Contents

6
Conjugate families

Conjugate families arise when the likelihood times the prior produces a recognizable posterior kernel

$$p(\theta \mid y) \propto \ell(\theta \mid y)\, p(\theta)$$

where the *kernel* is the characteristic part of the distribution function that depends on the random variable(s) (the part excluding any normalizing constants). For example, the density function for a univariate Gaussian or normal is

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$$

and its kernel (for σ known) is

$$\exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$$

as $\frac{1}{\sqrt{2\pi}\sigma}$ is a normalizing constant. Now, we discuss a few common conjugate family results[1] and uninformative prior results to connect with classical results.

[1] A more complete set of conjugate families are summarized in chapter 7 of *Accounting and Causal Effects: Econometric Challenges* as well as tabulated in an appendix at the end of the chapter.

6.1 Binomial - beta prior

A binomial likelihood with unknown success probability, θ,

$$\ell(\theta \mid s; n) = \binom{n}{s} \theta^s (1-\theta)^{n-s}$$

$$s = \textstyle\sum_{i=1}^{n} y_i, \quad y_i = \{0, 1\}$$

combines with a beta$(\theta; a, b)$ prior (i.e., with parameters a and b)

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

to yield

$$\begin{aligned} p(\theta \mid y) &\propto \theta^s (1-\theta)^{n-s} \theta^{a-1} (1-\theta)^{b-1} \\ &\propto \theta^{s+a-1} (1-\theta)^{n-s+b-1} \end{aligned}$$

which is the kernel of a beta distribution with parameters $(a+s)$ and $(b+n-s)$, beta$(\theta \mid y; a+s, b+n-s)$.

6.1.1 Uninformative priors

Suppose priors for θ are uniform over the interval zero to one or, equivalently, beta$(1,1)$.[2] Then, the likelihood determines the posterior distribution for θ.

$$p(\theta \mid y) \propto \theta^s (1-\theta)^{n-s}$$

which is beta$(\theta \mid y; 1+s, 1+n-s)$.

6.2 Gaussian (unknown mean, known variance)

A single draw from a Gaussian likelihood with unknown mean, θ, known standard deviation, σ,

$$\ell(\theta \mid y, \sigma) \propto \exp\left[-\frac{1}{2}\frac{(y-\theta)^2}{\sigma^2}\right]$$

combines with a Gaussian or normal prior for θ given σ^2 with prior mean θ_0 and prior variance τ_0^2

$$p\left(\theta \mid \sigma^2; \theta_0, \tau_0^2\right) \propto \exp\left[-\frac{1}{2}\frac{(\theta-\theta_0)^2}{\tau_0^2}\right]$$

[2] Some would utilize Jeffreys' prior, $p(\theta) \propto \text{beta}\left(\theta; \frac{1}{2}, \frac{1}{2}\right)$, which is invariant to transformation, as the uninformative prior.

or writing $\tau_0^2 \equiv \sigma^2/\kappa_0$, we have

$$p\left(\theta \mid \sigma^2; \theta_0, \sigma^2/\kappa_0\right) \propto \exp\left[-\frac{1}{2}\frac{\kappa_0\left(\theta-\theta_0\right)^2}{\sigma^2}\right]$$

to yield

$$p\left(\theta \mid y, \sigma, \theta_0, \sigma^2/\kappa_0\right) \propto \exp\left[-\frac{1}{2}\left(\frac{\left(y-\theta\right)^2}{\sigma^2}+\frac{\kappa_0\left(\theta-\theta_0\right)^2}{\sigma^2}\right)\right]$$

Expansion and rearrangement gives

$$p\left(\theta \mid y, \sigma, \theta_0, \sigma^2/\kappa_0\right) \propto \exp\left[-\frac{1}{2\sigma^2}\left(y^2+\kappa_0\theta_0^2-2y\theta+\theta^2+\kappa_0\left(\theta^2-2\theta\theta_0\right)\right)\right]$$

Any terms not involving θ are constants and can be discarded as they are absorbed on normalization of the posterior

$$p\left(\theta \mid y, \sigma, \theta_0, \sigma^2/\kappa_0\right) \propto \exp\left[-\frac{1}{2\sigma^2}\left(\theta^2\left(\kappa_0+1\right)-2\theta\left(\kappa_0\theta_0+y\right)\right)\right]$$

Completing the square (add and subtract $\frac{\left(\kappa_0\theta_0+y\right)^2}{\kappa_0+1}$), dropping the term subtracted (as it's all constants), and factoring out (κ_0+1) gives

$$p\left(\theta \mid y, \sigma, \theta_0, \sigma^2/\kappa_0\right) \propto \exp\left[-\frac{\kappa_0+1}{2\sigma^2}\left(\theta-\frac{\kappa_0\theta_0+y}{\kappa_0+1}\right)^2\right]$$

Finally, we have

$$p\left(\theta \mid y, \sigma, \theta_0, \sigma^2/\kappa_0\right) \propto \exp\left[-\frac{1}{2}\frac{\left(\theta-\theta_1\right)^2}{\tau_1^2}\right]$$

where $\theta_1 = \frac{\kappa_0\theta_0+y}{\kappa_0+1} = \frac{\frac{1}{\tau_0}\theta_0+\frac{1}{\sigma^2}y}{\frac{1}{\tau_0}+\frac{1}{\sigma^2}}$ and $\tau_1^2 = \frac{\sigma^2}{\kappa_0+1} = \frac{1}{\frac{1}{\tau_0}+\frac{1}{\sigma^2}}$, or the posterior distribution of the mean given the data and priors is Gaussian or normal. Notice, the posterior mean, θ_1, weights the data and prior beliefs by their relative precisions.

For a sample of n exchangeable draws, the likelihood is

$$\ell\left(\theta \mid y, \sigma\right) \propto \prod_{i=1}^{n} \exp\left[-\frac{1}{2}\frac{\left(y_i-\theta\right)^2}{\sigma^2}\right]$$

combined with the above prior yields

$$p\left(\theta \mid y, \sigma, \theta_0, \sigma^2/\kappa_0\right) \propto \exp\left[-\frac{1}{2}\frac{\left(\theta-\theta_n\right)^2}{\tau_n^2}\right]$$

where $\theta_n = \frac{\kappa_0\theta_0 + n\overline{y}}{\kappa_0 + n} = \frac{\frac{1}{\tau_0}\theta_0 + \frac{n}{\sigma^2}\overline{y}}{\frac{1}{\tau_0} + \frac{n}{\sigma^2}}$, $\overline{y}$ is the sample mean, and $\tau_n^2 = \frac{\sigma^2}{\kappa_0 + n} = \frac{1}{\frac{1}{\tau_0} + \frac{n}{\sigma^2}}$, or the posterior distribution of the mean, θ, given the data and priors is again Gaussian or normal and the posterior mean, θ_n, weights the data and priors by their relative precisions.

6.2.1 Uninformative prior

An uninformative prior for the mean, θ, is the (improper) uniform, $p\left(\theta \mid \sigma^2\right) = 1$. Hence, the likelihood

$$\begin{aligned}
\ell\left(\theta \mid y, \sigma\right) &\propto \prod_{i=1}^{n} \exp\left[-\frac{1}{2}\frac{\left(y_i - \theta\right)^2}{\sigma^2}\right] \\
&\propto \exp\left[-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{n} y_i^2 - 2n\overline{y}\theta + n\theta^2\right\}\right] \\
&\propto \exp\left[-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2 + n\left(\theta - \overline{y}\right)^2\right\}\right] \\
&\propto \exp\left[-\frac{1}{2\sigma^2} n\left(\theta - \overline{y}\right)^2\right]
\end{aligned}$$

determines the posterior

$$p\left(\theta \mid \sigma^2, y\right) \propto \exp\left[-\frac{n}{2}\frac{\left(\theta - \overline{y}\right)^2}{\sigma^2}\right]$$

which is the kernel for a Gaussian or $\mathrm{N}\left(\theta \mid \sigma^2, y; \overline{y}, \frac{\sigma^2}{n}\right)$, the classical result.

6.3 Gaussian (known mean, unknown variance)

For a sample of n exchangeable draws with known mean, μ, and unknown variance, θ, a Gaussian or normal likelihood is

$$\ell\left(\theta \mid y, \mu\right) \propto \prod_{i=1}^{n} \theta^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\frac{\left(y_i - \mu\right)^2}{\theta}\right]$$

combines with an inverted-gamma(a, b)

$$p\left(\theta; a, b\right) \propto \theta^{-(a+1)} \exp\left[-\frac{b}{\theta}\right]$$

to yield an inverted-gamma$\left(\frac{n+2a}{2}, b + \frac{1}{2}t\right)$ posterior distribution where

$$t = \sum_{i=1}^{n} \left(y_i - \mu\right)^2$$

Alternatively and conveniently (but equivalently), we could parameterize the prior as an inverted-chi square$\left(\nu_0, \sigma_0^2\right)$[3]

$$p\left(\theta; \nu_0, \sigma_0^2\right) \propto (\theta)^{-\left(\frac{\nu_0}{2}+1\right)} \exp\left[-\frac{\nu_0 \sigma_0^2}{2\theta}\right]$$

and combine with the above likelihood to yield

$$p\left(\theta \mid y\right) \propto \theta^{-\left(\frac{n+\nu_0}{2}+1\right)} \exp\left[-\frac{1}{2\theta}\left(\nu_0 \sigma_0^2 + t\right)\right]$$

an inverted chi-square$\left(\nu_0 + n, \frac{\nu_0 \sigma_0^2 + t}{\nu_0 + n}\right)$.

6.3.1 Uninformative prior

An uninformative prior for scale is

$$p\left(\theta\right) \propto \theta^{-1}$$

Hence, the posterior distribution for scale is

$$p\left(\theta \mid y\right) \propto \theta^{-\left(\frac{n}{2}+1\right)} \exp\left[-\frac{t}{2\theta}\right]$$

which is the kernel of an inverted-chi square$\left(\theta; n, \frac{t}{n}\right)$.

6.4 Gaussian (unknown mean, unknown variance)

For a sample of n exchangeable draws, a normal likelihood with unknown mean, θ, and unknown (but constant) variance, σ^2, is

$$\ell\left(\theta, \sigma^2 \mid y\right) \propto \prod_{i=1}^{n} \sigma^{-1} \exp\left[-\frac{1}{2}\frac{\left(y_i - \theta\right)^2}{\sigma^2}\right]$$

Expanding and rewriting the likelihood gives

$$\ell\left(\theta, \sigma^2 \mid y\right) \propto \sigma^{-n} \exp\left[\sum_{i=1}^{n} -\frac{1}{2}\frac{y_i^2 - 2y_i\theta + \theta^2}{\sigma^2}\right]$$

Adding and subtracting $\sum_{i=1}^{n} 2y_i\overline{y} = 2n\overline{y}^2$, we write

$$\ell\left(\theta, \sigma^2 \mid y\right) \propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\{\left(y_i^2 - 2y_i\overline{y} + \overline{y}^2\right) + \left(\overline{y}^2 - 2\overline{y}\theta + \theta^2\right)\right\}\right]$$

---

[3] $\frac{\sigma_0^2 \nu_0}{X}$ is a scaled, inverted-chi square$\left(\nu_0, \sigma_0^2\right)$ with scale σ_0^2 where X is a chi square(ν_0) random variable.

or

$$\ell\left(\theta,\sigma^2 \mid y\right) \propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\{\left(y_i-\overline{y}\right)^2+\left(\overline{y}-\theta\right)^2\right\}\right]$$

which can be rewritten as

$$\ell\left(\theta,\sigma^2 \mid y\right) \propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\left\{\left(n-1\right)s^2+n\left(\overline{y}-\theta\right)^2\right\}\right]$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2$. The above likelihood combines with a Gaussian or normal$\left(\theta \mid \sigma^2;\theta_0,\sigma^2/\kappa_0\right)$ × inverted-chi square$\left(\sigma^2;\nu_0,\sigma_0^2\right)$ prior[4]

$$\begin{aligned} p\left(\theta \mid \sigma^2;\theta_0,\sigma^2/\kappa_0\right) \times p\left(\sigma^2;\nu_0,\sigma_0^2\right) &\propto \frac{1}{\sigma}\exp\left[-\frac{\kappa_0\left(\theta-\theta_0\right)^2}{2\sigma^2}\right] \\ &\quad \times\left(\sigma^2\right)^{-(\nu_0/2+1)}\exp\left[-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right] \\ &\propto \left(\sigma^2\right)^{-\left(\frac{\nu_0+3}{2}\right)} \\ &\quad \times\exp\left[-\frac{\nu_0\sigma_0^2+\kappa_0\left(\theta-\theta_0\right)^2}{2\sigma^2}\right] \end{aligned}$$

to yield a normal$\left(\theta \mid \sigma^2;\theta_n,\sigma_n^2/\kappa_n\right)$*inverted-chi square$\left(\sigma^2;\nu_n,\sigma_n^2\right)$ joint posterior distribution[5] where

$$\begin{aligned} \nu_n &= \nu_0+n \\ \kappa_n &= \kappa_0+n \\ \nu_n\sigma_n^2 &= \nu_0\sigma_0^2+\left(n-1\right)s^2+\frac{\kappa_0 n}{\kappa_0+n}\left(\theta_0-\overline{y}\right)^2 \end{aligned}$$

That is, the joint posterior is

$$\begin{aligned} p\left(\theta,\sigma^2 \mid y;\theta_0,\sigma^2/\kappa_0,\nu_0,\sigma_0^2\right) &\propto \left(\sigma^2\right)^{-\frac{n+\nu_0+3}{2}} \\ &\quad \times\exp\left[-\frac{1}{2\sigma^2}\left\{\begin{array}{c}\nu_0\sigma_0^2+\left(n-1\right)s^2 \\ +\kappa_0\left(\theta-\theta_0\right)^2 \\ +n\left(\theta-\overline{y}\right)^2\end{array}\right\}\right] \end{aligned}$$

6.4.1 Completing the square

The expression for the joint posterior is written by completing the square. Completing the weighted square for θ centered around

$$\theta_n = \frac{1}{\kappa_0+n}\left(\kappa_0\theta_0+n\overline{y}\right)$$

---

[4] The prior for the mean, θ, is conditional on the scale of the data, σ^2.

