Bayesian inference in simple conjugate families

A) Let $\mathbf{x} := \langle x_1, \ldots, x_N \rangle$, and let k denote the number of successes in $\mathbf{x}$. Then

$$\begin{aligned} p(w|\mathbf{x}) &\propto p(\mathbf{x}|w)p(w) \\ &\propto w^k(1-w)^{N-k} w^{a-1}(1-w)^{b-1} \\ &= w^{a+k-1}(1-w)^{b+(n-k)-1}, \end{aligned}$$

which is the kernel of a Beta distribution with parameters $a+k-1$ and $b+(n-k)-1$. Thus, the posterior distribution is $\text{Beta}\big(a+k-1, b+(n-k)-1\big)$.

B) Let

$$Y_1 = \frac{X_1}{X_1 + X_2} \qquad \text{and} \qquad Y_2 = X_1 + X_2.$$

Then, solving for X_1 and X_2 in terms of Y_1 and Y_2, we get

$$X_1 = Y_1 Y_2 \qquad \text{and} \qquad X_2 = (1 - Y_1)Y_2.$$

Since we want to find the joint density of Y_1 and Y_2, we start by finding the Jacobian of the transformation:

$$\begin{aligned} J &= \begin{vmatrix} \partial x_1/\partial y_1 & \partial x_1/\partial y_2 \\ \partial x_2/\partial y_1 & \partial x_2/\partial y_2 \end{vmatrix} \\ &= \begin{vmatrix} y_2 & y_1 \\ -y_2 & 1-y_1 \end{vmatrix} \\ &= y_2(1-y_1) + y_1 y_2 \\ &= y_2. \end{aligned}$$

Note that since X_1 and X_2 are never negative, neither is Y_2, and therefore $|J| = J$. We can now express the joint distribution function in terms of f_{X_1,X_2} as follows:

$$\begin{aligned} f_{Y_1,Y_2}(y_1, y_2) &\propto (y_1 y_2)^{a_1-1} \cdot e^{-y_1 y_2} \big((1-y_1)y_2\big)^{a_2-1} \cdot e^{-(1-y_1)y_2} \cdot y_2 \\ &= \Big(y_1^{a_1-1}(1-y_1)^{a_2-1}\Big)\Big(y_2^{a_1+a_2-1} \cdot e^{-y_2}\Big) \end{aligned}$$

where the expression in the first pair of parentheses is the kernel of a Beta distribution and that in the second is the kernel of a Gamma distribution. We therefore see that

$$y_1 \sim \text{Beta}(a_1, a_2) \qquad \text{and} \qquad y_2 \sim \text{Gamma}(a_1 + a_2, 1).$$

This provides us with a means to simulate a Beta random variable with parameters a_1 and a_2 from Gamma random variables by simulating $X_1 \sim \text{Gamma}(a_1, 1)$ and $X_2 \sim \text{Gamma}(a_2, 1)$, and then computing the ratio $X_1/(X_1 + X_2)$.

C) Let $X_i \sim N(\theta, \sigma^2)$ be independent where the variance is known and $\theta \sim N(m, v)$. Let $\bar{x}$ denote the mean of $x_1, \ldots, x_N$. Then

$$
\begin{aligned}
p(\theta|x_1, \ldots, x_N) &\propto p(x_1, \ldots, x_N|\theta)p(\theta) \\
&= \left(\prod_{i=1}^{N} p(x_i|\theta)\right)p(\theta) \\
&\propto \exp\left(-\sum_{i=1}^{N} \frac{(x_i-\theta)^2}{2\sigma^2}\right)\exp\left(\frac{-(\theta-m)^2}{2v}\right) \\
&\propto \exp\left(\sum_{i=1}^{N} \frac{2\theta x_i - \theta^2}{2\sigma^2}\right)\exp\left(-\frac{\theta^2-2\theta m}{2v}\right) \\
&= \exp\left(\frac{2\theta n\bar{x} - n\theta^2}{2\sigma^2} - \frac{\theta^2-2\theta m}{2v}\right) \\
&= \exp\left(-\left(\frac{n}{2\sigma^2}+\frac{1}{2v}\right)\theta^2 + \left(\frac{n\bar{x}}{\sigma^2}+\frac{m}{v}\right)\theta\right)
\end{aligned}
$$

This last expression is complicated, but has the form $e^{-A\theta^2+B\theta}$, with which we can work until it is further simplified. Note that we still only care about things proportional to the expression involving θ.

