2\pi\lambda}} \exp\left(-\frac{x^2}{2\lambda}\right)$$

and, λ is the conditional variance of X. It is often simpler just to write

$$X = \sqrt{\lambda}\varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0,1)$. The distribution of $\sqrt{\lambda}$ determines the marginal distribution of X. Here are a number of examples of scale mixture distributions.

- **T-distribution.** The t-distribution arises in many Bayesian inference problems involving inverse Gamma priors and conditionally normally distributed likelihoods. If $p(y_t|\lambda) \sim \mathcal{N}(0,\lambda)$ and $p(\lambda) \sim \mathcal{IG}\left(\frac{b}{2}, \frac{B}{2}\right)$, then the marginal distribution of y_t is $t_b\left(0, \frac{B}{b}\right)$. The proof is direct by analytically computing the marginal distribution,

  $$p(y_t) = \int_0^\infty p(y_t|\lambda)\, p(\lambda)\, d\lambda.$$

  Using our integration results:

  $$p(y) = \int_0^\infty \underbrace{\frac{1}{\sqrt{2\pi}} \left(\frac{1}{\lambda}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{y^2}{\lambda}\right)}_{likelihood} \underbrace{\frac{(B/2)^{b/2}}{\Gamma\left(\frac{b}{2}\right)} \left(\frac{1}{\lambda}\right)^{\frac{b}{2}+1} \exp\left(-\frac{B}{2\lambda}\right)}_{prior} d\lambda$$
  $$\propto \int_0^\infty \left(\frac{1}{\lambda}\right)^{\left(\frac{b+1}{2}\right)+1} \exp\left(-\frac{1}{\lambda}\frac{y^2+B}{2}\right) d\lambda$$

  which is in the class of scale mixture integrals. We obtain

  $$p(y) \propto \left[1 + \frac{y^2}{B}\right]^{-\left(\frac{b+1}{2}\right)}$$

  Thus $y_t \sim t_b(0,B)$. Given μ, $y_t|\mu,\lambda,\sigma^2 \sim \mathcal{N}(\mu,\lambda\sigma^2)$ and $\lambda \sim \mathcal{IG}\left(\frac{b}{2}, \frac{B}{2}\right)$ which implies $p(y_t|\mu,\sigma^2) \sim t_b(\mu,B)$.

- **Double-exponential distribution**. The double exponential arises as a scale mix-

ture distribution: If $p\left(y_t|\lambda\right) \sim \mathcal{N}\left(0,\lambda\right)$ and $\lambda \sim \exp\left(2\right)$, then the marginal distribution of $y_t \sim \mathcal{DE}\left(0,1\right)$. The proof is again by direct integration using the results in our integration appendix:

$$\int_0^\infty \frac{1}{\sqrt{2\pi\lambda}} \exp\left\{-\frac{1}{2}\left(\frac{y^2}{\lambda} - \lambda\right)\right\} d\lambda = \frac{1}{2}\exp\left(-\left|y\right|\right),$$

More generally, if μ and σ^2 are known, then $y_t|\mu,\lambda,\sigma^2 \sim \mathcal{N}\left(\mu,\lambda\sigma^2\right)$ and $\lambda \sim \exp\left(2\right)$, then $p\left(y_t|\mu,\sigma^2\right) \sim \mathcal{DE}\left(\mu,\sigma^2\right)$ substituting $b = \left(y-\mu\right)/\sigma$ and multiplying both sides by σ^{-1}.