[5] The product of normal or Gaussian kernels produces a Gaussian kernel.

where $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ gives

$$\begin{aligned}(\kappa_0 + n)(\theta - \theta_n)^2 &= (\kappa_0 + n)\theta^2 - 2\theta(\kappa_0 + n)\theta_n + (\kappa_0 + n)\theta_n^2 \\ &= (\kappa_0 + n)\theta^2 - 2\theta(\kappa_0\theta_0 + n\overline{y}) + (\kappa_0 + n)\theta_n^2\end{aligned}$$

While expanding the exponent includes the square plus additional terms as follows

$$\begin{aligned}\kappa_0(\theta - \theta_0)^2 + n(\theta - \overline{y})^2 &= \kappa_0\left(\theta^2 - 2\theta\theta_0 + \theta_0^2\right) + n\left(\theta^2 - 2\theta\overline{y} + \overline{y}^2\right) \\ &= (\kappa_0 + n)\theta^2 - 2\theta(\kappa_0\theta_0 + n\overline{y}) + \kappa_0\theta_0^2 + n\overline{y}^2\end{aligned}$$

Add and subtract $(\kappa_0 + n)\theta_n^2$ and simplify.

$$\begin{aligned}\kappa_0(\theta - \theta_0)^2 + n(\theta - \overline{y})^2 &= (\kappa_0 + n)\theta^2 - 2\theta(\kappa_0 + n)\theta_n + (\kappa_0 + n)\theta_n^2 \\ &\quad - (\kappa_0 + n)\theta_n^2 + \kappa_0\theta_0^2 + n\overline{y}^2 \\ &= (\kappa_0 + n)(\theta - \theta_n)^2 \\ &\quad \frac{1}{(\kappa_0 + n)}\left\{\begin{array}{c}(\kappa_0 + n)\left(\kappa_0\theta_0^2 + n\overline{y}^2\right) \\ -(\kappa_0\theta_0 + n\overline{y})^2\end{array}\right\}\end{aligned}$$

Expand and simplify the last term.

$$\kappa_0(\theta - \theta_0)^2 + n(\theta - \overline{y})^2 = (\kappa_0 + n)(\theta - \theta_n)^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\theta_0 - \overline{y})^2$$

Now, the joint posterior can be rewritten as

$$\begin{aligned}p\left(\theta, \sigma^2 \mid y; \theta_0, \sigma^2/\kappa_0, \nu_0, \sigma_0^2\right) &\propto \left(\sigma^2\right)^{-\frac{n+\nu_0+3}{2}} \\ &\quad \times \exp\left[-\frac{1}{2\sigma^2}\left\{\begin{array}{c}\nu_0\sigma_0^2 + (n-1)s^2 \\ +\frac{\kappa_0 n}{\kappa_0+n}(\theta_0 - \overline{y})^2 \\ +(\kappa_0 + n)(\theta - \theta_n)^2\end{array}\right\}\right]\end{aligned}$$

or

$$\begin{aligned}p\left(\theta, \sigma^2 \mid y; \theta_0, \sigma^2/\kappa_0, \nu_0, \sigma_0^2\right) &\propto \left(\sigma^2\right)^{-\frac{n+\nu_0}{2}-1}\exp\left[-\frac{1}{2\sigma^2}\nu_n\sigma_n^2\right] \\ &\quad \times\sigma^{-1}\exp\left[-\frac{1}{2\sigma^2}(\kappa_0 + n)(\theta - \theta_n)^2\right]\end{aligned}$$

Hence, the conditional posterior distribution for the mean, θ, given σ^2 is Gaussian or normal$\left(\theta \mid \sigma^2; \theta_n, \frac{\sigma^2}{\kappa_0+n}\right)$.

6.4.2 Marginal posterior distributions

We're often interested in the marginal posterior distributions which are derived by integrating out the other parameter from the joint posterior. The

marginal posterior for the mean, θ, on integrating out σ^2 is a noncentral, scaled-Student $\mathrm{t}\left(\theta;\theta_n,\frac{\sigma_n^2}{\kappa_n},\nu_n\right)$[6] for the mean

$$p\left(\theta;\theta_n,\sigma_n^2,\kappa_n,\nu_n\right)\propto\left(\frac{\nu_n}{\nu_n+\frac{\kappa_n(\theta-\theta_n)^2}{\sigma_n^2}}\right)^{\frac{\nu_n+1}{2}}$$

or

$$p\left(\theta;\theta_n,\frac{\nu_n\sigma_n^2}{\kappa_n},\nu_n\right)\propto\left(1+\frac{\kappa_n\left(\theta-\theta_n\right)^2}{\nu_n\sigma_n^2}\right)^{-\frac{\nu_n+1}{2}}$$

and the marginal posterior for the variance, σ^2, is an inverted-chi square$\left(\sigma^2;\nu_n,\sigma_n^2\right)$ on integrating out θ.

$$p\left(\sigma^2;\nu_n,\sigma_n^2\right)\propto\left(\sigma^2\right)^{-(\nu_n/2+1)}\exp\left[-\frac{\nu_n\sigma_n^2}{2\sigma^2}\right]$$

Derivation of the marginal posterior for the mean, θ, is as follows. Let $z=\frac{A}{2\sigma^2}$ where

$$\begin{aligned}A &= \nu_0\sigma_0^2+(n-1)\,s^2+\frac{\kappa_0 n}{\kappa_0+n}\left(\theta_0-\overline{y}\right)^2+(\kappa_0+n)\left(\theta-\theta_n\right)^2\\ &= \nu_n\sigma_n^2+(\kappa_0+n)\left(\theta-\theta_n\right)^2\end{aligned}$$

The marginal posterior for the mean, θ, integrates out σ^2 from the joint posterior

$$\begin{aligned}p\left(\theta\mid y\right) &= \int_0^\infty p\left(\theta,\sigma^2\mid y\right)d\sigma^2\\ &= \int_0^\infty\left(\sigma^2\right)^{-\frac{n+\nu_0+3}{2}}\exp\left[-\frac{A}{2\sigma^2}\right]d\sigma^2\end{aligned}$$

Utilizing $\sigma^2=\frac{A}{2z}$ and $dz=-\frac{2z^2}{A}d\sigma^2$ or $d\sigma^2=-\frac{A}{2z^2}dz$,

$$\begin{aligned}p\left(\theta\mid y\right) &\propto \int_0^\infty\left(\frac{A}{2z}\right)^{-\frac{n+\nu_0+3}{2}}\frac{A}{2z^2}\exp\left[-z\right]dz\\ &\propto \int_0^\infty\left(\frac{A}{2z}\right)^{-\frac{n+\nu_0+1}{2}}z^{-1}\exp\left[-z\right]dz\\ &\propto A^{-\frac{n+\nu_0+1}{2}}\int_0^\infty z^{\frac{n+\nu_0+1}{2}-1}\exp\left[-z\right]dz\end{aligned}$$

---

[6] The noncentral, scaled-Student $\mathrm{t}\left(\theta;\theta_n,\sigma_n^2/\kappa_n,\nu_n\right)$ implies $\frac{\theta-\theta_n}{\sigma_n/\sqrt{\kappa_n}}$ has a standard Student-t(ν_n) distribution $p\left(\theta\mid y\right)\propto\left[1+\frac{\left(\frac{\theta-\theta_n}{\sigma_n/\sqrt{\kappa_n}}\right)^2}{\nu_n}\right]^{-\frac{\nu_n+1}{2}}$.

The integral $\int_0^\infty z^{\frac{n+\nu_0+1}{2}-1} \exp[-z]\, dz$ is a constant since it is the kernel of a gamma density and therefore can be ignored when deriving the kernel of the marginal posterior for the mean

$$\begin{aligned} p(\theta \mid y) &\propto A^{-\frac{n+\nu_0+1}{2}} \\ &\propto \left[\nu_n\sigma_n^2 + (\kappa_0+n)(\theta-\theta_n)^2\right]^{-\frac{n+\nu_0+1}{2}} \\ &\propto \left[1 + \frac{(\kappa_0+n)(\theta-\theta_n)^2}{\nu_n\sigma_n^2}\right]^{-\frac{n+\nu_0+1}{2}} \end{aligned}$$

which is the kernel for a noncentral, scaled Student $\mathrm{t}\left(\theta; \theta_n, \frac{\sigma_n^2}{\kappa_0+n}, n+\nu_0\right)$.

Derivation of the marginal posterior for σ^2 is somewhat simpler. Write the joint posterior in terms of the conditional posterior for the mean multiplied by the marginal posterior for σ^2.

$$p\left(\theta, \sigma^2 \mid y\right) = p\left(\theta \mid \sigma^2, y\right) p\left(\sigma^2 \mid y\right)$$

Marginalization of σ^2 is achieved by integrating out θ.

$$p\left(\sigma^2 \mid y\right) = \int_{-\infty}^{\infty} p\left(\sigma^2 \mid y\right) p\left(\theta \mid \sigma^2, y\right) d\theta$$

Since only the conditional posterior involves θ the marginal posterior for σ^2 is immediate.

$$\begin{aligned} p\left(\theta, \sigma^2 \mid y\right) &\propto \left(\sigma^2\right)^{-\frac{n+\nu_0+3}{2}} \exp\left[-\frac{A}{2\sigma^2}\right] \\ &\propto \left(\sigma^2\right)^{-\frac{n+\nu_0+2}{2}} \exp\left[-\frac{\nu_n\sigma_n^2}{2\sigma^2}\right] \sigma^{-1} \exp\left[-\frac{(\kappa_0+n)(\theta-\theta_n)^2}{2\sigma^2}\right] \end{aligned}$$

Integrating out θ yields

$$\begin{aligned} p\left(\sigma^2 \mid y\right) &\propto \left(\sigma^2\right)^{-\frac{n+\nu_0+2}{2}} \exp\left[-\frac{\nu_n\sigma_n^2}{2\sigma^2}\right] \\ &\quad \times \int_{-\infty}^{\infty} \sigma^{-1} \exp\left[-\frac{(\kappa_0+n)(\theta-\theta_n)^2}{2\sigma^2}\right] d\theta \\ &\propto \left(\sigma^2\right)^{-\left(\frac{\nu_n}{2}+1\right)} \exp\left[-\frac{\nu_n\sigma_n^2}{2\sigma^2}\right] \end{aligned}$$

which we recognize as the kernel of an inverted-chi square$\left(\sigma^2; \nu_n, \sigma_n^2\right)$.

6.4.3 Uninformative priors

The case of uninformative priors is relatively straightforward. Since priors convey no information, the prior for the mean is uniform (proportional to

a constant, $\kappa_0 \to 0$) and an uninformative prior for σ^2 has $\nu_0 \to 0$ degrees of freedom so that the joint prior is

$$p\left(\theta, \sigma^2\right) \propto \left(\sigma^2\right)^{-1}$$

The joint posterior is

$$\begin{aligned} p\left(\theta, \sigma^2 \mid y\right) &\propto \left(\sigma^2\right)^{-(n/2+1)} \exp\left[-\frac{1}{2\sigma^2}\left\{(n-1)\, s^2 + n\left(\theta - \overline{y}\right)^2\right\}\right] \\ &\propto \left(\sigma^2\right)^{-[(n-1)/2+1]} \exp\left[-\frac{\sigma_n^2}{2\sigma^2}\right] \\ &\quad \times \sigma^{-1} \exp\left[-\frac{n}{2\sigma^2}\left(\theta - \overline{y}\right)^2\right] \end{aligned}$$

where

$$\sigma_n^2 = (n-1)\, s^2$$

The conditional posterior for θ given σ^2 is Gaussian$\left(\overline{y}, \frac{\sigma^2}{n}\right)$. And, the marginal posterior for θ is noncentral, scaled Student t$\left(\overline{y}, \frac{s^2}{n}, n-1\right)$, the classical estimator.

Derivation of the marginal posterior proceeds as above. The joint posterior is

$$p\left(\theta, \sigma^2 \mid y\right) \propto \left(\sigma^2\right)^{-(n/2+1)} \exp\left[-\frac{1}{2\sigma^2}\left\{(n-1)\, s^2 + n\left(\theta - \overline{y}\right)^2\right\}\right]$$

Let $z = \frac{A}{2\sigma^2}$ where $A = (n-1)\, s^2 + n\left(\theta - \overline{y}\right)^2$. Now integrate σ^2 out of the joint posterior following the transformation of variables.

$$\begin{aligned} p\left(\theta \mid y\right) &\propto \int_0^\infty \left(\sigma^2\right)^{-(n/2+1)} \exp\left[-\frac{A}{2\sigma^2}\right] d\sigma^2 \\ &\propto A^{-n/2} \int_0^\infty z^{n/2-1} e^{-z} dz \end{aligned}$$

As before, the integral involves the kernel of a gamma density and therefore is a constant which can be ignored. Hence,

$$\begin{aligned} p\left(\theta \mid y\right) &\propto A^{-n/2} \\ &\propto \left[(n-1)\, s^2 + n\left(\theta - \overline{y}\right)^2\right]^{-\frac{n}{2}} \\ &\propto \left[1 + \frac{n\left(\theta - \overline{y}\right)^2}{(n-1)\, s^2}\right]^{-\frac{n-1+1}{2}} \end{aligned}$$

which we recognize as the kernel of a noncentral, scaled Student t$\left(\theta; \overline{y}, \frac{s^2}{n}, n-1\right)$.

6.5 Multivariate Gaussian (unknown mean, known variance)

More than one random variable (the multivariate case) with joint Gaussian or normal likelihood is analogous to the univariate case with Gaussian conjugate prior. Consider a vector of k random variables (the sample is comprised of n draws for each random variable) with unknown mean, θ, and known variance, Σ. For n exchangeable draws of the random vector (containing each of the m random variable), the multivariate Gaussian likelihood is

$$\ell\left(\theta \mid y, \Sigma\right) \propto \prod_{i=1}^{n} \exp\left[-\frac{1}{2}\left(y_i-\theta\right)^T \Sigma^{-1}\left(y_i-\theta\right)\right]$$

where superscript T refers to transpose, y_i and θ are k length vectors and Σ is a $k \times k$ variance-covariance matrix. A Gaussian prior for the mean vector, θ, with prior mean, θ_0, and prior variance, Υ_0,is

$$p\left(\theta \mid \Sigma; \theta_0, \Upsilon_0\right) \propto \exp\left[-\frac{1}{2}\left(\theta-\theta_0\right)^T \Upsilon_0^{-1}\left(\theta-\theta_0\right)\right]$$

The product of the likelihood and prior yields the kernel of a multivariate posterior Gaussian distribution for the mean

$$\begin{aligned} p\left(\theta \mid \Sigma, y; \theta_0, \Upsilon_0\right) \quad \propto \quad & \exp\left[-\frac{1}{2}\left(\theta-\theta_0\right)^T \Upsilon_0^{-1}\left(\theta-\theta_0\right)\right] \\ & \times \exp\left[\sum_{i=1}^{n} -\frac{1}{2}\left(y_i-\theta\right)^T \Sigma^{-1}\left(y_i-\theta\right)\right] \end{aligned}$$