$$
\begin{aligned}
\exp\big(-A\theta^2 + B\theta\big) &= \exp\left(\frac{-\theta^2 + (B/A)\theta}{1/A}\right) \\
&\propto \exp\left(\frac{-\big(\theta - (B/2A)\big)^2}{1/A}\right)
\end{aligned}
$$

This is the kernel of a normal distribution with mean and variance as follows:

$$
\mu_{\text{post}} = \frac{B}{2A} = \frac{\dfrac{n\bar{x}}{\sigma^2} + \dfrac{m}{v}}{\dfrac{n}{\sigma^2} + \dfrac{1}{v}}
$$

$$
\sigma^2_{\text{post}} = \frac{1}{2A} = \frac{1}{\dfrac{n}{\sigma^2} + \dfrac{1}{v}}
$$

D) Let $X_i \sim N(\theta, \sigma^2)$ be independent where the mean θ is known and $w := 1/\sigma^2 \sim \text{Gamma}(a, b)$. Let $\hat{\sigma}^2$ denote the standard deviation of $x_1, \ldots, x_N$. Then

$$\begin{aligned}
p(w|x_1, \ldots, x_N) &\propto p(x_1, \ldots, x_N|w)p(w) \\
&= \left(\prod_{i=1}^{N} p(x_i|w)\right)p(w) \\
&\propto w^{N/2} \exp\left(\frac{-w}{2}\sum_{i=1}^{N}(x_i - \theta)^2\right) w^{a-1}\exp(-bw) \\
&= w^{a+N/2-1}\exp\big(-w \cdot N\hat{\sigma}^2/2\big)\exp(-bw) \\
&= w^{a+N/2-1}\exp\big(-(N\hat{\sigma}^2/2 + b)w\big).
\end{aligned}$$

This is the kernel of a Gamma distribution, namely $\text{Gamma}\big(a + N/2, b + N\hat{\sigma}^2/2\big)$. Thus, the posterior distribution for σ^2 is $\text{IG}\big(a + N/2, b + N\hat{\sigma}^2/2\big)$

E) Let $X_i \sim N(\theta, \sigma_i^2)$ be independent where each σ_i^2 is known and $\theta \sim N(m, v)$. Then

$$\begin{aligned}
p(\theta|x_1, \ldots, x_N) &\propto p(x_1, \ldots, x_N|\theta)p(\theta) \\
&= \left(\prod_{i=1}^{N} p(x_i|\theta)\right)p(\theta) \\
&\propto \exp\left(-\sum_{i=1}^{N}\frac{(x_i - \theta)^2}{2\sigma_i^2}\right)\exp\left(\frac{-(\theta - m)^2}{2v}\right) \\
&\propto \exp\left(-\sum_{i=1}^{N}\frac{\theta^2 - 2x_i\theta}{\sigma_i^2}\right)\exp\left(\frac{-(\theta^2 - 2m\theta)}{2v}\right) \\
&= \exp\left(-\frac{\theta^2 - 2m\theta}{2v} - \sum_{i=1}^{N}\frac{\theta^2 - 2x_i\theta}{\sigma_i^2}\right) \\
&= \exp\left(-\left(\frac{1}{2v} + \sum_{i=1}^{N}\frac{1}{2\sigma_i^2}\right)\theta^2 + \left(\frac{m}{v} + \sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}\right)\theta\right)
\end{aligned}$$

As in (C) above, this has the form $e^{-A\theta^2 + B\theta}$, and we solve this in the same way. Thereby we find that $\theta|x_1, \ldots, x_N$ is normally distributed with mean and variance given by

$$\mu_{\text{post}} = \frac{\dfrac{m}{v} + \displaystyle\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}}{\dfrac{1}{v} + \displaystyle\sum_{i=1}^{N}\frac{1}{\sigma_i^2}}$$

$$\sigma^2_{\text{post}} = \frac{1}{\dfrac{1}{v} + \displaystyle\sum_{i=1}^{N}\frac{1}{\sigma_i^2}}$$