- **Asymmetric Laplacean.** The asymmetric Laplacean distribution is a scale mixture of normal distributions: if

$$p\left(y_t|\lambda\right) \sim \mathcal{N}\left(\left(2\tau-1\right)\lambda,\lambda\right) \text{ and } \lambda \sim \exp\left(\mu_\tau^{-1}\right),$$

then $y_t \sim \mathcal{CE}\left(\tau,0,1\right)$. The proof uses our integration appendix and

$$\int_0^\infty \frac{1}{\sqrt{2\pi\lambda}} \exp\left\{-\frac{1}{2\lambda}\left(y+\left(2\tau-1\right)\lambda\right)^2 - 2\tau\left(1-\tau\right)\lambda\right\} d\lambda = \frac{1}{2}\exp\left(-\left|y\right| - \left(2\tau-1\right)y\right)$$

In general, if μ and σ^2 are known, then

$$p(y_t|\mu,\lambda,\sigma^2) \sim \mathcal{N}\left(\mu+\left(1-2\tau\right)\lambda,\lambda\sigma^2\right) \text{ and } \lambda \sim \exp\left(\mu_\tau^{-1}\right)$$

which leads to $p\left(y_t|\mu,\sigma^2\right) \sim \mathcal{CE}\left(\tau,\mu,\sigma^2\right)$.

- **Exponential power family.** This family of distributions (Box and Tiao, 1973) $p(x|\tau,\gamma)$ is given by

$$p(y|\sigma,\gamma) = \left(2\sigma\right)^{-1} c(\gamma) \exp\left(-\left|\frac{y}{\sigma}\right|^\gamma\right)$$

where $c(\gamma) = \Gamma(1+\gamma^{-1})^{-1}$. Following West (1987) we have the following scale mixtures of normals representation

$$p(y|\sigma=1,\gamma) = c(\gamma)\exp\left(-|\beta|^\gamma\right) = \int_0^\infty \frac{1}{\sqrt{2\pi\lambda}} \exp\left(-\frac{y^2}{2\lambda}\right) p(\lambda|\gamma) d\lambda$$

where $p(\lambda|\gamma) \propto \lambda^{-3/2}\mathrm{St}^{+}_{\frac{\gamma}{2}}(\lambda^{-1})$ and St^{+}_{a} is the (analytically intractable) density of a positive stable distribution.

Factorization Method Another method that is useful in some multivariate settings is known as factorization. The rules of probability imply that a joint density, $p(x_1, ..., x_n)$, can always be factored as

$$\begin{aligned} p(x_1, ..., x_n) &= p(x_n|x_1, ..., x_{n-1})\, p(x_1, ..., x_{n-1}) \\ &= p(x_1)\, p(x_2|x_1) \cdots p(x_n|x_1, ..., x_{n-1}). \end{aligned}$$

In this case, simulating $X_1 \sim p(x_1)$, $X_2 \sim p(x_2|X_1)$, ... $X_n \sim p(x_n|X_1, ..., X_{n-1})$ generates a draw from the joint distribution. This procedure is common in Bayesian statistics, and the distribution of X_1 is a marginal distribution and the other distributions are conditional distributions. This is used repeatedly to express a joint posterior in terms of lower dimensional conditional posteriors.

Rejection sampling The final general method discussed here is the accept-reject method, developed by von Neumann. Suppose that the goal is to generate a sample from $f(x)$, where it is assumed that f is bounded, that is, $f(x) \leq cg(x)$ for some c. The accept-reject algorithm is a two-step procedure

Step 1: Draw $U \sim \mathcal{U}[0, 1]$ and $X \sim g$

Step 2: Accept $Y = X$ if $U \leq \dfrac{f(X)}{cg(X)}$, otherwise return to Step 1.

Rejection sampling simulates repeatedly until a draw that satisfies $U \leq \frac{f(X)}{cg(X)}$ is found. By direct calculation, it is clear that Y has density f:

$$\begin{aligned} \text{Prob}(Y \leq y) &= \text{Prob}\left(X \leq y | U \leq \frac{f(X)}{cg(X)}\right) = \frac{\text{Prob}\left(X \leq y, U \leq \frac{f(X)}{cg(X)}\right)}{\text{Prob}\left(U \leq \frac{f(X)}{cg(X)}\right)} \\ &= \frac{\int_{-\infty}^{y}\left(\int_{-\infty}^{f(x)/cg(X)} du\right) g(x)\, dx}{\int_{-\infty}^{-\infty}\left(\int_{-\infty}^{f(x)/cg(X)} du\right) g(x)\, dx} = \frac{\frac{1}{c}\int_{-\infty}^{y} f(x)\, dx}{\frac{1}{c}\int_{-\infty}^{-\infty} f(x)\, dx} = \int_{-\infty}^{y} f(x)\, dx. \end{aligned}$$