6.5.1 Completing the square

Expanding terms in the exponent leads to

$$\begin{aligned} & \left(\theta-\theta_0\right)^T \Upsilon_0^{-1}\left(\theta-\theta_0\right) + \sum_{i=1}^{n}\left(y_i-\theta\right)^T \Sigma^{-1}\left(y_i-\theta\right) \\ = \quad & \theta^T\left(\Upsilon_0^{-1}+n\Sigma^{-1}\right)\theta - 2\theta^T\left(\Upsilon_0^{-1}\theta_0 + n\Sigma^{-1}\overline{y}\right) \\ & + \theta_0^T \Upsilon_0^{-1}\theta_0 + \sum_{i=1}^{n} y_i^T \Sigma^{-1} y_i \end{aligned}$$

where $\overline{y}$ is the sample average. While completing the (weighted) square centered around

$$\overline{\theta} = \left(\Upsilon_0^{-1}+n\Sigma^{-1}\right)^{-1}\left(\Upsilon_0^{-1}\theta_0 + n\Sigma^{-1}\overline{y}\right)$$

leads to

$$\begin{aligned}\left(\theta-\overline{\theta}\right)^{T}\left(\Upsilon_{0}^{-1}+n\Sigma^{-1}\right)\left(\theta-\overline{\theta}\right) &= \theta^{T}\left(\Upsilon_{0}^{-1}+n\Sigma^{-1}\right)\theta \\ &\quad -2\theta^{T}\left(\Upsilon_{0}^{-1}+n\Sigma^{-1}\right)\overline{\theta} \\ &\quad +\overline{\theta}^{T}\left(\Upsilon_{0}^{-1}+n\Sigma^{-1}\right)\overline{\theta}\end{aligned}$$

Thus, adding and subtracting $\overline{\theta}^{T}\left(\Upsilon_{0}^{-1}+n\Sigma^{-1}\right)\overline{\theta}$ in the exponent completes the square (with three extra terms).

$$\begin{aligned}&\left(\theta-\theta_{0}\right)^{T}\Upsilon_{0}^{-1}\left(\theta-\theta_{0}\right)+\sum_{i=1}^{n}\left(y_{i}-\theta\right)^{T}\Sigma^{-1}\left(y_{i}-\theta\right) \\ = \;& \theta^{T}\left(\Upsilon_{0}^{-1}+n\Sigma^{-1}\right)\theta-2\theta^{T}\left(\Upsilon_{0}^{-1}+n\Sigma^{-1}\right)\overline{\theta}+\overline{\theta}^{T}\left(\Upsilon_{0}^{-1}+n\Sigma^{-1}\right)\overline{\theta} \\ & -\overline{\theta}^{T}\left(\Upsilon_{0}^{-1}+n\Sigma^{-1}\right)\overline{\theta}+\theta_{0}^{T}\Upsilon_{0}^{-1}\theta_{0}+\sum_{i=1}^{n}y_{i}^{T}\Sigma^{-1}y_{i} \\ = \;& \left(\theta-\overline{\theta}\right)^{T}\left(\Upsilon_{0}^{-1}+n\Sigma^{-1}\right)\left(\theta-\overline{\theta}\right) \\ & -\overline{\theta}^{T}\left(\Upsilon_{0}^{-1}+n\Sigma^{-1}\right)\overline{\theta}+\theta_{0}^{T}\Upsilon_{0}^{-1}\theta_{0}+\sum_{i=1}^{n}y_{i}^{T}\Sigma^{-1}y_{i}\end{aligned}$$

Dropping constants (the last three extra terms unrelated to θ) gives

$$p\left(\theta\mid\Sigma,y;\theta_{0},\Upsilon_{0}\right)\propto\exp\left[-\frac{1}{2}\left(\theta-\overline{\theta}\right)^{T}\left(\Upsilon_{0}^{-1}+n\Sigma^{-1}\right)\left(\theta-\overline{\theta}\right)\right]$$

Hence, the posterior for the mean θ has expected value $\overline{\theta}$ and variance

$$Var\left[\theta\mid y,\Sigma,\theta_{0},\Upsilon_{0}\right]=\left(\Upsilon_{0}^{-1}+n\Sigma^{-1}\right)^{-1}$$

As in the univariate case, the data and prior beliefs are weighted by their relative precisions.

6.5.2 Uninformative priors

Uninformative priors for θ are proportional to a constant. Hence, the likelihood determines the posterior

$$\ell\left(\theta\mid\Sigma,y\right)\propto\exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(y_{i}-\theta\right)^{T}\Sigma^{-1}\left(y_{i}-\theta\right)\right]$$

Expanding the exponent and adding and subtracting $n\overline{y}^T\Sigma^{-1}\overline{y}$ (to complete the square) gives

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\theta)^T\Sigma^{-1}(y_i-\theta) &= \sum_{i=1}^{n}y_i^T\Sigma^{-1}y_i - 2n\theta^T\Sigma^{-1}\overline{y}+n\theta^T\Sigma^{-1}\theta \\ &\quad +n\overline{y}^T\Sigma^{-1}\overline{y}-n\overline{y}^T\Sigma^{-1}\overline{y} \\ &= n(\overline{y}-\theta)^T\Sigma^{-1}(\overline{y}-\theta) \\ &\quad +\sum_{i=1}^{n}y_i^T\Sigma^{-1}y_i - n\overline{y}^T\Sigma^{-1}\overline{y}\end{aligned}$$

The latter two terms are constants, hence, the posterior kernel is

$$p(\theta \mid \Sigma, y) \propto \exp\left[-\frac{n}{2}(\overline{y}-\theta)^T\Sigma^{-1}(\overline{y}-\theta)\right]$$

which is Gaussian or $\mathrm{N}\left(\theta;\overline{y},\frac{1}{n}\Sigma\right)$, the classical result.

6.6 Multivariate Gaussian (unknown mean, unknown variance)

When both the mean, θ, and variance, Σ, are unknown, the multivariate Gaussian cases remains analogous to the univariate case. Specifically, a Gaussian likelihood

$$\begin{aligned}\ell(\theta,\Sigma \mid y) &\propto \prod_{i=1}^{n}|\Sigma|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(y_i-\theta)^T\Sigma^{-1}(y_i-\theta)\right] \\ &\propto |\Sigma|^{-\frac{n}{2}}\exp\left[-\frac{1}{2}\left\{\begin{array}{c}\sum_{i=1}^{n}(y_i-\overline{y})^T\Sigma^{-1}(y_i-\overline{y}) \\ +n(\overline{y}-\theta)^T\Sigma^{-1}(\overline{y}-\theta)\end{array}\right\}\right] \\ &\propto |\Sigma|^{-\frac{n}{2}}\exp\left[-\frac{1}{2}\left\{(n-1)s^2+n(\overline{y}-\theta)^T\Sigma^{-1}(\overline{y}-\theta)\right\}\right]\end{aligned}$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^T\Sigma^{-1}(y_i-\overline{y})$ combines with a Gaussian-inverted Wishart prior

$$\begin{aligned}p\left(\theta \mid \Sigma;\theta_0,\frac{\Sigma}{\kappa_0}\right)\times p\left(\Sigma^{-1};\nu,\Psi\right) &\propto |\Sigma|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\theta-\theta_0)^T\kappa_0\Sigma^{-1}(\theta-\theta_0)\right] \\ &\quad \times|\Psi|^{\frac{\nu}{2}}|\Sigma|^{-\frac{\nu+k+1}{2}}\exp\left[-\frac{tr\left(\Psi\Sigma^{-1}\right)}{2}\right]\end{aligned}$$

where $tr(\cdot)$ is the trace of the matrix and ν is degrees of freedom, to produce

$$\begin{aligned} p(\theta, \Sigma \mid y) &\propto |\Psi|^{\frac{\nu}{2}} |\Sigma|^{-\frac{\nu+n+k+1}{2}} \exp\left[-\frac{tr\left(\Psi\Sigma^{-1}\right)}{2}\right] \\ &\quad \times |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left\{\begin{array}{c}(n-1)s^2 + n(\overline{y}-\theta)^T\Sigma^{-1}(\overline{y}-\theta) \\ +\kappa_0(\theta-\theta_0)^T\Sigma^{-1}(\theta-\theta_0)\end{array}\right\}\right] \end{aligned}$$

6.6.1 Completing the square

Completing the square involves the matrix analog to the univariate unknown mean and variance case. Consider the exponent (in braces)

$$\begin{aligned} & (n-1)s^2 + n(\overline{y}-\theta)^T\Sigma^{-1}(\overline{y}-\theta) + \kappa_0(\theta-\theta_0)^T\Sigma^{-1}(\theta-\theta_0) \\ = \;& (n-1)s^2 + n\overline{y}^T\Sigma^{-1}\overline{y} - 2n\theta^T\Sigma^{-1}\overline{y} + n\theta^T\Sigma^{-1}\theta \\ & +\kappa_0\theta^T\Sigma^{-1}\theta - 2\kappa_0\theta^T\Sigma^{-1}\theta_0 + \kappa_0\theta_0^T\Sigma^{-1}\theta_0 \\ = \;& (n-1)s^2 + (\kappa_0+n)\theta^T\Sigma^{-1}\theta - 2\theta^T\Sigma^{-1}(\kappa_0\theta_0 + n\overline{y}) + (\kappa_0+n)\theta_n^T\Sigma^{-1}\theta_n \\ & -(\kappa_0+n)\theta_n^T\Sigma^{-1}\theta_n + \kappa_0\theta_0^T\Sigma^{-1}\theta_0 + n\overline{y}^T\Sigma^{-1}\overline{y} \\ = \;& (n-1)s^2 + (\kappa_0+n)(\theta-\theta_n)^T\Sigma^{-1}(\theta-\theta_n) \\ & +\frac{\kappa_0 n}{\kappa_0+n}(\theta_0-\overline{y})^T\Sigma^{-1}(\theta_0-\overline{y}) \end{aligned}$$

Hence, the joint posterior can be rewritten as

$$\begin{aligned} p(\theta, \Sigma \mid y) &\propto |\Psi|^{\frac{\nu}{2}} |\Sigma|^{-\frac{\nu+n+k+1}{2}} \exp\left[-\frac{tr\left(\Psi\Sigma^{-1}\right)}{2}\right] \\ &\quad \times |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left\{\begin{array}{c}(\kappa_0+n)(\theta-\theta_n)^T\Sigma^{-1}(\theta-\theta_n) \\ +(n-1)s^2 \\ +\frac{\kappa_0 n}{\kappa_0+n}(\theta_0-\overline{y})^T\Sigma^{-1}(\theta_0-\overline{y})\end{array}\right\}\right] \\ &\propto |\Psi|^{\frac{\nu}{2}} |\Sigma|^{-\frac{\nu+n+k+1}{2}} \exp\left[-\frac{1}{2}\left\{\begin{array}{c}tr\left(\Psi\Sigma^{-1}\right) + (n-1)s^2 \\ +\frac{\kappa_0 n}{\kappa_0+n}(\theta_0-\overline{y})^T\Sigma^{-1}(\theta_0-\overline{y})\end{array}\right\}\right] \\ &\quad \times |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left\{(\kappa_0+n)(\theta-\theta_n)^T\Sigma^{-1}(\theta-\theta_n)\right\}\right] \end{aligned}$$

6.6.2 Inverted-Wishart kernel

We wish to identify the exponent with Gaussian by inverted-Wishart kernels where the inverted-Wishart involves the trace of a square, symmetric matrix, call it Ψ_n, multiplied by Σ^{-1}.

To make this connection we utilize the following general results. Since a quadratic form, say $x^T\Sigma^{-1}x$, is a scalar, it's equal to its trace,

$$x^T\Sigma^{-1}x = tr\left(x^T\Sigma^{-1}x\right)$$

Further, for conformable matrices A, B and C, D,

$$tr\left(A\right) + tr\left(B\right) = tr\left(A + B\right)$$

and

$$tr\left(CD\right) = tr\left(DC\right)$$

We immediately have the results

$$tr\left(x^T x\right) = tr\left(xx^T\right)$$

and

$$tr\left(x^T \Sigma^{-1} x\right) = tr\left(\Sigma^{-1} xx^T\right) = tr\left(xx^T \Sigma^{-1}\right)$$

Therefore, the above joint posterior can be rewritten as a $\mathrm{N}\Big(\theta; \theta_n, \left(\kappa_0 + n\right)^{-1} \Sigma\Big) \times$ inverted-Wishart$\left(\Sigma^{-1}; \nu + n, \Psi_n\right)$

$$\begin{aligned} p\left(\theta, \Sigma \mid y\right) \quad \propto \quad & \left|\Psi_n\right|^{\frac{\nu+n}{2}} \left|\Sigma\right|^{-\frac{\nu+n+k+1}{2}} \exp\left[-\frac{1}{2} tr\left(\Psi_n \Sigma^{-1}\right)\right] \\ & \times \left|\Sigma\right|^{-\frac{1}{2}} \exp\left[-\frac{\kappa_0 + n}{2} \left(\theta - \theta_n\right)^T \Sigma^{-1} \left(\theta - \theta_n\right)\right] \end{aligned}$$

where

$$\theta_n = \frac{1}{\kappa_0 + n} \left(\kappa_0 \theta_0 + n\overline{y}\right)$$

and

$$\Psi_n = \Psi + \sum_{i=1}^{n} \left(y_i - \overline{y}\right) \left(y_i - \overline{y}\right)^T + \frac{\kappa_0 n}{\kappa_0 + n} \left(\overline{y} - \theta_0\right) \left(\overline{y} - \theta_0\right)^T$$

Now, it's apparent the conditional posterior for θ given Σ is $\mathrm{N}\Big(\theta_n, \left(\kappa_0 + n\right)^{-1} \Sigma\Big)$

$$p\left(\theta \mid \Sigma, y\right) \propto \exp\left[-\frac{\kappa_0 + n}{2} \left(\theta - \theta_n\right)^T \Sigma^{-1} \left(\theta - \theta_n\right)\right]$$

6.6.3 Marginal posterior distributions

Integrating out the other parameter gives the marginal posteriors, a multivariate Student t for the mean,

$$\text{Student t}_k\left(\theta; \theta_n, \Gamma, \nu + n - k + 1\right)$$

and an inverted-Wishart for the variance,

$$\text{I-W}\left(\Sigma^{-1}; \nu + n, \Psi_n\right)$$

where

$$\Gamma = \left(\kappa_0 + n\right)^{-1} \left(\nu + n - k + 1\right)^{-1} \Psi_n$$

Marginalization of the mean derives from the following identities (see Box and Tiao [1973], p. 427, 441). Let Z be a $m \times m$ positive definite symmetric matrix consisting of $\frac{1}{2}m(m+1)$ distinct random variables z_{ij} $(i,j=1,\ldots,m; i \geq j)$. And let $q > 0$ and B be an $m \times m$ positive definite symmetric matrix. Then, the distribution of z_{ij},

$$p(Z) \propto |Z|^{\frac{1}{2}q-1} \exp\left[-\tfrac{1}{2}tr(ZB)\right], \quad Z > 0$$

is a multivariate generalization of the χ^2 distribution obtained by Wishart [1928]. Integrating out the distinct z_{ij} produces the first identity.

$$\begin{aligned} \int_{Z>0} |Z|^{\frac{1}{2}q-1} \exp\left[-\frac{1}{2}tr(ZB)\right] dZ &= |B|^{-\frac{1}{2}(q+m-1)} \\ &\quad \times 2^{\frac{1}{2}(q+m-1)} \Gamma_m\left(\frac{q+m-1}{2}\right) \end{aligned} \tag{I.1}$$

where $\Gamma_p(b)$ is the generalized gamma function (Siegel [1935])

$$\Gamma_p(b) = \left[\Gamma\left(\tfrac{1}{2}\right)\right]^{\frac{1}{2}p(p-1)} \prod_{\alpha=1}^{p} \Gamma\left(b + \tfrac{\alpha-p}{2}\right), \quad b > \tfrac{p-1}{2}$$

and

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$$

or for integer n,

$$\Gamma(n) = (n-1)!$$

The second identity involves the relationship between determinants that allows us to express the above as a quadratic form. The identity is

$$|I_k - PQ| = |I_l - QP| \tag{I.2}$$

for P a $k \times l$ matrix and Q a $l \times k$ matrix.

