A. Suppose that we take independent observations $x_1, ..., x_N$ from a Bernoulli sampling model with unknown probability w. That is, the x_i are the results of flipping a coin with unknown bias. Suppose that w is given a Beta(a,b) prior distribution. Derive the posterior distribution $p(w|x_1, ..., x_N)$.

We would like to obtain the posterior distribution for w:

$$\begin{aligned} p(w|x_1, ..., x_N) &\propto \left(\prod_{i=1}^{N} p(x_i|w) \right) p(w) \\ &\propto \left(\prod_{i=1}^{N} w^{x_i}(1-w)^{1-x_i} \right) w^{a-1}(1-w)^{b-1} \\ &\propto \underbrace{w^{\sum_{i=1}^{N} x_i + a - 1}(1-w)^{N - \sum_{i=1}^{N} x_i + b - 1}}_{\text{kernel of Beta}\left(\sum_{i=1}^{N} x_i + a, N - \sum_{i=1}^{N} x_i + b\right)} \end{aligned}$$

Therefore, the posterior distribution for w is

$$p(x_1, ..., x_N) = \text{Beta}\left(\sum_{i=1}^{N} x_i + a, N - \sum_{i=1}^{N} x_i + b \right)$$

B. Suppose that $x_1 \sim Ga(a_1, 1)$ and $x_2 \sim Ga(a_2, 1))$. Define two new random variables $y_1 = \frac{x_1}{x_1+x_2}$ and $y_2 = x_1 + x_2$. Find the joint distribution for (y_1, y_2) using a direct PDF transformation. Use this to characterize the marginals $p(y_1)$ and $p(y_2)$, and propose a method that exploits this result to simulate beta random variables, assuming you have a source of gamma random variables.

We know that

$$\begin{aligned} y_1 &= \frac{x_1}{x_1 + x_2} \\ y_2 &= x_1 + x_2 \end{aligned}$$

implies that

$$\begin{aligned} x_1 &= y_1 y_2 \\ x_2 &= y_2(1 - y_1) \\ J &= \left| \begin{bmatrix} y_2 & y_1 \\ -y_2 & 1 - y_1 \end{bmatrix} \right| \\ &= y_2 - y_1 y_2 + y_1 y_2 = y_2 \end{aligned}$$

Now, we can easily obtain the joint distribution of (y_1, y_2), using the independence of x_1 and x_2:

$$\begin{aligned} f_{y_1,y_2}(y_1, y_2) &= f_{x_1,x_2}(y_1 y_2, y_2(1 - y_1))|J| \\ &= f_{x_1}(y_1 y_2) f_{x_2}(y_2(1 - y_1)) y_2 \\ &\propto e^{-y_1 y_2}(y_1 y_2)^{a_1 - 1} e^{-y_2 + y_1 y_2}(y_2(1 - y_1))^{a_2 - 1} y_2 \\ &\propto \underbrace{y_1^{a_1 - 1}(1 - y_1)^{a_2 - 1}}_{\text{kernel of Beta}(a_1, a_2)} \quad \underbrace{y_2^{a_1 + a_2 - 1} e^{-y_2}}_{\text{kernel of Ga}(a_1 + a_2, 1)} \end{aligned}$$

Therefore, the joint distribution is

$$p(y_1, y_2) = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} y_1^{a_1 - 1}(1 - y_1)^{a_2 - 1} \cdot \frac{1}{\Gamma(a_1 + a_2)} e^{-y_2} y_2^{a_1 + a_2 - 1}$$

By definition, $y_1 \perp y_2$ (c.f. Casella & Berger; definition 4.2.5). Thus, we can easily obtain the following marginals:

$$\begin{aligned} p(y_1) &= \text{Beta}(a_1, a_2) \\ p(y_2) &= \text{Gamma}(a_1 + a_2, 1) \end{aligned}$$

Proposed Method

Assume we have gamma random variables, and we want to simulate beta random variables. So long as the rate parameter for the gamma random variables

is 1, we can take a value less than the simulated gamma random variables, divide it by the simulated gamma variables, and then we would obtain a beta random variable with shape parameters summing to the gamma's shape parameter.

C. Suppose that we take independent observations $x_1, ..., x_N$ from a normal sampling model with unknown mean θ and known variance σ^2 : $x_i \sim N(\theta, \sigma^2)$. Suppose that θ is given a normal prior distribution with mean m and variance v. Derive the posterior distribution $p(\theta|x_1, ..., x_N)$.