Rejection sampling requires (a) a bounding or dominating density g, (b) an ability to evaluate the ratio f/g; (c) an ability to simulate i.i.d. draws from g, and (d) the bounding constant c. Rejection sampling does not require that the normalization constant $\int f(x)\,dx$ be known, since the algorithm only requires knowledge of f/c. For continuous densities on a bounded support, it is to satisfy (a) and (c) (a uniform density works), but for continuous densities on unbounded support it can be more difficult since we need to find a density with heavier tails and higher peaks. Setting $c = \sup_x \frac{f(x)}{g(x)}$ maximizes the acceptance probability. In practice, finding the constant is difficult because f generally depends on a multi-dimensional parameter vector, $f(x|\theta_f)$, and thus the bounding is over x and θ_f.

Rejection sampling is often used to generate random variables from various recognizable distributions, such as the Gamma or beta distributions. In these cases, the structure of the densities is well known and the bounding density can be tailored to generate fast and efficient rejection sampling algorithms. In many of these cases, the densities are log-concave (e.g., Normal, Double exponential, Gamma, and Beta). A density f is log-concave if $\ln(f(x))$. For differentiable densities, this is equivalent to assuming that $\frac{dlnf(x)}{dx} = \frac{f\prime(x)}{f(x)}$ is non-increasing in x and $\frac{d^2lnf(x)}{dx^2} < 0$. Under these conditions, it is possible to develop "black-box" generation methods that perform well in many settings. Another modification that "adapts" the dominating densities works well for log-concave densities.

The basis of the rejection sampling is the simple fact that since

$$f(x) = \int_0^{f(x)} du = \int_0^{f(x)} U(du),$$

where U is the uniform distribution function, the density f is a marginal distribution from the joint distribution

$$(X, U) \sim \mathcal{U}((x, u) : 0 \leq u \leq f(x)).$$

More generally, if $X \in R^d$ is a random vector with density f and U is an independent random variable distributed $\mathcal{U}[0, 1]$, then $(X, cUf(X))$ is uniformly distributed on the set $A = \{(x, u) : x \in R^d, 0 \leq u \leq cf(x)\}$. Conversely, if (X, U) is uniformly distributed on A, then the density of X is $f(X)$.

Multinomial Resampling Sampling N values from a discrete distribution (x_i, p_i) can be done by simulating standard uniforms U_i and then using binary search to find the

value of j, and hence x_j, corresponding to

$$q_{j-1} < U_i \leq q_j$$

where $q_j = \sum_{l=0}^{j} p_j$ and $q_0 = 0$. This algorithm is inefficient, although commonly used due to its simplicity, as requires $O(N \ln N)$ operations. The $\ln N$ operations comes from the binary search.

A more efficient method (particularly suited to the particle filtering methods) is to simulate $N+1$ exponentially distributed variables $z_0, \ldots, z_N$ via $z_i = -\ln U_i$ and calculating the totals $Z_j = \sum_{l=0}^{j} z_l$ and then merging Z_j and q_j in the sense that if $q_j Z_N > Z_i$ then output x_j. This is an $O(N)$ algorithm.