If we transform the joint posterior to $p\left(\theta, \Sigma^{-1} \mid y\right)$, the above identities can be applied to marginalize the joint posterior. The key to transformation is

$$p\left(\theta, \Sigma^{-1} \mid y\right) = p(\theta, \Sigma \mid y) \left|\frac{\partial \Sigma}{\partial \Sigma^{-1}}\right|$$

where $\left|\frac{\partial \Sigma}{\partial \Sigma^{-1}}\right|$is the (absolute value of the) determinant of the Jacobian or

$$\begin{aligned} \left|\frac{\partial \Sigma}{\partial \Sigma^{-1}}\right| &= \left|\frac{\partial(\sigma_{11}, \sigma_{12}, \ldots, \sigma_{kk})}{\partial(\sigma^{11}, \sigma^{12}, \ldots, \sigma^{kk})}\right| \\ &= |\Sigma|^{k+1} \end{aligned}$$

with σ_{ij} the elements of Σ and σ^{ij} the elements of Σ^{-1}. Hence,

$$\begin{aligned} p(\theta, \Sigma \mid y) &\propto |\Sigma|^{-\frac{\nu+n+k+1}{2}} \exp\left[-\frac{1}{2} tr\left(\Psi_n \Sigma^{-1}\right)\right] \\ &\quad \times |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{\kappa_0+n}{2}(\theta-\theta_n)^T \Sigma^{-1}(\theta-\theta_n)\right] \\ &\propto |\Sigma|^{-\frac{\nu+n+k}{2}-1} \exp\left[-\frac{1}{2} tr\left(S(\theta)\Sigma^{-1}\right)\right] \end{aligned}$$

where $S(\theta) = \Psi_n + (\kappa_0+n)(\theta-\theta_n)(\theta-\theta_n)^T$, can be rewritten

$$\begin{aligned} p\left(\theta, \Sigma^{-1} \mid y\right) &\propto |\Sigma|^{-\frac{\nu+n+k+2}{2}} \exp\left[-\frac{1}{2} tr\left(S(\theta)\Sigma^{-1}\right)\right] |\Sigma|^{\frac{2k+2}{2}} \\ &\propto \left|\Sigma^{-1}\right|^{\frac{\nu+n-k}{2}} \exp\left[-\frac{1}{2} tr\left(S(\theta)\Sigma^{-1}\right)\right] \end{aligned}$$

Now, applying the first identity yields

$$\begin{aligned} \int_{\Sigma^{-1}>0} p\left(\theta, \Sigma^{-1} \mid y\right) d\Sigma^{-1} &\propto |S(\theta)|^{-\frac{1}{2}(\nu+n+1)} \\ &\propto \left|\Psi_n + (\kappa_0+n)(\theta-\theta_n)(\theta-\theta_n)^T\right|^{-\frac{1}{2}(\nu+n+1)} \\ &\propto \left|I + (\kappa_0+n)\Psi_n^{-1}(\theta-\theta_n)(\theta-\theta_n)^T\right|^{-\frac{1}{2}(\nu+n+1)} \end{aligned}$$

And the second identity gives

$$p(\theta \mid y) \propto \left[1 + (\kappa_0+n)(\theta-\theta_n)^T \Psi_n^{-1}(\theta-\theta_n)\right]^{-\frac{1}{2}(\nu+n+1)}$$

We recognize this is the kernel of a multivariate Student $\mathrm{t}_k(\theta; \theta_n, \Gamma, \nu+n-k+1)$ distribution.

6.6.4 Uninformative priors

The joint uninformative prior (with a locally uniform prior for θ) is

$$p(\theta, \Sigma) \propto |\Sigma|^{-\frac{k+1}{2}}$$

and the joint posterior is

$$\begin{aligned} p(\theta, \Sigma \mid y) &\propto |\Sigma|^{-\frac{k+1}{2}} |\Sigma|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\left\{(n-1)s^2 + n(\overline{y}-\theta)^T \Sigma^{-1}(\overline{y}-\theta)\right\}\right] \\ &\propto |\Sigma|^{-\frac{n+k+1}{2}} \exp\left[-\frac{1}{2}\left\{(n-1)s^2 + n(\overline{y}-\theta)^T \Sigma^{-1}(\overline{y}-\theta)\right\}\right] \\ &\propto |\Sigma|^{-\frac{n+k+1}{2}} \exp\left[-\frac{1}{2} tr\left(S(\theta)\Sigma^{-1}\right)\right] \end{aligned}$$

where now $S(\theta) = \sum_{i=1}^{n} (\overline{y} - y_i)(\overline{y} - y_i)^T + n(\overline{y} - \theta)(\overline{y} - \theta)^T$. Then, the conditional posterior for θ given Σ is $\mathrm{N}\left(\overline{y}, n^{-1}\Sigma\right)$

$$p(\theta \mid \Sigma, y) \propto \exp\left[-\frac{n}{2}(\theta - \overline{y})^T \Sigma^{-1}(\theta - \overline{y})\right]$$

The marginal posterior for θ is derived analogous to the above informed conjugate prior case. Rewriting the posterior in terms of Σ^{-1} yields

$$\begin{aligned} p\left(\theta, \Sigma^{-1} \mid y\right) &\propto |\Sigma|^{-\frac{n+k+1}{2}} \exp\left[-\frac{1}{2} tr\left(S(\theta)\Sigma^{-1}\right)\right] |\Sigma|^{\frac{2k+2}{2}} \\ &\propto \left|\Sigma^{-1}\right|^{\frac{n-k-1}{2}} \exp\left[-\frac{1}{2} tr\left(S(\theta)\Sigma^{-1}\right)\right] \end{aligned}$$

$$\begin{aligned} p(\theta \mid y) &\propto \int_{\Sigma^{-1}>0} p\left(\theta, \Sigma^{-1} \mid y\right) d\Sigma^{-1} \\ &\propto \int_{\Sigma^{-1}>0} \left|\Sigma^{-1}\right|^{\frac{n-k-1}{2}} \exp\left[-\frac{1}{2} tr\left(S(\theta)\Sigma^{-1}\right)\right] d\Sigma^{-1} \end{aligned}$$

The first identity (I.1) produces

$$\begin{aligned} p(\theta \mid y) &\propto |S(\theta)|^{-\frac{n}{2}} \\ &\propto \left|\sum_{i=1}^{n} (\overline{y} - y_i)(\overline{y} - y_i)^T + n(\overline{y} - \theta)(\overline{y} - \theta)^T\right|^{\frac{n}{2}} \\ &\propto \left|I + n\left[\sum_{i=1}^{n} (\overline{y} - y_i)(\overline{y} - y_i)^T\right]^{-1} (\overline{y} - \theta)(\overline{y} - \theta)^T\right|^{\frac{n}{2}} \end{aligned}$$

The second identity (I.2) identifies the marginal posterior for θ as (multivariate) Student $\mathrm{t}_k\left(\theta; \overline{y}, \frac{1}{n}s^2, n-k\right)$

$$p(\theta \mid y) \propto \left[1 + \frac{n}{(n-k)s^2}(\overline{y} - \theta)^T (\overline{y} - \theta)^T\right]^{-\frac{n}{2}}$$

where $(n-k)s^2 = \sum_{i=1}^{n} (\overline{y} - y_i)^T (\overline{y} - y_i)$. The marginal posterior for the variance is $\text{I-W}\left(\Sigma^{-1}; n, \Psi_n\right)$ where now $\Psi_n = \sum_{i=1}^{n} (\overline{y} - y_i)(\overline{y} - y_i)^T$.

6.7 Bayesian linear regression

Linear regression is the starting point for more general data modeling strategies, including nonlinear models. Hence, Bayesian linear regression is foundational. Suppose the data are generated by

$$y = X\beta + \varepsilon$$

where X is a $n \times p$ full column rank matrix of (weakly exogenous) regressors and $\varepsilon \sim N\left(0, \sigma^2 I_n\right)$ and $E\left[\varepsilon \mid X\right] = 0$. Then, the sample conditional density is $\left(y \mid X, \beta, \sigma^2\right) \sim N\left(X\beta, \sigma^2 I_n\right)$.

6.7.1 Known variance

If the error variance, $\sigma^2 I_n$, is known and we have informed Gaussian priors for β conditional on σ^2,

$$p\left(\beta \mid \sigma^2\right) \sim N\left(\beta_0, \sigma^2 V_0\right)$$

where we can think of $V_0 = \left(X_0^T X_0\right)^{-1}$ as if we had a prior sample (y_0, X_0) such that

$$\beta_0 = \left(X_0^T X_0\right)^{-1} X_0^T y_0$$

then the conditional posterior for β is

$$p\left(\beta \mid \sigma^2, y, X; \beta_0, V_0\right) \sim N\left(\overline{\beta}, V_\beta\right)$$

where

$$\overline{\beta} = \left(X_0^T X_0 + X^T X\right)^{-1} \left(X_0^T X_0 \beta_0 + X^T X \widehat{\beta}\right)$$

$$\widehat{\beta} = \left(X^T X\right)^{-1} X^T y$$

and

$$V_\beta = \sigma^2 \left(X_0^T X_0 + X^T X\right)^{-1}$$

The variance expression follows from rewriting the estimator

$$\begin{aligned}
\overline{\beta} &= \left(X_0^T X_0 + X^T X\right)^{-1} \left(X_0^T X_0 \beta_0 + X^T X \widehat{\beta}\right) \\
&= \left(X_0^T X_0 + X^T X\right)^{-1} \left(X_0^T X_0 \left(X_0^T X_0\right)^{-1} X_0^T y_0 + X^T X \left(X^T X\right)^{-1} X^T y\right) \\
&= \left(X_0^T X_0 + X^T X\right)^{-1} \left(X_0^T y_0 + X^T y\right)
\end{aligned}$$

Since the DGP is

$$\begin{array}{cc}
y_0 = X_0 \beta + \varepsilon_0, & \varepsilon_0 \sim N\left(0, \sigma^2 I_{n_0}\right) \\
y = X\beta + \varepsilon, & \varepsilon \sim N\left(0, \sigma^2 I_n\right)
\end{array}$$

then

$$\overline{\beta} = \left(X_0^T X_0 + X^T X\right)^{-1} \left(X_0^T X_0 \beta + X_0^T \varepsilon_0 + X^T X \beta + X^T \varepsilon\right)$$

The conditional (and by iterated expectations, unconditional) expected value of the estimator is

$$E\left[\overline{\beta} \mid X, X_0\right] = \left(X_0^T X_0 + X^T X\right)^{-1} \left(X_0^T X_0 + X^T X\right) \beta = \beta$$

Hence,

$$\begin{aligned}\overline{\beta} - E\left[\overline{\beta} \mid X, X_0\right] &= \overline{\beta} - \beta \\ &= \left(X_0^T X_0 + X^T X\right)^{-1}\left(X_0^T \varepsilon_0 + X^T \varepsilon\right)\end{aligned}$$

so that

$$\begin{aligned}V_\beta &\equiv Var\left[\overline{\beta} \mid X, X_0\right] \\ &= E\left[\left(\overline{\beta} - \beta\right)\left(\overline{\beta} - \beta\right)^T \mid X, X_0\right] \\ &= E\left[\begin{array}{c}\left(X_0^T X_0 + X^T X\right)^{-1}\left(X_0^T \varepsilon_0 + X^T \varepsilon\right)\left(X_0^T \varepsilon_0 + X^T \varepsilon\right)^T \\ \times \left(X_0^T X_0 + X^T X\right)^{-1} \mid X, X_0\end{array}\right] \\ &= E\left[\begin{array}{c}\left(X_0^T X_0 + X^T X\right)^{-1}\left(\begin{array}{c}X_0^T \varepsilon_0 \varepsilon_0^T X_0 + X^T \varepsilon \varepsilon_0^T X_0 \\ + X_0^T \varepsilon_0 \varepsilon^T X + X^T \varepsilon^T \varepsilon^T X\end{array}\right) \\ \times \left(X_0^T X_0 + X^T X\right)^{-1} \mid X, X_0\end{array}\right] \\ &= \left(X_0^T X_0 + X^T X\right)^{-1}\left(X_0^T \sigma^2 I X_0 + X^T \sigma^T I X\right)\left(X_0^T X_0 + X^T X\right)^{-1} \\ &= \sigma^2\left(X_0^T X_0 + X^T X\right)^{-1}\left(X_0^T X_0 + X^T X\right)\left(X_0^T X_0 + X^T X\right)^{-1} \\ &= \sigma^2\left(X_0^T X_0 + X^T X\right)^{-1}\end{aligned}$$

Now, let's backtrack and derive the conditional posterior as the product of conditional priors and the likelihood function. The likelihood function for known variance is

$$\ell\left(\beta \mid \sigma^2, y, X\right) \propto \exp\left[-\frac{1}{2\sigma^2}\left(y - X\beta\right)^T\left(y - X\beta\right)\right]$$

Conditional Gaussian priors are

$$p\left(\beta \mid \sigma^2\right) \propto \exp\left[-\frac{1}{2\sigma^2}\left(\beta - \beta_0\right)^T V_0^{-1}\left(\beta - \beta_0\right)\right]$$

The conditional posterior is the product of the prior and likelihood

$$\begin{aligned}p\left(\beta \mid \sigma^2, y, X\right) &\propto \exp\left[-\frac{1}{2\sigma^2}\left\{\begin{array}{c}\left(y - X\beta\right)^T\left(y - X\beta\right) \\ + \left(\beta - \beta_0\right)^T V_0^{-1}\left(\beta - \beta_0\right)\end{array}\right\}\right] \\ &= \exp\left[-\frac{1}{2\sigma^2}\left\{\begin{array}{c}y^T y - 2y^T X\beta + \beta^T X^T X\beta \\ + \beta^T X_0^T X_0 \beta - 2\beta_0^T X_0^T X_0 \beta \\ + \beta_0^T X_0^T X_0 \beta_0\end{array}\right\}\right]\end{aligned}$$

The first and last terms in the exponent do not involve β (are constants) and can ignored as they are absorbed through normalization. This leaves

$$\begin{aligned}p\left(\beta \mid \sigma^2, y, X\right) &\propto \exp\left[-\frac{1}{2\sigma^2}\left\{\begin{array}{c}-2y^T X\beta + \beta^T X^T X\beta + \beta^T X_0^T X_0 \beta \\ -2\beta_0^T X_0^T X_0 \beta\end{array}\right\}\right] \\ &= \exp\left[-\frac{1}{2\sigma^2}\left\{\begin{array}{c}\beta^T\left(X_0^T X_0 + X^T X\right)\beta \\ -2\left(y^T X + \beta_0^T X_0^T X_0\right)\beta\end{array}\right\}\right]\end{aligned}$$

which can be recognized as the expansion of the conditional posterior claimed above.