F) Suppose that $X \sim N(0, \sigma^2)$, and that $1/\sigma^2 \sim \text{Gamma}(a, b)$. Let us find $f(x)$, the marginal distribution of x. As above, let $\omega = 1/\sigma^2$. Then:

$$
\begin{aligned}
f(x) &= \int_0^\infty p(x|\omega)\, p(\omega)\, d\omega \\
&\propto \int_0^\infty \omega^{1/2} e^{-\omega x^2/2} \omega^{a-1} e^{-bw}\, d\omega \\
&= \int_0^\infty \omega^{a-1/2} \exp\bigl(-(b + x^2/2)\omega\bigr)\, d\omega \qquad \text{(the integral of the kernel of a Gamma)} \\
&\propto \frac{\Gamma(a + 1/2)}{\bigl(b + x^2/2\bigr)^{a+1/2}} \\
&\propto \frac{1}{\left(1 + \dfrac{x^2}{2b}\right)^{a+1/2}} \\
&\propto \frac{1}{\left(1 + \dfrac{x^2}{(b/a)2a}\right)^{(2a+1)/2}}
\end{aligned}
$$

This is a student-t distribution with $\nu = 2a$ and $\sigma^2 = b/a$.

With the alternative parameterization $x|\omega \sim N(\mu, \omega^{-1})$ and $\omega \sim \text{IG}(d/2, d\,\tau^2/2)$, we get a center μ, scale τ, and degrees of freedom d.

The multivariate normal distribution

Basics

A)

$$\begin{aligned}
\operatorname{cov}(x) &:= E\big((x-\mu)(x-\mu)^T\big) \\
&= E\big((x-\mu)(x^T-\mu^T)\big) \\
&= E\big(xx^T - \mu x^T - x\mu^T + \mu\mu^T\big) \\
&= E\big(xx^T\big) - E\big(\mu x^T\big) - E\big(x\mu^T\big) + E\big(\mu\mu^T\big) \\
&= E\big(xx^T\big) - \mu E\big(x^T\big) - E\big(x\big)\mu^T + \mu\mu^T E(I_N) \\
&= E\big(xx^T\big) - \mu\mu^T - \mu\mu^T + \mu\mu^T \\
&= E\big(xx^T\big) - \mu\mu^T
\end{aligned}$$

$$\begin{aligned}
\operatorname{cov}(Ax+b) &:= E\Big(\big(Ax+b-E(Ax+b)\big)\big(Ax+b-E(Ax+b)\big)^T\Big) \\
&= E\Big(\big(Ax-E(Ax)\big)\big(Ax-E(Ax)\big)^T\Big) \\
&= E\Big(A(x-\mu)\big(A(x-\mu)\big)^T\Big) \\
&= E\Big(A(x-\mu)(x-\mu)^T A^T\Big) \\
&= AE\Big((x-\mu)(x-\mu)^T\Big)A^T \\
&= A\operatorname{cov}(x)A^T
\end{aligned}$$

D) Suppose that z has a standard multivariate normal distribution, let L be a $p \times p$ matrix of full rank, and μ a vector. Let $x := Lz + \mu$. Then for any vector a,

$$a^T x = a^T Lz + a^T \mu = (L^T a)^T z + a^T \mu.$$

By assumption, $(L^T a)^T z$ has a normal distribution, and therefore so does $(L^T a)^T z + a^T$. Thus, x has a multivariate normal distribution. By part (A), we know that $\operatorname{cov}(x) = L\operatorname{cov}(z)L^T$.

E) Suppose that $a^T x$ is normal for all vectors a. Let $\alpha_i, \mu_i \in \mathbb{R}$ be such that $a_i^T x - \mu_i$ is standard normal, where $a_i := \langle 0, \ldots, \alpha_i, \ldots, 0\rangle$. (The α_i term is in the i^{th} position.) Let

$$L^{-1} := \begin{bmatrix} a_1^T \\ \vdots \\ a_n^T \end{bmatrix}$$

Then $L^{-1}x - \mu^* = z$, so $x = Lz + L\mu^*$.

[clean up, define μ^*]

F) The PDF of the standard multivariate normal distribution is

$$f_Z(z) = (2\pi)^{-p/2} \exp\bigl(-z^T z/2\bigr).$$

Suppose that x is multivariate normal. Then $x = Lz + \mu$ for some invertible matrix L, and so $z = L^{-1}(x-\mu)$. By the change of variables formula,

$$\begin{aligned} f_X(x) &= f_Z\bigl(L^{-1}(x-\mu)\bigr) \det(J) \\ &= (2\pi)^{-p/2} \exp\Bigl(-(L^{-1}(x-\mu))^T L^{-1}(x-\mu)/2\Bigr) \det(L^{-1}) \\ &= (2\pi)^{-p/2} \det(L^{-1}) \exp\Bigl(-(x-\mu)^T \bigl((L^{-1})^T L^{-1}\bigr)(x-\mu)/2\Bigr) \end{aligned}$$