Slice Sampling Slice sampling can be used to sample the conditional posterior when the scale mixture is a stable distribution. Here we have

$$p\left(\lambda_t|\sigma^2, y_t\right) \sim p(y_t|\lambda_t, \sigma^2) p\left(\lambda_t\right),$$

where $p\left(y_t|\lambda_t, \sigma^2\right) = N\left(0, \sigma^2\lambda_t\right)$. To do this, introduce a uniform random variable u_t and consider the joint distribution

$$p\left(\lambda_t, u_t|\sigma^2\right) \propto p\left(\lambda_t\right) U_{u_t}\left[0, \phi\left(y_t', \lambda_t\sigma^2\right)\right].$$

Then the algorithm is:

$$\begin{aligned} &p\left(\lambda_t|u_t, \sigma^2\right) \propto p\left(\lambda_t\right) \text{ on } \mathcal{N}\left(y_t; 0, \sigma^2\lambda_t\right) > u_t \\ &p\left(u_t|\lambda_t, \sigma^2\right) \sim \mathcal{U}\left[0, \phi\left(y_t; \lambda_t\sigma^2\right)\right]. \end{aligned}$$

The accept/reject sampling alternative is as follows. Since we have the upper bound $\phi\left(y_t; 0, \lambda_t\sigma^2\right) \leq \left(2\pi y_t^2\right)^{-1/2} \exp(-1/2)$, the sample λ_t via

$$\begin{aligned} \lambda_t &\sim p\left(\lambda_t\right) \\ u &\sim U\left[0, \left(2\pi y_t^2\right)^{-1/2} e^{-1/2}\right], \end{aligned}$$

and if $U > \phi\left(y_t, \sigma^2\lambda_t\right)$, repeat.

3.2 Integration results

There are a number of useful integration results that are repeatedly used for normalising probability distributions. The first identity is

$$\int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx = \sqrt{2\pi\sigma^2},$$

for all μ which defines the univariate normal distribution. The second identity is for the Gamma function, defined as

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy,$$

which is important for a number of distributions, most notably the Gamma and inverse Gamma. This implies that $y^{\alpha-1}e^{-y}/\Gamma(\alpha)$ is a proper density. Changing variables to $x = \beta y$, gives the standard form of the Gamma distribution.

Integration by parts implies $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, so $\Gamma(\alpha) = (\alpha-1)!$ with $\Gamma(1) = 1$ and $\Gamma(1/2) = \sqrt{\pi}$. For other fractional values

$$\Gamma(\alpha)\Gamma\left(\alpha+\frac{1}{2}\right) = 2^{1-2\alpha}\sqrt{\pi}\Gamma(2\alpha),$$

A number of integrals are useful for analytic characterization of distributions, either priors or posteriors, that are scale mixtures of normals. As discussed in the previous appendix, scale mixtures involve integrals that are a product of two distributions. The following integral identities are useful for analyzing these specifications. For any given $p, a, b > 0$,

$$\int_0^{\infty} x^{p-1} \exp\left(-ax^b\right) dx = \frac{1}{b} a^{-\frac{p}{b}} \Gamma\left(\frac{p}{b}\right)$$
$$\int_0^{\infty} \left(\frac{1}{x}\right)^{p+1} \exp\left(-ax^{-b}\right) dx = \frac{1}{b} a^{-\frac{p}{b}} \Gamma\left(\frac{p}{b}\right)$$

These integrals are useful when combining Gamma or inverse Gamma distributions prior distributions with normal likelihood functions. Second, for any a and b,

$$\int_0^{\infty} \frac{a}{\sqrt{2\pi x}} \exp\left\{-\frac{1}{2}\left(a^2 x + \frac{b^2}{x}\right)\right\} dx = \exp\left(-|ab|\right),$$

which are useful for double exponential. A related integral, which is useful for the check

exponential distribution is

$$\int_0^\infty \frac{a}{\sqrt{2\pi x}} \exp\left\{-\frac{1}{2}\left(\frac{b^2}{x} + 2(2\tau-1)b + a^2 x\right)\right\} dx = \exp\left(-|ab| - (2\tau-1)b\right).$$

3.3 Useful algebraic formulae

One of the most useful algebraic tricks for calculating posterior distribution is *completing the square* (have fun showing the algebra for the second part!).