$$\begin{aligned}
p\left(\beta \mid \sigma^2, y, X\right) &\sim N\left(\overline{\beta}, V_\beta\right) \\
&\propto \exp\left[-\frac{1}{2}\left(\beta-\overline{\beta}\right)^T V_\beta^{-1}\left(\beta-\overline{\beta}\right)\right] \\
&= \exp\left[-\frac{1}{2\sigma^2}\left(\beta-\overline{\beta}\right)^T\left(X_0^T X_0 + X^T X\right)\left(\beta-\overline{\beta}\right)\right] \\
&= \exp\left[-\frac{1}{2\sigma^2}\left\{\begin{array}{c}\beta^T\left(X_0^T X_0 + X^T X\right)\beta \\ -2\overline{\beta}^T\left(X_0^T X_0 + X^T X\right)\beta \\ +\overline{\beta}^T\left(X_0^T X_0 + X^T X\right)\overline{\beta}\end{array}\right\}\right] \\
&= \exp\left[-\frac{1}{2\sigma^2}\left\{\begin{array}{c}\beta^T\left(X_0^T X_0 + X^T X\right)\beta \\ -2\left(X_0^T X_0\beta_0 + X^T y\right)^T\beta \\ +\overline{\beta}^T\left(X_0^T X_0 + X^T X\right)\overline{\beta}\end{array}\right\}\right]
\end{aligned}$$

The last term in the exponent is all constants (does not involve β) so its absorbed through normalization and disregarded for comparison of kernels. Hence,

$$\begin{aligned}
p\left(\beta \mid \sigma^2, y, X\right) &\propto \exp\left[-\frac{1}{2}\left(\beta-\overline{\beta}\right)^T V_\beta^{-1}\left(\beta-\overline{\beta}\right)\right] \\
&\propto \exp\left[-\frac{1}{2\sigma^2}\left\{\begin{array}{c}\beta^T\left(X_0^T X_0 + X^T X\right)\beta \\ -2\left(y^T X + \beta_0^T X_0^T X_0\right)\beta\end{array}\right\}\right]
\end{aligned}$$

as claimed.

Uninformative priors

If the prior for β is uniformly distributed conditional on known variance, σ^2, $p\left(\beta \mid \sigma^2\right) \propto 1$, then it's as if $X_0^T X_0 \to 0$ (the information matrix for the prior is null) and the posterior for β is

$$p\left(\beta \mid \sigma^2, y, X\right) \sim N\left(\widehat{\beta}, \sigma^2\left(X^T X\right)^{-1}\right)$$

equivalent to the classical parameter estimators.

To see this intuition holds, recognize combining the likelihood with the uninformative prior indicates the posterior is proportional to the likelihood.

$$p\left(\beta \mid \sigma^2, y, X\right) \propto \exp\left[-\frac{1}{2\sigma^2}\left(y-X\beta\right)^T\left(y-X\beta\right)\right]$$

Expanding this expression yields

$$p\left(\beta \mid \sigma^2, y, X\right) \propto \exp\left[-\frac{1}{2\sigma^2}\left(y^T y - 2y^T X\beta + \beta^T X^T X\beta\right)\right]$$

The first term in the exponent doesn't depend on β and can be dropped as it's absorbed via normalization. This leaves

$$p\left(\beta \mid \sigma^2, y, X\right) \propto \exp\left[-\frac{1}{2\sigma^2}\left(-2y^T X\beta + \beta^T X^T X\beta\right)\right]$$

Now, write $p\left(\beta \mid \sigma^2, y, X\right) \sim N\left(\widehat{\beta}, \sigma^2\left(X^T X\right)^{-1}\right)$

$$p\left(\beta \mid \sigma^2, y, X\right) \propto \exp\left[-\frac{1}{2\sigma^2}\left(\beta - \widehat{\beta}\right)^T X^T X\left(\beta - \widehat{\beta}\right)\right]$$

and expand

$$p\left(\beta \mid \sigma^2, y, X\right) \propto \exp\left[-\frac{1}{2\sigma^2}\left(\beta^T X^T X\beta - 2\beta^T X^T X\widehat{\beta} + \widehat{\beta} X^T X\widehat{\beta}\right)\right]$$

The last term in the exponent doesn't depend on β and is absorbed via normalization. This leaves

$$\begin{aligned} p\left(\beta \mid \sigma^2, y, X\right) &\propto \exp\left[-\frac{1}{2\sigma^2}\left(\beta^T X^T X\beta - 2\beta^T X^T X\widehat{\beta}\right)\right] \\ &\propto \exp\left[-\frac{1}{2\sigma^2}\left(\beta^T X^T X\beta - 2\beta^T X^T X\left(X^T X\right)^{-1} X^T y\right)\right] \\ &\propto \exp\left[-\frac{1}{2\sigma^2}\left(\beta^T X^T X\beta - 2\beta^T X^T y\right)\right] \end{aligned}$$

As this latter expression matches the simplified likelihood expression, the demonstration is complete, $p\left(\beta \mid \sigma^2, y, X\right) \sim N\left(\widehat{\beta}, \sigma^2\left(X^T X\right)^{-1}\right)$.

6.7.2 Unknown variance

In the usual case where the variance as well as the regression coefficients, β, are unknown, the likelihood function can be expressed as

$$\ell\left(\beta, \sigma^2 \mid y, X\right) \propto \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\left(y - X\beta\right)^T\left(y - X\beta\right)\right]$$

Rewriting gives

$$\ell\left(\beta, \sigma^2 \mid y, X\right) \propto \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\varepsilon^T\varepsilon\right]$$

since $\varepsilon = y - X\beta$. The estimated model is $y = Xb + e$, therefore $X\beta + \varepsilon = Xb + e$ where $b = \left(X^T X\right)^{-1} X^T y$ and $e = y - Xb$ are estimates of β and ε, respectively. This implies $\varepsilon = e - X\left(\beta - b\right)$ and

$$\ell\left(\beta, \sigma^2 \mid y, X\right) \propto \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\left\{\begin{array}{c} e^T e - 2\left(\beta - b\right)^T X^T e \\ + \left(\beta - b\right)^T X^T X\left(\beta - b\right) \end{array}\right\}\right]$$

Since, $X^T e = 0$ by construction, this simplifies as

$$\ell\left(\beta, \sigma^2 \mid y, X\right) \propto \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\left\{e^T e + (\beta - b)^T X^T X (\beta - b)\right\}\right]$$

or

$$\ell\left(\beta, \sigma^2 \mid y, X\right) \propto \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\left\{(n-p) s^2 + (\beta - b)^T X^T X (\beta - b)\right\}\right]$$

where $s^2 = \frac{1}{n-p} e^T e$.[7]

The conjugate prior for linear regression is the Gaussian$\left(\beta \mid \sigma^2; \beta_0, \sigma^2 \Omega_0^{-1}\right)$-inverse chi square$\left(\sigma^2; \nu_0, \sigma_0^2\right)$

$$\begin{aligned} p\left(\beta \mid \sigma^2; \beta_0, \sigma^2 \Omega_0^{-1}\right) \times p\left(\sigma^2; \nu_0, \sigma_0^2\right) \quad &\propto \quad \sigma^{-p} \exp\left[-\frac{(\beta - \beta_0)^T \Omega_0 (\beta - \beta_0)}{2\sigma^2}\right] \\ &\quad \times \sigma^{-(\nu_0/2+1)} \exp\left[-\frac{\nu_0 \sigma_0^2}{2\sigma^2}\right] \end{aligned}$$

Combining the prior with the likelihood gives a joint Gaussian$\left(\overline{\beta}, \sigma^2 \Omega_n^{-1}\right)$-inverse chi square$\left(\nu_0 + n, \sigma_n^2\right)$ posterior

$$\begin{aligned} p\left(\beta, \sigma^2 \mid y, X; \beta_0, \sigma^2 \Omega_0^{-1}, \nu_0, \sigma_0^2\right) \quad &\propto \quad \sigma^{-n} \exp\left[-\frac{(n-p) s^2}{2\sigma^2}\right] \\ &\quad \times \exp\left[-\frac{(\beta - b)^T X^T X (\beta - b)}{2\sigma^2}\right] \\ &\quad \times \sigma^{-p} \exp\left[-\frac{(\beta - \beta_0)^T \Omega_0 (\beta - \beta_0)}{2\sigma^2}\right] \\ &\quad \times \left(\sigma^2\right)^{-(\nu_0/2+1)} \exp\left[-\frac{\nu_0 \sigma_0^2}{2\sigma^2}\right] \end{aligned}$$

---

[7] Notice, the univariate Gaussian case is subsumed by linear regression where $X = \iota$ (a vector of ones). Then, the likelihood as described earlier,

$$\ell\left(\beta, \sigma^2 \mid y, X\right) \propto \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\left\{(n-p) s^2 + (\beta - b)^T X^T X (\beta - b)\right\}\right]$$

becomes

$$\ell\left(\beta = \theta, \sigma^2 \mid y, X = \iota\right) \propto \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\left\{(n-1) s^2 + n (\theta - \overline{y})^2\right\}\right]$$

where $\theta = \beta$, $b = \left(X^T X\right)^{-1} X^T y = \overline{y}$, $p = 1$, and $X^T X = n$.

Collecting terms and rewriting, we have

$$\begin{aligned} p\left(\beta,\sigma^2 \mid y,X;\beta_0,\sigma^2\Omega_0^{-1},\nu_0,\sigma_0^2\right) &\propto \left(\sigma^2\right)^{-[(\nu_0+n)/2+1]}\exp\left[-\frac{\sigma_n^2}{2\sigma^2}\right] \\ &\quad \times\sigma^{-p}\exp\left[-\frac{1}{2\sigma^2}\left(\beta-\overline{\beta}\right)^T\Omega_n\left(\beta-\overline{\beta}\right)\right] \end{aligned}$$

where

$$\begin{aligned} \overline{\beta} &= \left(\Omega_0+X^TX\right)^{-1}\left(\Omega_0\beta_0+X^TXb\right) \\ \Omega_n &= \left(\Omega_0+X^TX\right) \end{aligned}$$

and

$$\nu_n\sigma_n^2 = \nu_0\sigma_0^2+(n-p)\,s^2+\left(\beta_0-\overline{\beta}\right)^T\Omega_0\left(\beta_0-\overline{\beta}\right)+\left(\widehat{\beta}-\overline{\beta}\right)^TX^TX\left(\widehat{\beta}-\overline{\beta}\right)$$

where $\nu_n=\nu_0+n$. The conditional posterior of β given σ^2 is Gaussian$\left(\overline{\beta},\sigma^2\Omega_n^{-1}\right)$.

Completing the square

The derivation of the above joint posterior follows from the matrix version of completing the square where Ω_0 and X^TX are square, symmetric, full rank $p\times p$ matrices. The exponents from the prior for the mean and likelihood are

$$\left(\beta-\beta_0\right)^T\Omega_0\left(\beta-\beta_0\right)+\left(\beta-\widehat{\beta}\right)^TX^TX\left(\beta-\widehat{\beta}\right)$$

Expanding and rearranging gives

$$\beta^T\left(\Omega_0+X^tX\right)\beta-2\left(\Omega_0\beta_0+X^TX\widehat{\beta}\right)^T\beta+\beta_0^T\Omega_0\beta_0+\widehat{\beta}^TX^TX\widehat{\beta} \quad (6.1)$$

The latter two terms are constants not involving β (and can be ignored when writing the kernel for the conditional posterior) which we'll add to when we complete the square. Now, write out the square centered around $\overline{\beta}$

$$\begin{aligned} \left(\beta-\overline{\beta}\right)^T\left(\Omega_0+X^TX\right)\left(\beta-\overline{\beta}\right) &= \beta^T\left(\Omega_0+X^TX\right)\beta \\ &\quad -2\overline{\beta}^T\left(\Omega_0+X^TX\right)\beta+\overline{\beta}^T\left(\Omega_0+X^TX\right)\overline{\beta} \end{aligned}$$

Substitute for $\overline{\beta}$ in the second term on the right hand side and the first two terms are identical to the two terms in equation (6.1). Hence, the exponents from the prior for the mean and likelihood in (6.1) are equal to

$$\begin{aligned} &\left(\beta-\overline{\beta}\right)^T\left(\Omega_0+X^TX\right)\left(\beta-\overline{\beta}\right) \\ &-\overline{\beta}^T\left(\Omega_0+X^TX\right)\overline{\beta}+\beta_0^T\Omega_0\beta_0+\widehat{\beta}^TX^TX\widehat{\beta} \end{aligned}$$

which can be rewritten as

$$\begin{aligned}&\left(\beta-\overline{\beta}\right)^{T}\left(\Omega_{0}+X^{T}X\right)\left(\beta-\overline{\beta}\right)\\&+\left(\beta_{0}-\overline{\beta}\right)^{T}\Omega_{0}\left(\beta_{0}-\overline{\beta}\right)+\left(\widehat{\beta}-\overline{\beta}\right)^{T}X^{T}X\left(\widehat{\beta}-\overline{\beta}\right)\end{aligned}$$

or (in the form analogous to the univariate Gaussian case)

$$\begin{aligned}&\left(\beta-\overline{\beta}\right)^{T}\left(\Omega_{0}+X^{T}X\right)\left(\beta-\overline{\beta}\right)\\&+\left(\beta_{0}-\widehat{\beta}\right)^{T}\left(\Omega_{1}\Omega_{n}^{-1}\Omega_{0}\Omega_{n}^{-1}\Omega_{1}+\Omega_{0}\Omega_{n}^{-1}\Omega_{1}\Omega_{n}^{-1}\Omega_{0}\right)\left(\beta_{0}-\widehat{\beta}\right)\end{aligned}$$

where $\Omega_1 = X^T X$.

Stacked regression

Bayesian linear regression with conjugate priors works as if we have a prior sample $\{X_0, y_0\}$, $\Omega_0 = X_0^T X_0$, and initial estimates

$$\beta_0 = \left(X_0^T X_0\right)^{-1} X_0^T y_0$$

Then, we combine this initial "evidence" with new evidence to update our beliefs in the form of the posterior. Not surprisingly, the posterior mean is a weighted average of the two "samples" where the weights are based on the relative precision of the two "samples".