We want to obtain the posterior distribution for θ:

$$\begin{aligned}
p(\theta|x_1, ..., x_N) &\propto \left(\prod_{i=1}^{N} p(x_i|\theta)\right) p(\theta) \\
&\propto \left(\prod_{i=1}^{N} \exp\left(-\frac{1}{2\sigma^2}(x_i-\theta)^2\right)\right) \cdot \exp\left(-\frac{1}{2v}(\theta-m)^2\right) \\
&\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\theta)^2 - \frac{1}{2v}(\theta-m)^2\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\frac{\sum_{i=1}^{N} x_i^2 - 2\theta\sum_{i=1}^{N} x_i + N\theta^2}{\sigma^2} + \frac{\theta^2 - 2m\theta}{v}\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\left(-2\left(\underbrace{\frac{\sum_{i=1}^{N} x_i}{\sigma^2} + \frac{m}{v}}_{b^T}\right)\theta + \left(\underbrace{\frac{N}{\sigma^2} + \frac{1}{v}}_{A}\right)\theta^2\right)\right]
\end{aligned}$$

Therefore, we see that the posterior for θ is

$$p(\theta|x_1, ..., x_N) = \mathrm{N}\left(\left(\frac{N}{\sigma^2} + \frac{1}{v}\right)^{-1}\left(\frac{\sum_{i=1}^{N} x_i}{\sigma^2} + \frac{m}{v}\right), \left(\frac{N}{\sigma^2} + \frac{1}{v}\right)^{-1}\right)$$

D. Suppose that we take independent observations x_i from a normal sampling model with known mean θ but unknown variance σ^2. Suppose that precision parameter ω has a gamma prior with parameters a and b, implying that σ^2 has what is called an inverse-gamma prior. Derive the posterior distribution for ω. Re-express this as a posterior for σ^2, the variance.

First, we derive the posterior distribution for ω:

$$\begin{aligned} p(\omega|x_1,...,x_N) &\propto \left(\prod_{i=1}^{N} p(x_i|\omega)\right) p(\omega) \\ &\propto \prod_{i=1}^{N} \left(\frac{\omega}{2\pi}\right)^{1/2} \exp\left[-\frac{\omega}{2}(x_i-\theta)^2\right] \cdot e^{-b\omega}\omega^{a-1} \\ &\propto \underbrace{\omega^{N/2+a-1}\exp\left[-\omega\left(\frac{\sum_{i=1}^{N}(x_i-\theta)^2}{2}+b\right)\right]}_{\text{kernel of Ga}\left(\frac{N}{2}+a,\frac{\sum_{i=1}^{N}(x_i-\theta)^2}{2}+b\right)} \end{aligned}$$

Therefore, the posterior distributions for ω and σ^2 are

$$p(\omega|x_1,...,x_N) = \text{Ga}\left(\frac{N}{2}+a, \frac{\sum_{i=1}^{N}(x_i-\theta)^2}{2}+b\right)$$

$$p(\sigma^2|x_1,...,x_N) = \text{Inv-Ga}\left(\frac{N}{2}+a, \frac{\sum_{i=1}^{N}(x_i-\theta)^2}{2}+b\right)$$

E. Suppose that, as above, we take independent observations x_i from a normal sampling model with unknown, common mean θ. This time, however, each observation has its own idiosyncratic (but known) variance: $x_i \sim N(\theta, \sigma_i^2)$. Suppose that θ is given a normal prior distribution with mean m and variance v. Derive the posterior distribution for θ. Express the posterior mean in a form that is clearly interpretable as a weighted average of the observations and the prior mean.

We want to find the posterior distribution for θ:

$$\begin{aligned}
p(\theta|x_1,...,x_N) &\propto \left(\prod_{i=1}^{N} p(x_i|\theta)\right) p(\theta) \\
&\propto \left(\prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma_i^2}(x_i-\theta)^2\right]\right) \cdot \exp\left[-\frac{1}{2v}(\theta-m)^2\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\sum_{i=1}^{N}\frac{(x_i-\theta)^2}{\sigma_i^2} + \frac{(\theta-m)^2}{v}\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\sum_{i=1}^{N}\frac{x_i^2}{\sigma_i^2} - 2\theta\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2} + \theta^2\sum_{i=1}^{N}\frac{1}{\sigma_i^2} + \frac{\theta^2-2\theta m}{v}\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\left(-2\left(\underbrace{\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}+\frac{m}{v}}_{b^T}\right)\theta + \theta^2\left(\underbrace{\sum_{i=1}^{N}\frac{1}{\sigma_i^2}+\frac{1}{v}}_{A}\right)\right)\right]
\end{aligned}$$