In the scalar case, we have the identity

$$\frac{(x-\mu_1)^2}{\Sigma_1} + \frac{(x-\mu_2)^2}{\Sigma_2} = \frac{(x-\mu_3)^2}{\Sigma_3} + \frac{(\mu_1-\mu_2)^2}{\Sigma_1+\Sigma_2}$$

where $\mu_3 = \Sigma_3\left(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2\right)$ and $\Sigma_3 = \left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)^{-1}$.

A shrinkage interpretation is the following. When analyzing shrinkage and it is common to define the weight $w = (\Sigma_1 + \Sigma_2)^{-1}\Sigma_1$. We can now re-write *completing the square* as

$$\frac{(x-\mu_1)^2}{\Sigma_1} + \frac{(x-\mu_2)^2}{\Sigma_2} = \frac{(\mu_1-\mu_2)^2}{\Sigma_1(1-w)^{-1}} + \frac{(x-(1-w)\mu_1 - w\mu_2)^2}{\Sigma_2(1-w)}.$$

In vector-matrix case, completing the square becomes

$$(x-\mu_1)'\Sigma_1^{-1}(x-\mu_1) + (x-\mu_2)'\Sigma_2^{-1}(x-\mu_2) = (x-\mu_3)'\Sigma_3^{-1}(x-\mu_3) + (\mu_1-\mu_2)'\Sigma_4^{-1}(\mu_1-\mu_2)$$

where $\mu_3 = \Sigma_3\left(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2\right)$ and

$$\begin{aligned}\Sigma_3 &= \left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)^{-1} \\ \Sigma_4 &= \Sigma_1^{-1}\left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)^{-1}\Sigma_2^{-1}.\end{aligned}$$

Completing the square for the convolution of two normal densities is also useful when interpreting the fundamental identity of likelihood times prior is equal to posterior times marginal. The identity yields

$$\phi(\mu_1, \Sigma_1)\,\phi(\mu_2, \Sigma_2) = \phi(\mu, \Sigma)\, c(\mu_1, \Sigma_1, \mu_2, \Sigma_2)$$

with ϕ is the standard normal density and

$$c(\mu_1, \Sigma_1, \mu_2, \Sigma_2) = (2\pi)^{-\frac{k}{2}} |\Sigma_1 + \Sigma_2|^{-\frac{k}{2}} \exp\left(-\frac{1}{2}(\mu_2 - \mu_1)'(\Sigma_1 + \Sigma_2)^{-1}(\mu_2 - \mu_1)\right)$$

with parameters $\mu = \Sigma\left(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2\right)$ and $\Sigma = \left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)^{-1}$

3.4 The EM, ECM, and ECME algorithms

MCMC methods have been used extensively to perform numerical integration. There is also interest in using simulation-based methods to optimise functions. The EM algorithm is a algorithms in a general class of Q-maximisation algorithms that finds a (deterministic) sequence $\{\theta^{(g)}\}$ converging to $\arg\max_{\theta\in\theta} Q(\theta)$.

First, define a function $Q(\theta, \phi)$ such that $Q(\theta) = Q(\theta, \theta)$ and it satisfies a convexity constraint $Q(\theta, \phi) \geq Q(\theta, \theta)$. Then define

$$\theta^{(g+1)} = \arg\max_{\theta\in\theta} Q(\theta, \theta^{(g)})$$

This satisfies the convexity constraint $Q(\theta, \theta) \geq Q(\theta, \varphi)$ for any φ. In order to prove convergence. you get a sequence of inequalities

$$Q(\theta^{(0)}, \theta^{(0)}) \leq Q(\theta^{(1)}, \theta^{(0)}) \leq Q(\theta^{(1)}, \theta^{(1)}) \leq \ldots \leq Q$$

In many models we have to deal with a latent variable and require estimation where integration is also involved. For example, suppose that we have a triple (y, z, θ) with joint probability specification $p(y, z, \theta) = p(y|z, \theta)p(z, \theta)$. This can occur in missing data problems and estimation problems in mixture models.