Marginal posterior distributions

The marginal posterior for β on integrating out σ^2 is noncentral, scaled multivariate Student $\mathrm{t}_p\left(\overline{\beta}, \sigma_n^2\Omega_n^{-1}, \nu_0+n\right)$

$$\begin{aligned}p\left(\beta \mid y, X\right) &\propto \left[\nu_n\sigma_n^2+\left(\beta-\overline{\beta}\right)^{T}\Omega_n\left(\beta-\overline{\beta}\right)\right]^{-\frac{\nu_0+n+p}{2}}\\&\propto \left[1+\frac{1}{\nu_n\sigma_n^2}\left(\beta-\overline{\beta}\right)^{T}\Omega_n\left(\beta-\overline{\beta}\right)\right]^{-\frac{\nu_0+n+p}{2}}\end{aligned}$$

where $\Omega_n = \Omega_0 + X^T X$. This result corresponds with the univariate Gaussian case and is derived analogously by transformation of variables where $z = \frac{A}{2\sigma^2}$ where $A = \sigma_n^2 + \left(\beta-\overline{\beta}\right)^{T}\Omega_n\left(\beta-\overline{\beta}\right)$. The marginal posterior for σ^2 is inverted-chi square$\left(\sigma^2; \nu_n, \sigma_n^2\right)$.

Derivation of the marginal posterior for β is as follows.

$$\begin{aligned}p\left(\beta \mid y\right) &= \int_0^\infty p\left(\beta, \sigma^2 \mid y\right) d\sigma^2\\&= \int_0^\infty \left(\sigma^2\right)^{-\frac{n+\nu_0+p+2}{2}} \exp\left[-\frac{A}{2\sigma^2}\right] d\sigma^2\end{aligned}$$

Utilizing $\sigma^2 = \frac{A}{2z}$ and $dz = -\frac{2z^2}{A}d\sigma^2$ or $d\sigma^2 = -\frac{A}{2z^2}dz$, ($-1$ and 2 are constants and can be ignored when deriving the kernel)

$$\begin{aligned} p(\beta \mid y) &\propto \int_0^\infty \left(\frac{A}{2z}\right)^{-\frac{n+\nu_0+p+2}{2}} \frac{A}{2z^2}\exp[-z]\,dz \\ &\propto A^{-\frac{n+\nu_0+p}{2}} \int_0^\infty z^{\frac{n+\nu_0+p}{2}-1}\exp[-z]\,dz \end{aligned}$$

The integral $\int_0^\infty z^{\frac{n+\nu_0+k}{2}-1}\exp[-z]\,dz$ is a constant since it is the kernel of a gamma density and therefore can be ignored when deriving the kernel of the marginal posterior for beta

$$\begin{aligned} p(\beta \mid y) &\propto A^{-\frac{n+\nu_0+p}{2}} \\ &\propto \left[\nu_n\sigma_n^2 + \left(\beta-\overline{\beta}\right)^T\Omega_n\left(\beta-\overline{\beta}\right)\right]^{-\frac{n+\nu_0+p}{2}} \\ &\propto \left[1+\frac{\left(\beta-\overline{\beta}\right)^T\Omega_n\left(\beta-\overline{\beta}\right)}{\nu_n\sigma_n^2}\right]^{-\frac{n+\nu_0+p}{2}} \end{aligned}$$

the kernel for a noncentral, scaled (multivariate) Student $\mathrm{t}_p\left(\beta;\overline{\beta},\sigma_n^2\Omega_n^{-1},n+\nu_0\right)$.

6.7.3 Uninformative priors

Again, the case of uninformative priors is relatively straightforward. Since priors convey no information, the prior for the mean is uniform (proportional to a constant, $\Omega_0 \to 0$) and the prior for σ^2 has $\nu_0 \to 0$ degrees of freedom so that the joint prior is $p\left(\beta,\sigma^2\right) \propto \left(\sigma^2\right)^{-1}$.

The joint posterior is

$$p\left(\beta,\sigma^2 \mid y\right) \propto \left(\sigma^2\right)^{-[n/2+1]}\exp\left[-\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)\right]$$

Since $y = Xb + e$ where $b = \left(X^TX\right)^{-1}X^Ty$, the joint posterior can be written

$$p\left(\beta,\sigma^2 \mid y\right) \propto \left(\sigma^2\right)^{-[n/2+1]}\exp\left[-\frac{1}{2\sigma^2}\left\{(n-p)s^2 + (\beta-b)^TX^TX(\beta-b)\right\}\right]$$

Or, factoring into the conditional posterior for β and marginal for σ^2, we have

$$\begin{aligned} p\left(\beta,\sigma^2 \mid y\right) &\propto p\left(\sigma^2 \mid y\right)p\left(\beta \mid \sigma^2,y\right) \\ &\propto \left(\sigma^2\right)^{-[(n-p)/2+1]}\exp\left[-\frac{\sigma_n^2}{2\sigma^2}\right] \\ &\quad \times\sigma^{-p}\exp\left[-\frac{1}{2\sigma^2}(\beta-b)^TX^TX(\beta-b)\right] \end{aligned}$$

where

$$\sigma_n^2 = (n-p)\, s^2$$

Hence, the conditional posterior for β given σ^2 is Gaussian$\left(b, \sigma^2 \left(X^T X\right)^{-1}\right)$.

The marginal posterior for β is multivariate Student $\mathrm{t}_p\left(\beta; b, s^2 \left(X^T X\right)^{-1}, n-p\right)$, the classical estimator. Derivation of the marginal posterior for β is analogous to that above. Let $z = \frac{A}{2\sigma^2}$ where $A = (n-p)\, s^2 + (\beta - b)^T X^T X (\beta - b)$. Integrating σ^2 out of the joint posterior produces the marginal posterior for β.

$$\begin{aligned} p(\beta \mid y) &\propto \int p\left(\beta, \sigma^2 \mid y\right) d\sigma^2 \\ &\propto \int \left(\sigma^2\right)^{-\frac{n+2}{2}} \exp\left[-\frac{A}{2\sigma^2}\right] d\sigma^2 \end{aligned}$$

Substitution yields

$$\begin{aligned} p(\beta \mid y) &\propto \int \left(\frac{A}{2z}\right)^{-\frac{n+2}{2}} \frac{A}{2z^2} \exp[-z]\, dz \\ &\propto A^{-\frac{n}{2}} \int z^{\frac{n}{2}-1} \exp[-z]\, dz \end{aligned}$$

As before, the integral involves the kernel of a gamma distribution, a constant which can be ignored. Therefore, we have

$$\begin{aligned} p(\beta \mid y) &\propto A^{-\frac{n}{2}} \\ &\propto \left[(n-p)\, s^2 + (\beta - b)^T X^T X (\beta - b)\right]^{-\frac{n}{2}} \\ &\propto \left[1 + \frac{(\beta - b)^T X^T X (\beta - b)}{(n-p)\, s^2}\right]^{-\frac{n}{2}} \end{aligned}$$

which is multivariate Student $\mathrm{t}_p\left(\beta; b, s^2 \left(X^T X\right)^{-1}, n-p\right)$.

6.8 Bayesian linear regression with general error structure

Now, we consider Bayesian regression with a more general error structure. That is, the *DGP* is

$$y = X\beta + \varepsilon, \quad (\varepsilon \mid X) \sim N(0, \Sigma)$$

First, we consider the known variance case, then take up the unknown variance case.

6.8.1 Known variance

If the error variance, Σ, is known, we simply repeat the Bayesian linear regression approach discussed above for the known variance case after transforming all variables via the Cholesky decomposition of Σ. Let

$$\Sigma = \Gamma\Gamma^T$$

and

$$\Sigma^{-1} = \left(\Gamma^T\right)^{-1}\Gamma^{-1}$$

Then, the *DGP* is

$$\Gamma^{-1}y = \Gamma^{-1}X\beta + \Gamma^{-1}\varepsilon$$

where

$$\Gamma^{-1}\varepsilon \sim N\left(0, I_n\right)$$

With informed priors for β, $p\left(\beta \mid \Sigma\right) \sim N\left(\beta_0, \Sigma_\beta\right)$ where it is as if $\Sigma_\beta = \left(X_0^T\Sigma_0^{-1}X_0\right)^{-1}$, the posterior distribution for β conditional on Σ is

$$p\left(\beta \mid \Sigma, y, X; \beta_0, \Sigma_\beta\right) \sim N\left(\overline{\beta}, V_\beta\right)$$

where

$$\begin{aligned}
\overline{\beta} &= \left(X_0^T\Sigma_0^{-1}X_0 + X^T\Sigma^{-1}X\right)^{-1}\left(X_0^T\Sigma_0^{-1}X_0\beta_0 + X^T\Sigma^{-1}X\widehat{\beta}\right) \\
&= \left(\Sigma_\beta^{-1} + X^T\Sigma^{-1}X\right)^{-1}\left(\Sigma_\beta^{-1}\beta_0 + X^T\Sigma^{-1}X\widehat{\beta}\right) \\
\widehat{\beta} &= \left(X^T\Sigma^{-1}X\right)^{-1}X^T\Sigma^{-1}y
\end{aligned}$$

and

$$\begin{aligned}
V_\beta &= \left(X_0^T\Sigma_0^{-1}X_0 + X^T\Sigma^{-1}X\right)^{-1} \\
&= \left(\Sigma_\beta^{-1} + X^T\Sigma^{-1}X\right)^{-1}
\end{aligned}$$

It is instructive to once again backtrack to develop the conditional posterior distribution. The likelihood function for known variance is

$$\ell\left(\beta \mid \Sigma, y, X\right) \propto \exp\left[-\frac{1}{2}\left(y - X\beta\right)^T\Sigma^{-1}\left(y - X\beta\right)\right]$$

Conditional Gaussian priors are

$$p\left(\beta \mid \Sigma\right) \propto \exp\left[-\frac{1}{2\sigma^2}\left(\beta - \beta_0\right)^T V_\beta^{-1}\left(\beta - \beta_0\right)\right]$$

The conditional posterior is the product of the prior and likelihood

$$\begin{aligned}
p\left(\beta \mid \sigma^2, y, X\right) &\propto \exp\left[-\frac{1}{2\sigma^2}\left\{\begin{array}{c}\left(y - X\beta\right)^T\Sigma^{-1}\left(y - X\beta\right) \\ +\left(\beta - \beta_0\right)^T V_\beta^{-1}\left(\beta - \beta_0\right)\end{array}\right\}\right] \\
&= \exp\left[-\frac{1}{2\sigma^2}\left\{\begin{array}{c}y^T\Sigma^{-1}y - 2y^T\Sigma^{-1}X\beta + \beta^T X^T\Sigma^{-1}X\beta \\ +\beta^T V_\beta^{-1}\beta - 2\beta_0^T V_\beta^{-1}\beta + \beta_0^T V_\beta^{-1}\beta_0\end{array}\right\}\right]
\end{aligned}$$

The first and last terms in the exponent do not involve β (are constants) and can ignored as they are absorbed through normalization. This leaves

$$\begin{aligned} p\left(\beta \mid \sigma^2, y, X\right) &\propto \exp\left[-\frac{1}{2\sigma^2}\left\{\begin{array}{c} -2y^T\Sigma^{-1}X\beta + \beta^T X^T\Sigma^{-1}X\beta \\ +\beta^T V_\beta^{-1}\beta - 2\beta_0^T V_\beta^{-1}\beta \end{array}\right\}\right] \\ &= \exp\left[-\frac{1}{2\sigma^2}\left\{\begin{array}{c} \beta^T\left(V_\beta^{-1} + X^T\Sigma^{-1}X\right)\beta \\ -2\left(y^T\Sigma^{-1}X + \beta_0^T V_\beta^{-1}\right)\beta \end{array}\right\}\right] \end{aligned}$$

which can be recognized as the expansion of the conditional posterior claimed above.

$$\begin{aligned} p\left(\beta \mid \Sigma, y, X\right) &\sim N\left(\overline{\beta}, V_\beta\right) \\ &\propto \exp\left[-\frac{1}{2}\left(\beta-\overline{\beta}\right)^T V_\beta^{-1}\left(\beta-\overline{\beta}\right)\right] \\ &= \exp\left[-\frac{1}{2}\left(\beta-\overline{\beta}\right)^T\left(V_\beta^{-1} + X^T\Sigma^{-1}X\right)\left(\beta-\overline{\beta}\right)\right] \\ &= \exp\left[-\frac{1}{2}\left\{\begin{array}{c} \beta^T\left(V_\beta^{-1} + X^T\Sigma^{-1}X\right)\beta \\ -2\overline{\beta}^T\left(V_\beta^{-1} + X^T\Sigma^{-1}X\right)\beta \\ +\overline{\beta}^T\left(V_\beta^{-1} + X^T\Sigma^{-1}X\right)\overline{\beta} \end{array}\right\}\right] \\ &= \exp\left[-\frac{1}{2}\left\{\begin{array}{c} \beta^T\left(X_0^T X_0 + X^T X\right)\beta \\ -2\left(y^T\Sigma^{-1}X + \beta_0^T V_\beta^{-1}\right)^T\beta \\ +\overline{\beta}^T\left(X_0^T X_0 + X^T X\right)\overline{\beta} \end{array}\right\}\right] \end{aligned}$$

The last term in the exponent is all constants (does not involve β) so its absorbed through normalization and disregarded for comparison of kernels. Hence,

$$\begin{aligned} p\left(\beta \mid \sigma^2, y, X\right) &\propto \exp\left[-\frac{1}{2}\left(\beta-\overline{\beta}\right)^T V_\beta^{-1}\left(\beta-\overline{\beta}\right)\right] \\ &\propto \exp\left[-\frac{1}{2\sigma^2}\left\{\begin{array}{c} \beta^T\left(X_0^T X_0 + X^T X\right)\beta \\ -2\left(y^T\Sigma^{-1}X + \beta_0^T V_\beta^{-1}\right)^T\beta \end{array}\right\}\right] \end{aligned}$$

as claimed.