which has the required form when

$$C = (2\pi)^{-p/2} \det(L^{-1}) \qquad \text{and}$$

$$Q(x-\mu) = -(x-\mu)^T \bigl((L^{-1})^T L^{-1}\bigr)(x-\mu).$$

G) Let $x_1 \sim N(\mu_1, \Sigma_1)$ and $x_2 \sim N(\mu_2, \Sigma_2)$ be independent. Let A and B be matrices with full column rank such that $y = Ax_1 + Bx_2$. Then the MGF of y is the product of two known MGFs:

$$\begin{aligned} M_y(t) &:= E\bigl(e^{t^T y}\bigr) \\ &= E\bigl(e^{t^T(Ax_1+Bx_2)}\bigr) \\ &= E\bigl(e^{t^T Ax_1} e^{t^T Bx_2}\bigr) \\ &= E\bigl(e^{t^T Ax_1}\bigr) E\bigl(e^{t^T Bx_2}\bigr) \\ &= E\bigl(e^{(A^T t)^T x_1}\bigr) E\bigl(e^{(B^T t)^T x_2}\bigr) \\ &= \exp\Bigl((A^T t)^T \mu_1 + (A^T t)^T \Sigma_1 (A^T t)/2\Bigr) \exp\Bigl((B^T t)^T \mu_2 + (B^T t)^T \Sigma_2 (B^T t)/2\Bigr) \\ &= \exp\Bigl((A^T t)^T \mu_1 + (B^T t)^T \mu_2 + (A^T t)^T \Sigma_1 (A^T t)/2 + (B^T t)^T \Sigma_2 (B^T t)/2\Bigr) \\ &= \exp\Bigl(t^T A\mu_1 + t^T B\mu_2 + t^T A\Sigma_1 A^T t/2 + t^T B\Sigma_2 B^T t/2\Bigr) \\ &= \exp\Bigl(t^T(A\mu_1 + B\mu_2) + t^T(A\Sigma_1 A^T + B\Sigma_2 B^T)t/2\Bigr) \end{aligned}$$

This is the MGF of a multivariate normal distribution; that is, $y \sim N(A\mu_1 + B\mu_2, A\Sigma_1 A^T + B\Sigma_2 B^T)$.

Conditionals and marginals

A) Let x, x_1, μ, and Σ be as given. Let A denote the $k \times p$ matrix given by the partition $\left(I_{k\times k}\ 0_{k\times(p-k)}\right)$. Then $x_1 = Ax$, and we can apply the results in (G) above to get

$$\begin{aligned} x_1 &\sim N(A\mu, A\Sigma A^T) \\ &\sim N(\mu_1, \Sigma_{11}) \end{aligned}$$

B) Following Gentle's *Matrix Algebra* text, we get

$$\begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12}^T & \Omega_{22} \end{pmatrix} = \begin{pmatrix} \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}Z^{-1}\Sigma_{21}\Sigma_{11}^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}Z^{-1} \\ -Z^{-1}\Sigma_{21}\Sigma_{11}^{-1} & Z^{-1} \end{pmatrix}$$

where $Z = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. That is,

$$\Omega_{11} = \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}Z^{-1}\Sigma_{21}\Sigma_{11}^{-1}$$

$$\Omega_{12} = -\Sigma_{11}^{-1}\Sigma_{12}Z^{-1}$$

$$\Omega_{21} = -Z^{-1}\Sigma_{21}\Sigma_{11}^{-1}$$

$$\Omega_{22} = Z^{-1}$$

Multiple regression: three classical principles for inference

A) Let us first find the least squares estimate. We can write the quantity that we wish to minimize as

$$f(\beta) = (y - X\beta)^T (y - X\beta),$$

and we can differentiate to get

$$\frac{\partial f}{\partial \beta} = -2X^T (y - X\beta).$$

Assuming that $X^T X$ is invertible, we can ignore the -2, set this equal to zero, and solve for β:

$$\begin{aligned} 0 &= X^T (y - X\beta) \\ X^T y &= X^T X\beta \\ \beta &= (X^T X)^{-1} X^T y \end{aligned}$$

Thus, the least squares estimate for β is $\hat{\beta} := (X^T X)^{-1} X^T y$.