A standard application of the EM algorithm is to find

$$\arg\max_{\theta\in\theta} \int_z p(y|z, \theta)p(z|\theta)dz$$

As we are just finding an optimum, you do not need the prior specification $p(\theta)$. The EM algorithm finds a sequence of parameter values $\theta^{(g)}$ by alternating between an expectation and a maximsation step. This still requires the numerical (or analytical) computation of the criteria function $Q(\theta, \theta^{(g)})$ described below.

EM algorithms have been used in extensively in mixture models and missing data problems. The EM algorithm uses the particular choice where

$$Q(\theta) = \log p(y|\theta) = \log \int p(y, z|\theta) dz$$

Here the likelihood has a mixture representation where z is the latent variable (missing data, state variable etc). This is termed a Q-maximization algorithm with:

$$Q(\theta, \theta^{(g)}) = \int \log p(y|z, \theta) p(z|\theta^{(g)}, y) dz = E_{z|\theta^{(g)}, y}[\log p(y|z, \theta)]$$

To implement EM you need to be able to calculate $Q(\theta, \theta^{(g)})$ and optimize at each iteration.

The EM algorithm and its extensions ECM and ECME are methods of computing maximum likelihood estimates or posterior modes in the presence of missing data. Let the objective function be $l(\theta) = \log p(\theta|y) + c(y)$, where $c(y)$ is a possibly unknown normalizing constant that does not depend on β and y denotes observed data. We have a mixture representation,

$$p(\theta|y) = \int p(\theta, z|y) dz = \int p(\theta|z, y) p(z|y) dz$$

where distribution of the latent variables is $p(z|\theta, y) = p(y|\theta, z)p(z|\theta)/p(y|\theta)$.

In some cases the complete data log-posterior is simple enough for $\arg\max_\theta \log p(\theta|z, y)$ to be computed in closed form. The EM algorithm alternates between the Expectation and Maximization steps for which it is named. The E-step and M-step computes

$$Q(\beta|\beta^{(g)}) = E_{z|\beta^{(g)}, y}\left[\log p(y, z|\beta)\right] = \int \log p(y, z|\beta) p(z|\beta^{(g)}, y) dz$$
$$\beta^{(g+1)} = \arg\max_\beta Q(\beta|\beta^{(g)})$$

This has an important monotonicity property that ensures $\ell(\beta^{(g)}) \leq \ell(\beta^{(g+1)})$ for all g. In fact, the monotonicity proof given by Dempster et al (1977) shows that any β with $Q(\beta, \beta^{(g)}) \geq Q(\beta^{(g)}, \beta^{(g)})$ also satisfies the log-likelihood inequality $\ell(\beta) \geq \ell(\beta^{(g)})$.

In problems with many parameters the M-step of EM may be difficult. In this case θ may be partitioned into components $(\theta_1, \ldots, \theta_k)$ in such a way that maximizing $\log p(\theta_j|\theta_{-j}, z, y)$ is easy. The ECM algorithm pairs the EM algorithm's E-step with k conditional maximization (CM) steps, each maximizing Q over one component θ_j with each component of θ_{-j}

fixed at the most recent value. Due to the fact that each CM step increases Q the ECM algorithm retains the monotonicity property. The ECME algorithm replaces some of ECM's CM steps with maximizations over l instead of Q. Liu and Rubin (1994) show that doing so can greatly increase the rate of convergence.