6.8.2 Unknown variance

Bayesian linear regression with unknown general error structure, Σ, is something of a composite of ideas developed for exchangeable ($\sigma^2 I_n$ error structure) Bayesian regression and the multivariate Gaussian case with mean

and variance unknown where each draw is an element of the y vector and X is an $n \times p$ matrix of regressors. A Gaussian likelihood is

$$\begin{aligned}
\ell(\beta, \Sigma \mid y, X) &\propto |\Sigma|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}(y-\beta)^T \Sigma^{-1}(y-\beta)\right] \\
&\propto |\Sigma|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\left\{\begin{array}{c}(y-Xb)^T \Sigma^{-1}(y-Xb) \\ +(b-\beta)^T X^T \Sigma^{-1} X (b-\beta)\end{array}\right\}\right] \\
&\propto |\Sigma|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\left\{(n-p)s^2 + (b-\beta)^T X^T \Sigma^{-1} X (b-\beta)\right\}\right]
\end{aligned}$$

where $b = \left(X^T \Sigma^{-1} X\right)^{-1} X^T \Sigma^{-1} y$ and $s^2 = \frac{1}{n-p}(y-Xb)^T \Sigma^{-1}(y-Xb)$. Combine the likelihood with a Gaussian-inverted Wishart prior

$$\begin{aligned}
p(\beta \mid \Sigma; \beta_0, \Sigma_\beta) \times p\left(\Sigma^{-1}; \nu, \Psi\right) &\propto \exp\left[-\frac{1}{2}(\beta-\beta_0)^T \Sigma_\beta^{-1}(\beta-\beta_0)\right] \\
&\quad \times |\Psi|^{\frac{\nu}{2}} |\Sigma|^{-\frac{\nu+p+1}{2}} \exp\left[-\frac{tr\left(\Psi\Sigma^{-1}\right)}{2}\right]
\end{aligned}$$

where $tr(\cdot)$ is the trace of the matrix, it is as if $\Sigma_\beta = \left(X_0^T \Sigma_0^{-1} X_0\right)^{-1}$, and ν is degrees of freedom to produce the joint posterior

$$\begin{aligned}
p(\beta, \Sigma \mid y, X) &\propto |\Psi|^{\frac{\nu}{2}} |\Sigma|^{-\frac{\nu+n+p+1}{2}} \exp\left[-\frac{tr\left(\Psi\Sigma^{-1}\right)}{2}\right] \\
&\quad \times \exp\left[-\frac{1}{2}\left\{\begin{array}{c}(n-p)s^2 \\ +(b-\beta)^T X^T \Sigma^{-1} X (b-\beta) \\ +(\beta-\beta_0)^T \Sigma_\beta^{-1}(\beta-\beta_0)\end{array}\right\}\right]
\end{aligned}$$

Completing the square

Completing the square involves the matrix analog to the univariate unknown mean and variance case. Consider the exponent (in braces)

$$\begin{aligned}
& (n-p)s^2 + (b-\beta)^T X^T \Sigma^{-1} X (b-\beta) + (\beta-\beta_0)^T \Sigma_\beta^{-1}(\beta-\beta_0) \\
= \;& (n-p)s^2 + b^T X^T \Sigma^{-1} X b - 2\beta^T X^T \Sigma^{-1} X b + \beta^T X^T \Sigma^{-1} X \beta \\
& + \beta^T \Sigma_\beta^{-1} \beta - 2\beta^T \Sigma_\beta^{-1} \beta_0 + \beta_0^T \Sigma_\beta^{-1} \beta_0 \\
= \;& (n-p)s^2 + \beta^T \left(\Sigma_\beta^{-1} + X^T \Sigma^{-1} X\right) \beta \\
& - 2\beta^T V_\beta^{-1} \overline{\beta} + b^T X^T \Sigma^{-1} X b + \beta_0^T \Sigma_\beta^{-1} \beta_0 \\
= \;& (n-p)s^2 + \beta^T V_\beta^{-1} \beta - 2\beta^T V_\beta^{-1} \overline{\beta} + b^T X^T \Sigma^{-1} X b + \beta_0^T \Sigma_\beta^{-1} \beta_0
\end{aligned}$$

where

$$\begin{aligned}
\overline{\beta} &= \left(\Sigma_\beta^{-1} + X^T \Sigma^{-1} X\right)^{-1} \left(\Sigma_\beta^{-1} \beta_0 + X^T \Sigma^{-1} X b\right) \\
&= V_\beta \left(\Sigma_\beta^{-1} \beta_0 + X^T \Sigma^{-1} X b\right)
\end{aligned}$$

and $V_\beta = \left(\Sigma_\beta^{-1} + X^T\Sigma^{-1}X\right)^{-1}$.

Variation in β around $\overline{\beta}$ is

$$\left(\beta - \overline{\beta}\right)^T V_\beta^{-1}\left(\beta - \overline{\beta}\right) = \beta^T V_\beta^{-1}\beta - 2\beta^T V_\beta^{-1}\overline{\beta} + \overline{\beta}^T V_\beta^{-1}\overline{\beta}$$

The first two terms are identical to two terms in the posterior involving β and there is apparently no recognizable kernel from these expressions. The joint posterior is

$$\begin{aligned}
& p\left(\beta, \Sigma \mid y, X\right) \\
\propto\ & |\Psi|^{\frac{\nu}{2}} |\Sigma|^{-\frac{\nu+n+p+1}{2}} \exp\left[-\frac{tr\left(\Psi\Sigma^{-1}\right)}{2}\right] \\
& \times \exp\left[-\frac{1}{2}\left\{\begin{array}{c} \left(\beta - \overline{\beta}\right)^T V_\beta^{-1}\left(\beta - \overline{\beta}\right) \\ +\left(n-p\right)s^2 - \overline{\beta}^T V_\beta^{-1}\overline{\beta} \\ +b^T X^T\Sigma^{-1}Xb + \beta_0^T\Sigma_\beta^{-1}\beta_0 \end{array}\right\}\right] \\
\propto\ & |\Psi|^{\frac{\nu}{2}} |\Sigma|^{-\frac{\nu+n+p+1}{2}} \exp\left[-\frac{1}{2}\left\{\begin{array}{c} tr\left(\Psi\Sigma^{-1}\right) + \left(n-p\right)s^2 \\ -\overline{\beta}^T V_\beta^{-1}\overline{\beta} \\ +b^T X^T\Sigma^{-1}Xb + \beta_0^T\Sigma_\beta^{-1}\beta_0 \end{array}\right\}\right] \\
& \times \exp\left[-\frac{1}{2}\left\{\ \left(\beta - \overline{\beta}\right)^T V_\beta^{-1}\left(\beta - \overline{\beta}\right)\ \right\}\right]
\end{aligned}$$

Therefore, we write the conditional posteriors for the parameters of interest. First, we focus on β then we take up Σ.

The conditional posterior for β conditional on Σ involves collecting all terms involving β. Hence, the conditional posterior for β is $\left(\beta \mid \Sigma\right) \sim N\left(\overline{\beta}, V_\beta\right)$ or

$$p\left(\beta \mid \Sigma, y, X\right) \propto \exp\left[-\frac{1}{2}\left\{\ \left(\beta - \overline{\beta}\right)^T V_\beta^{-1}\left(\beta - \overline{\beta}\right)\ \right\}\right]$$

Inverted-Wishart kernel

Now, we gather all terms involving Σ and write the conditional posterior for Σ.

$$\begin{aligned}
& p\left(\Sigma \mid \beta, y, X\right) \\
\propto\ & |\Psi|^{\frac{\nu}{2}} |\Sigma|^{-\frac{\nu+n+p+1}{2}} \exp\left[-\frac{1}{2}\left\{\begin{array}{c} tr\left(\Psi\Sigma^{-1}\right) + \left(n-p\right)s^2 \\ +\left(b-\beta\right)^T X^T\Sigma^{-1}X\left(b-\beta\right) \end{array}\right\}\right] \\
\propto\ & |\Psi|^{\frac{\nu}{2}} |\Sigma|^{-\frac{\nu+n+p+1}{2}} \exp\left[-\frac{1}{2}\left\{\begin{array}{c} tr\left(\Psi\Sigma^{-1}\right) + \\ \left(y-Xb\right)^T\Sigma^{-1}\left(y-Xb\right) \\ +\left(b-\beta\right)^T X^T\Sigma^{-1}X\left(b-\beta\right) \end{array}\right\}\right] \\
\propto\ & |\Psi|^{\frac{\nu}{2}} |\Sigma|^{-\frac{\nu+n+p+1}{2}} \exp\left[-\frac{1}{2}\left\{tr\left(\left[\begin{array}{c} \Psi + \left(y-Xb\right)^T\left(y-Xb\right) \\ +\left(b-\beta\right)^T X^T X\left(b-\beta\right) \end{array}\right]\Sigma^{-1}\right)\right\}\right]
\end{aligned}$$

We can identify the kernel as an inverted-Wishart involving the trace of a square, symmetric matrix, call it Ψ_n, multiplied by Σ^{-1}.

The above joint posterior can be rewritten as an inverted-Wishart$\left(\Sigma^{-1}; \nu + n, \Psi_n\right)$

$$p\left(\theta, \Sigma \mid y\right) \propto \left|\Psi_n\right|^{\frac{\nu+n}{2}} \left|\Sigma\right|^{-\frac{\nu+n+p+1}{2}} \exp\left[-\frac{1}{2} tr\left(\Psi_n \Sigma^{-1}\right)\right]$$

where

$$\Psi_n = \Psi + \left(y - Xb\right)^T \left(y - Xb\right) + \left(b - \beta\right)^T X^T X \left(b - \beta\right)$$

With conditional posteriors in hand, we can employ *McMC* strategies (namely, a Gibbs sampler) to draw inferences around the parameters of interest, β and Σ. That is, we sequentially draw β conditional on Σ and Σ, in turn, conditional on β. We discuss *McMC* strategies (both the Gibbs sampler and its generalization, the Metropolis-Hastings algorithm) later.

6.8.3 (Nearly) uninformative priors

As discussed by Gelman, et al [2004] uninformative priors for this case is awkward, at best. What does it mean to posit uninformative priors for a regression with general error structure? Consistent probability assignment suggests that either we have some priors about the correlation structure or heteroskedastic nature of the errors (informative priors) or we know nothing about the error structure (uninformative priors). If priors are uninformative, then maximum entropy probability assignment suggests we assign independent and unknown homoskedastic errors. Hence, we discuss nearly uninformative priors for this general error structure regression.

The joint uninformative prior (with a locally uniform prior for β) is

$$p\left(\beta, \Sigma\right) \propto \left|\Sigma\right|^{-\frac{1}{2}}$$

and the joint posterior is

$$\begin{aligned} p\left(\beta, \Sigma \mid y, X\right) &\propto \left|\Sigma\right|^{-\frac{1}{2}} \left|\Sigma\right|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\left\{\left(n-p\right)s^2 + \left(b-\beta\right)^T X^T \Sigma^{-1} X \left(b-\beta\right)\right\}\right] \\ &\propto \left|\Sigma\right|^{-\frac{n+1}{2}} \exp\left[-\frac{1}{2}\left\{\left(n-p\right)s^2 + \left(b-\beta\right)^T X^T \Sigma^{-1} X \left(b-\beta\right)\right\}\right] \\ &\propto \left|\Sigma\right|^{-\frac{n+1}{2}} \exp\left[-\frac{1}{2} tr\left(S\left(\beta\right)\Sigma^{-1}\right)\right] \end{aligned}$$

where now $S\left(\beta\right) = \left(y - Xb\right)^T \left(y - Xb\right) + \left(b - \beta\right)^T X^T X \left(b - \beta\right)$. Then, the conditional posterior for β given Σ is $N\left(b, \left(X^T \Sigma^{-1} X\right)^{-1} \mid \Sigma\right)$

$$p\left(\beta \mid \Sigma, y, X\right) \propto \exp\left[-\frac{n}{2}\left(\beta - b\right)^T X^T \Sigma^{-1} X \left(\beta - b\right)\right]$$

The conditional posterior for Σ given β is inverted-Wishart$\left(\Sigma^{-1}; n, \Psi_n\right)$

$$p\left(\beta, \Sigma \mid y\right) \propto \left|\Psi_n\right|^{\frac{n}{2}} \left|\Sigma\right|^{-\frac{n+1}{2}} \exp\left[-\frac{1}{2} tr\left(\Psi_n \Sigma^{-1}\right)\right]$$

where

$$\Psi_n = \left(y - Xb\right)^T \left(y - Xb\right) + \left(b - \beta\right)^T X^T X \left(b - \beta\right)$$

As with informed priors, a Gibbs sampler (sequential draws from the conditional posteriors) can be employed to draw inferences for the uninformative prior case.

Next, we discuss posterior simulation, a convenient and flexible strategy for drawing inference from the evidence and (conjugate) priors.

6.9 Appendix: summary of conjugacy

focal parameter(s) θ	*prior* $\pi(\theta)$	*likelihood* $\ell(\theta \mid y)$	*posterior* $\pi(\theta \mid y)$
discrete data:			
beta-binomial p	beta $\propto p^{a-1}(1-p)^{b-1}$	binomial $\propto p^{s}(1-p)^{n-s}$	beta $\propto p^{a+s-1}(1-p)^{b+n-s-1}$
gamma-poisson λ	gamma $\propto \lambda^{a-1}e^{-b\lambda}$	poisson $\propto \lambda^{s}e^{-n\lambda}$	gamma $\propto \lambda^{a+s-1}e^{-(b+n)\lambda}$
gamma-exponential θ	gamma $\propto \theta^{a-1}e^{-b\theta}$	exponential $\propto \theta^{n}e^{-s\theta}$	gamma $\propto \theta^{a+n-1}e^{-(b+s)\theta}$
beta-negative binomial p	beta $\propto p^{a-1}(1-p)^{b-1}$	negative binomial $\propto p^{nr}(1-p)^{s}$	beta $\propto p^{a+nr-1}(1-p)^{b+s-1}$
beta-binomial-hypergeometric[8] k k unknown population success N known population size n known sample size x known sample success	beta-binomial $\binom{n}{x}\frac{\Gamma(a+x)\Gamma(b+n-x)\Gamma(a+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+n)}$, $x=0,1,2,\ldots,n$	hypergeometric $\frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$ sampling without replacement	beta-binomial $\binom{N-n}{k-x}\frac{\Gamma(a+k)\Gamma(b+N-k)\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(b+n-x)\Gamma(a+b+N)}$, $k=x,x+1,\ldots,$ $x+N-n$
multinomial-Dirichlet θ (vector)	Dirichlet $\propto \prod_{i=1}^{K}\theta_i^{a_i-1}$	multinomial $\propto \theta_1^{s_1}\cdots\theta_K^{s_K}$	Dirichlet $\propto \prod_{i=1}^{K}\theta_i^{a_i+s_i-1}$

$s=\sum_{i=1}^{n} y_i$, $\binom{n}{x}=\frac{n!}{x!(n-x)!}$,

$\Gamma(z)=\int_0^\infty e^{-t}t^{z-1}dt$, $\Gamma(n)=(n-1)!$ for n a positive integer, $B(a,b)=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

[8] See Dyer and Pierce [1993].