Let us now show that the maximum likelihood estimate is the same. Given our assumption about ε_i, note that $p(y_i|\beta, \sigma^2)$ is normal. In matrix notation (and unnormalized), we want to maximize

$$\exp\Big((y - X\beta)^T \Sigma^{-1} (y - X\beta)/2\Big)$$

where Σ is the covariance matrix—in this case, the diagonal matrix whose entries are σ^2. We can maximize this by maximizing the logarithm of it instead, which gives us a function of β:

$$f(\beta) = (y - X\beta)^T \Sigma^{-1} (y - X\beta)/2$$

which we differentiate and set equal to zero:

$$\begin{aligned} 0 &= X^T \big(\Sigma^{-1} + (\Sigma^{-1})^T\big)(y - X\beta) \\ \beta &= (X^T X)^{-1} X^T y \end{aligned}$$

Note that we can disregard the $\Sigma^{-1} + (\Sigma^{-1})^T$ factor, since is it corresponds to multiplication by a constant. We therefore obtain the same estimate as with least squares. Finally, let us find the estimate using the method of moments. Since covariance does not depend on the expected value, we can assume that $E(x_i) = 0$ for all i. We want to have $\text{cov}(\varepsilon_i, x_i) = 0$ for all i, which means

B) Suppose now that we want to find the weighted least squares estimate:

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} w_i (y_i - x_i^T \beta)^2.$$

As above, we can write this in matrix form:

$$\begin{aligned} f(\beta) &= (y - X\beta)^T W (y - X\beta) \\ &= y^T W y - y^T W X\beta - \beta^T X^T W y + \beta^T X^T W X\beta, \end{aligned}$$

where W is the diagonal matrix of weights. Setting the derivative equal to zero, we get

$$\begin{aligned} 0 &= -y^T W X - X^T W y + 2X^T W X\beta \\ &= -2X^T W y + 2X^T W X\beta \\ X^T W y &= X^T W X\beta \\ \beta &= (X^T W X)^{-1} X^T W y, \end{aligned}$$

assuming that $X^T W X$ is invertible.

Now we must show that this is the same estimate as obtained using maximum likelihood under heteroscedastic Gaussian error. As in part (A), we want to maximize

$$f(\beta) = (y - X\beta)^T \Sigma^{-1} (y - X\beta)/2$$

where Σ is the covariance matrix, which is now the diagonal matrix such that $(\Sigma)_{ii} = \sigma_i^2$. We differentiate this and set it equal to zero and solve for β to get

$$\begin{aligned}
0 &= X^T \left(\Sigma^{-1} + (\Sigma^{-1})^T\right)(y - X\beta) \\
&= X^T \left(2\Sigma^{-1}\right)(y - X\beta) \\
&= X^T \Sigma^{-1} (y - X\beta) \\
&= X^T \Sigma^{-1} (y - X\beta) \\
X^T \Sigma^{-1} X\beta &= X^T \Sigma^{-1} y \\
\beta &= \left(X^T \Sigma^{-1} X\right)^{-1} X^T \Sigma^{-1} y.
\end{aligned}$$

Thus, we get the same solution as with the weighted least squares where the weights are given by the inverse covariance matrix; that is, the weights are the precisions: $w_i = 1/\sigma_i^2$.

Quantifying uncertainty: some basic frequentist ideas

In linear regression

A) Given $y = X\beta + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2 I)$, we want to find the sampling distribution of $\hat{\beta}$. Since β is a linear combination of multivariate normals, we know that its sampling distribution is also multivariate normal. Therefore, we need only find the mean and the covariance matrix. Note that $y \sim N(X\beta, \sigma^2 I)$, and thus, since $\hat{\beta} = (X^T X)^{-1} X^T y$, the mean of $\hat{\beta}$ is

$$(X^T X)^{-1} X^T X\beta = \beta.$$

Now we find the covariance matrix:

$$\begin{aligned}
\operatorname{cov}(\hat{\beta}) &= \operatorname{cov}\big((X^T X)^{-1} X^T y\big) \\
&= (X^T X)^{-1} X^T \operatorname{cov}(y) \big((X^T X)^{-1} X^T\big)^T \\
&= (X^T X)^{-1} X^T \sigma^2 I \big((X^T X)^{-1} X^T\big)^T \\
&= \sigma^2 (X^T X)^{-1} X^T X \big((X^T X)^{-1}\big)^T \\
&= \sigma^2 \big((X^T X)^{-1}\big)^T \\
&= \sigma^2 (X^T X)^{-1}
\end{aligned}$$

Note that in the third line, we have used the fact proven above that $\operatorname{cov}(Ax) = A\operatorname{cov}(x)A^T$. Therefore,

$$\hat{\beta} \sim N\big(\beta, (X^T X)^{-1}\big).$$