In many cases we will have a parameter vector $\theta = (\beta, \nu)$ partitioned into its components and a missing data vector $z = (\lambda, \omega)$. Then we compute the $Q(\beta, \nu|\beta^{(g)}, \nu^{(g)})$ objective function and then compute E- and M steps from this to provide an iterative algorithm for updating parameters. To update the hyperparameter ν we can maximize the fully data posterior $p(\beta, \nu|y)$ with β fixed at $\beta^{(g+1)}$. The algorithm can be summarized as follows

$$\beta^{(g+1)} = \text{argmax}_\beta Q(\beta|\beta^{(g)}, \nu^{(g)}) \text{ where } Q(\beta|\beta^{(g)}, \nu^{(g)}) = E_{z\ \beta^{(g)},\nu^{(g)},y}\left[\log p(y, z|\beta, \nu^{(g)})\right]$$
$$\nu^{(g+1)} = \text{argmax}_\nu \log p(\beta^{(g+1)}, \nu|y)$$

Simulated Annealing (SA) A simulation-based approach to finding $\hat{\theta} = \arg\max_{\theta\in\Theta} H(\theta)$ is to sample a sequence densities

$$\pi_J(\theta) = \frac{e^{-JH(\theta)}}{\int e^{JH(\theta)} d\mu(\theta)}$$

where J is a temperature parameter. Instead of looking at derivatives and performing gradient-based optimization you can simulate from the sequence of densities. This forms a time-homogeneous Markov chain and under suitable regularly conditions on the relaxation schedule for the temperature we have $\theta^{(g)} \to \hat{\theta}$. The main caveat is that we need to know the criterion function $H(\theta)$ to evaluate the Metropolis probability for sampling from the sequence of densities. This is not always available.

An interesting generalisation which is appropriate in latent variable mixture models is the following. Suppose that $H(\theta) = E_{z|\theta}\{H(z, \theta)\}$ is unavailable in closed-form where without loss of generality we assume that $H(z, \theta) \geq 0$. In this case we can use latent variable simulated annealing (LVSA) methods. Define a joint probability distribution for $z^J = (z_1, \ldots, z_J)$ as

$$\pi_J(z^J, \theta) \propto \prod_{j=1}^{J} H(z_j, \theta) p(z_j|\theta)\mu(\theta)$$

for some measure μ which ensures integrability of the joint. This distribution has the

property that its marginal distribution on θ is given by

$$\pi_J(\theta) \propto E_{z|\theta}\left\{H(z,\theta)\right\}^J \mu(\theta) = e^{J \ln H(\theta)} \mu(\theta)$$

By the simulated annealing argument we see that this marginal collapses on the maximum of $\ln H(\theta)$. The advantage of this approach is that it is typically straightforward to sample with MCMC from the conditionals

$$\pi_J(z_i|\theta) \sim H(z_j,\theta)p(z_j|\theta) \text{ and } \pi_J(\theta|z) \sim \prod_{j=1}^{J} H(z_j,\theta)p(z_j|\theta)\mu(\theta)$$

Jacquier, Johannes and Polson (2007) apply this to finding MLE estimates for commonly encountered latent variable models.

3.5 Notes and Discussion

Geman (1987) provides a monograph discussion of Markov chain optimisation methods. Pincus (1968) provides an early application of the Metropolis algorithm to simulated annealing optimisation. Standard references on simulated annealing include Kirkpatrick (1984), Kirkpatrick et al, (1983), Aarts and Korst (1989) Van Laarhoven and Aarts, (1987). Muller (2000) and Mueller et al (2004) propose this version of latent variable simulated annealing for dealing with the joint problem of integration and optimisation. Slice sampling applications in Statistics are dicussed in Besag and Green (1993), Polson, (1996), Damien et al, (1999) and Neal (2003) provides a general algorithm. General strategies based on the ratio-of-uniforms method are in Wakefield et al (1991). Devroye (1986) and Ripley (1987) discuss many tailored made algorithms for simulation.

The EM algorithm for missing data problems was originally developed by Dempster, Laird and Rubin (1977) but has its roots in the hidden Markov model literature and the Baum-Welch algorithm. Liu and Rubin (1994) consider faster ECME algorothms. Liang and Wong (2001) and Liu, Liang and Wong (2000) consider extensions including evolutionary MCMC algorithms.