focal parameter(s) θ	*prior* $\pi(\theta)$	*likelihood* $\ell(\theta \mid y)$	*posterior* $\pi(\theta \mid y)$
continuous data:			
Pareto-uniform w w unknown upper bound, 0 known lower bound	Pareto $\frac{ab^a}{w^{a+1}}$	uniform $\frac{1}{w^n}$, $w > \max(x_i)$	Pareto $\propto \frac{(a+n)\max[b,x_i]^{a+n}}{w^{a+n+1}}$
Pareto-Pareto β β unknown precision, α known shape	Pareto $\frac{ab^{a\alpha}}{\beta^{a+1}}$, $\beta > b$	Pareto $\propto \beta^{n\alpha}$, $0 < \beta < \min(x_i)$	Pareto $\propto \frac{(a-\alpha n)b^{(a-\alpha n)\alpha}}{\beta^{a-\alpha n+1}}$, $a > \alpha n, \beta > b$
gamma-Pareto α α unknown shape, β known precision	gamma $\propto \frac{\alpha^{a-1}e^{-\alpha/b}}{b^a\Gamma(a)}$, $\alpha > 0$	Pareto $\propto \frac{\alpha^n\beta^{n\alpha}}{m^{\alpha+1}}$, $m = \prod_{i=1}^n x_i$, $0 < \beta < \min(x_i)$	gamma $\propto \frac{\alpha^{a+n-1}e^{-\alpha/b'}}{(b')^{a+n}\Gamma(a+n)}$, $b' = \frac{1}{\frac{1}{b}+\log m - n\log\beta} > 0$
gamma-exponential θ	gamma $\propto \frac{\theta^{a-1}e^{-\theta/b}}{\Gamma(a)b^a}$	exponential $\propto \theta^n e^{-s\theta}$, $s = \sum_{i=1}^n x_i$	gamma $\propto \frac{\theta^{a+n-1}e^{-\theta/b'}}{\Gamma(a+n)(b')^{a+n}}$, $b' = \frac{b}{1+bs}$
inverse gamma-gamma β β unknown rate, α known shape	inverse gamma $\propto \frac{\beta^{1-a}e^{-1/\beta b}}{\Gamma(a)b^a}$	gamma $\propto \frac{e^{-s/\beta}}{\beta^{\alpha n}}$, $s = \sum_{i=1}^n x_i$	inverse gamma $\propto \frac{\beta^{1-a-\alpha n}e^{-1/\beta b'}}{\Gamma(a+\alpha n)(b')^{a+\alpha n}}$, $b' = \frac{b}{1+bs}$
conjugate prior-gamma α α unknown shape, β known rate	nonstandard $\propto \frac{a^{\alpha-1}\beta^{\alpha c}}{\Gamma(\alpha)^b}$, $a, b, c > 0$ $\alpha > 0$	gamma $\propto \frac{m^{\alpha-1}}{\beta^{-\alpha n}\Gamma(\alpha)^n}$, $m = \prod_{i=1}^n x_i$, $x_i > 0$	nonstandard $\propto \frac{(am)^{\alpha-1}\beta^{\alpha(c+n)}}{\Gamma(\alpha)^{b+n}}$

focal parameter(s) θ	*prior* $\pi(\theta)$	*likelihood* $\ell(\theta \mid y)$	*posterior* $\pi(\theta \mid y)$
continuous data:			
normal-normal μ	normal $\propto \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right]$, $\sigma_0^2 = \frac{\sigma^2}{\kappa_0}$	normal $\propto \prod_{i=1}^{n} \exp\left[-\frac{(y_i-\mu)^2}{2\sigma^2}\right] = \exp\left[-\frac{ss}{2\sigma^2}\right]$	normal $\propto \exp\left[-\frac{(\mu-\mu_n)^2}{2\sigma_n^2}\right]$, $\mu_n = \frac{\kappa_0\mu_0+n\overline{y}}{\kappa_0+n}$, $\sigma_n^2 = \frac{\sigma^2}{\kappa_0+n}$
inverse gamma -normal σ^2	inverse gamma $\propto \left(\sigma^2\right)^{-(a+1)} \exp\left[-\frac{b}{\sigma^2}\right]$	normal $\propto \frac{1}{(\sigma^2)^{n/2}} \exp\left[-\frac{ss}{2\sigma^2}\right]$	inverse gamma $\propto \left(\sigma^2\right)^{-\left(\frac{n+2a}{2}+1\right)} \exp\left[-\frac{b+\frac{1}{2}ss}{\sigma^2}\right]$
$\left(\text{normal} \mid \sigma^2\right) \times$ inverse gamma- normal μ, σ^2	$\left(\text{normal} \mid \sigma^2\right) \times$ inverse gamma $\propto \sigma_0^{-1} \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right] \times \left(\sigma^2\right)^{-(a+1)} \exp\left[-\frac{b}{\sigma^2}\right]$, $\sigma_0^2 = \frac{\sigma^2}{\kappa_0}$	normal $\propto \frac{1}{(\sigma^2)^{n/2}} \exp\left[-\frac{ss}{2\sigma^2}\right]$	joint posterior: $\left(\text{normal} \mid \sigma^2\right) \times$ inverse gamma $\propto \sqrt{\frac{\kappa_0'}{\sigma^2}} \exp\left[-\frac{\kappa_0'\left(\mu-\mu_0'\right)^2}{2\sigma^2}\right] \times \left(\sigma^2\right)^{-\left(a'+1\right)} \exp\left[-\frac{b'}{\sigma^2}\right]$; Student t marginal posterior for μ: $\propto \left(1+\frac{\kappa_0 b'\left(\mu-\mu_0'\right)^2}{2}\right)^{-\frac{2a'+1}{2}}$; inverse gamma marginal posterior for σ^2: $\propto \left(\sigma^2\right)^{-\left(a'+1\right)} \exp\left[-\frac{b'}{\sigma^2}\right]$, $a' = a + \frac{n}{2}, \kappa_0' = \kappa_0 + n$, $b' = \frac{1}{b} + \frac{ss}{2} + \frac{\kappa_0 n(\overline{y}-\mu_0)^2}{2(\kappa_0+n)}$, $\mu_0' = \frac{\kappa_0\mu_0+n\overline{y}}{\kappa_0+n}$

$ss = \sum_{i=1}^{n} (y_i - \mu)^2$

focal parameter(s) θ	*prior* $\pi(\theta)$	*likelihood* $\ell(\theta \mid y)$	*posterior* $\pi(\theta \mid y)$
continuous data:			
bilateral bivariate Pareto-uniform l, u	bilateral bivariate Pareto $\frac{a(a+1)(r_2-r_1)^a}{(u-l)^{a+2}}$, $l < r_1, u > r_2$	uniform $\left(\frac{1}{u-l}\right)^n$	bilateral bivariate Pareto $(a+n)(a+n+1)$ $\frac{\left(r_2^{'}-r_1^{'}\right)^{a+n}}{(u-l)^{a+n+2}}$, $r_1^{'} = \min(r_1, x_i)$, $r_2^{'} = \max(r_2, x_i)$
normal-lognormal μ	normal $\propto \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right]$, $\sigma_0^2 = \frac{\sigma^2}{\kappa_0}$	log normal $\propto \prod_{i=1}^{n} \exp\left[-\frac{(\log y_i-\mu)^2}{2\sigma^2}\right]$ $= \exp\left[-\frac{lss}{2\sigma^2}\right]$	normal $\propto \exp\left[-\frac{(\mu-\mu_n)^2}{2\sigma_n^2}\right]$, $\mu_n = \frac{\kappa_0\mu_0+n\log y}{\kappa_0+n}$, $\sigma_n^2 = \frac{\sigma^2}{\kappa_0+n}$
inverse gamma-lognormal σ^2	inverse gamma $\propto \left(\sigma^2\right)^{-(a+1)}$ $\exp\left[-\frac{b}{\sigma^2}\right]$	normal $\propto \frac{1}{(\sigma^2)^{n/2}}$ $\exp\left[-\frac{lss}{2\sigma^2}\right]$	inverse gamma $\propto \left(\sigma^2\right)^{-\left(\frac{n+2a}{2}+1\right)}$ $\exp\left[-\frac{b+\frac{1}{2}lss}{\sigma^2}\right]$

$lss = \sum_{i=1}^{n} (\log y_i - \mu)^2$

continuous data: multivariate normal $\times$ inverted Wishart- multivariate normal μ, Σ	
prior $\pi(\mu, \Sigma)$	
multivariate normal $\pi(\mu \mid \Sigma)$	$\propto \lvert\Sigma\rvert^{-\frac{1}{2}} \exp\left[-\frac{\kappa_0}{2}(\mu-\mu_0)^T \Sigma^{-1}(\mu-\mu_0)\right]$
$\times$ inverted Wishart $\pi(\Sigma)$	$\times \lvert\Psi\rvert^{\frac{\nu}{2}} \lvert\Sigma\rvert^{-\frac{\nu+k+1}{2}} \exp\left[-\frac{tr\left(\Psi\Sigma^{-1}\right)}{2}\right]$
likelihood $\ell(\mu, \Sigma \mid y)$	
multivariate normal	$\propto \lvert\Sigma\rvert^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\left\{(n-1)s^2 + (\overline{y}-\mu)^T \Sigma^{-1}(\overline{y}-\mu)\right\}\right]$
joint posterior $\pi(\mu, \Sigma \mid y)$	
multivariate normal $\pi(\mu \mid \Sigma, y)$	$\propto \exp\left[-\frac{\kappa_0+n}{2}(\mu-\mu_n)^T \Sigma^{-1}(\mu-\mu_n)\right]$
$\times$ inverted Wishart $\pi(\Sigma \mid y)$	$\times \lvert\Psi\rvert^{\frac{\nu+n}{2}} \lvert\Sigma\rvert^{-\frac{\nu+n+k+1}{2}} \exp\left[-\frac{tr\left(\Psi_n\Sigma^{-1}\right)}{2}\right]$
marginal posterior	
multivariate Student t $\pi(\mu \mid y)$	$\propto \left[I + (\kappa_0+n)\left((\mu-\mu_n)^T \Psi_n^{-1}(\mu-\mu_n)\right)\right]^{-\frac{1}{2}(\nu+n+1)}$
inverted Wishart $\pi(\Sigma \mid y)$	$\propto \lvert\Psi\rvert^{\frac{\nu+n}{2}} \lvert\Sigma\rvert^{-\frac{\nu+n+k+1}{2}} \exp\left[-\frac{tr\left(\Psi_n\Sigma^{-1}\right)}{2}\right]$
where	$tr(\cdot)$ is the trace of a matrix,
$\mu_n = \frac{\kappa_0\mu_0+n\overline{y}}{\kappa_0+n}$,	$s^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i-\overline{y})^T \Sigma^{-1}(y_i-\overline{y})$,
$\Psi_n = \Psi + \sum_{i=1}^n (y_i-\overline{y})^T (y_i-\overline{y}) + \frac{\kappa_0 n}{\kappa_0+n}(\mu_0-\overline{y})^T(\mu_0-\overline{y})$	

continuous data:
linear regression
normal $\times$ inverse chi square-normal
β, σ^2

prior $\boldsymbol{\pi}\left(\beta, \sigma^2\right)$	
normal $\pi\left(\beta \mid \sigma^2\right)$	$\propto \sigma^{-p} \exp\left[-\frac{1}{2\sigma^2}\left(\beta-\beta_0\right)^T \Omega_0\left(\beta-\beta_0\right)\right]$
$\times$ inverse chi square $\pi\left(\sigma^2\right)$	$\times\ \sigma^{-(\nu_0/2+1)} \exp\left[-\frac{\nu_0 \sigma_0^2}{2\sigma^2}\right]$
normal likelihood $\ell\left(\beta, \sigma^2 \mid y, X\right)$	
normal	$\propto \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\left\{e^T e+\left(\beta-b\right)^T X^T X\left(\beta-b\right)\right\}\right]$
joint posterior $p\left(\beta, \sigma^2 \mid y, X\right)$	
normal $p\left(\beta \mid \sigma^2, y, X\right)$	$\propto \sigma^{-p} \exp\left[-\frac{1}{2\sigma^2}\left(\beta-\overline{\beta}\right)^T \Omega_n\left(\beta-\overline{\beta}\right)\right]$
$\times$ inverse chi square $\pi\left(\sigma^2 \mid y, X\right)$	$\times\left(\sigma^2\right)^{-[(\nu_0+n)/2+1]} \exp\left[-\frac{\sigma_n^2}{2\sigma^2}\right]$
marginal posterior	
Student t $\pi\left(\beta \mid \sigma^2, y, X\right)$	$\propto\left[1+\frac{1}{\nu_n \sigma_n^2}\left(\beta-\overline{\beta}\right)^T \Omega_n\left(\beta-\overline{\beta}\right)\right]^{-\frac{\nu_0+n+p}{2}}$
$\times$ inverse chi square $\pi\left(\sigma^2 \mid y, X\right)$	$\propto\left(\sigma^2\right)^{-[(\nu_0+n)/2+1]} \exp\left[-\frac{\sigma_n^2}{2\sigma^2}\right]$

where

$$e=y-Xb, \qquad b=\left(X^T X\right)^{-1} X^T y,$$

$$\overline{\beta}=\left(\Omega_0+X^T X\right)^{-1}\left(\Omega_0 \beta_0+X^T X b\right), \qquad \Omega_n=\Omega_0+X^T X,$$

$$\nu_n \sigma_n^2=\nu_0 \sigma_0^2+e^T e+\left(\beta_0-\overline{\beta}\right)^T \Omega_0\left(\beta_0-\overline{\beta}\right)+\left(\widehat{\beta}-\overline{\beta}\right)^T X^T X\left(\widehat{\beta}-\overline{\beta}\right),$$

and $\nu_n=\nu_0+n$

continuous data: linear regression with general variance	
normal $\times$ inverted Wishart-normal	
β, Σ	
prior $\boldsymbol{\pi}(\beta, \Sigma)$	
normal $\pi(\beta \mid \Sigma)$	$\propto \exp\left[-\frac{1}{2}(\beta-\beta_0)^T \Sigma_\beta^{-1}(\beta-\beta_0)\right]$
$\times$ inverted Wishart $\pi(\Sigma)$	$\times \lvert\Psi\rvert^{\frac{\nu}{2}} \lvert\Sigma\rvert^{-\frac{\nu+p+1}{2}} \exp\left[-\frac{tr\left(\Psi\Sigma^{-1}\right)}{2}\right]$
normal likelihood $\ell(\beta, \Sigma \mid y, X)$	
normal	$\propto \lvert\Sigma\rvert^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\left\{(n-p)s^2 + (\beta-b)^T X^T\Sigma^{-1}X(\beta-b)\right\}\right]$
conditional posterior	
normal $p(\beta \mid \Sigma, y, X)$	$\propto \exp\left[-\frac{1}{2}\left(\beta-\overline{\beta}\right)^T V_\beta^{-1}\left(\beta-\overline{\beta}\right)\right]$
inverted Wishart $\pi(\Sigma \mid \beta, y, X)$	$\propto \lvert\Psi\rvert^{\frac{\nu+n}{2}} \lvert\Sigma\rvert^{-\frac{\nu+n+p+1}{2}} \exp\left[-\frac{tr\left(\Psi_n\Sigma^{-1}\right)}{2}\right]$
where	$tr(\cdot)$ is the trace of a matrix,

$$s^2 = \frac{1}{n-p}(y-Xb)^T\Sigma^{-1}(y-Xb), \qquad b = \left(X^T\Sigma^{-1}X\right)^{-1}X^T\Sigma^{-1}y,$$

$$V_\beta = \left(\Sigma_\beta^{-1} + X^T\Sigma^{-1}X\right)^{-1}, \qquad \overline{\beta} = \left(\Sigma_\beta^{-1} + X^T\Sigma^{-1}X\right)^{-1}\left(\Sigma_\beta^{-1}\beta_0 + X^T\Sigma^{-1}Xb\right),$$

$$\Psi_n = \Psi + (y-Xb)^T(y-Xb) + \left(b-\overline{\beta}\right)^T X^TX\left(b-\overline{\beta}\right),$$

$$\text{and } \Sigma_\beta = \left(X_0^T\Sigma_0^{-1}X_0\right)^{-1}$$