Therefore, the posterior distribution for θ is

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

Note that we can express the posterior mean as

$$\left(\sum_{i=1}^{N}\frac{1}{\sigma_i^2}+\frac{1}{v}\right)^{-1}\left(\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}+\frac{m}{v}\right) = \frac{\sum_{i=1}^{N}\left(\frac{1}{\sigma_i^2}\right)x_i + \frac{1}{v}m}{\sum_{i=1}^{N}\frac{1}{\sigma_i^2}+\frac{1}{v}},$$

which is a weighted average of the observations x_i and the prior mean m with respective weights $\frac{1}{\sigma_i^2}$ and $\frac{1}{v}$.

F. Suppose that $(x|\omega) \sim \mathbf{N}(m, \omega^{-1})$, and that $\omega \sim \mathbf{Ga}(a/2, b/2)$ prior. Show that the marginal distribution of x is Student's t with a degrees of freedom, center m, and scale parameter $(b/a)^{1/2}$. This is why the t distribution is often referred to as a scale mixture of normals.

We want to obtain the marginal data distribution:

$$\begin{aligned}
p(x) &= \int_\Omega p(x, \omega) d\omega \\
&= \int_\Omega p(x|\omega) p(\omega) d\omega \\
&= \frac{1}{\sqrt{2\pi}} \frac{(b/2)^{a/2}}{\Gamma(a/2)} \int_\Omega \underbrace{\omega^{1/2+a/2-1} \exp\left[-\omega\left(\frac{1}{2}(x-m)^2 + \frac{b}{2}\right)\right]}_{\text{kernel of Ga}\left(\frac{a+1}{2}, \frac{1}{2}(x-m)^2 + \frac{b}{2}\right)} d\omega \\
&= \frac{1}{\sqrt{2\pi}} \frac{(b/2)^{a/2}}{\Gamma(a/2)} \cdot \frac{\Gamma((a+1)/2)}{(\frac{1}{2}(x-m)^2 + b/2)^{\frac{a+1}{2}}} \\
&= \frac{\Gamma((a+1)/2)}{\Gamma(a/2)\sqrt{2\pi}} \left(\frac{b}{2}\right)^{a/2-(a+1)/2} \left(\left[\frac{1}{2}(x-m)^2 + \frac{b}{2}\right]\frac{2}{b}\right)^{-\frac{a+1}{2}} \\
&= \frac{\Gamma((a+1)/2)}{\Gamma(a/2)\sqrt{b\pi}} \left(\frac{1}{b}(x-m)^2 + 1\right)^{-\frac{a+1}{2}},
\end{aligned}$$

which uniquely characterizes this distribution as Student's t distribution with a degrees of freedom, center m, and scale parameter $(b/a)^{1/2}$.

A. The covariance matrix $\mathbf{cov}(x)$ of a vector-valued random variable x is defined as the matrix whose (i, j) entry is the covariance between x_i and x_j. In matrix notation, $\mathbf{cov}(x) = E[(x-\mu)(x-\mu)^T]$, where μ is the mean vector whose ith component is $E(x_i)$. Prove the following: (1) $\mathbf{cov}(x) = E(xx^T) - \mu\mu^T$; and (2) $\mathbf{cov}(Ax+b) = A\mathbf{cov}(x)A^T$ for matrix A and vector b.

First, let's obtain $\text{cov}(x)$:

$$\begin{aligned}
\text{cov}(x) &= E[(x-\mu)(x-\mu)^T] \\
&= E[(x-\mu)(x^T-\mu^T)] \\
&= E[xx^T - x\mu^T - \mu x^T + \mu\mu^T] \\
&= E[xx^T] - E[x\mu^T] - E[\mu x^T] + E[\mu\mu^T] \\
&= E[xx^T] - E[x]\mu^T - \mu E[x^T] + \mu\mu^T \\
&= E[xx^T] - \mu\mu^T - \mu\mu^T + \mu\mu^T \\
&= E[xx^T] - \mu\mu^T
\end{aligned}$$

Now, let's find the covariance of $Ax+b$:

$$\begin{aligned}
\text{cov}(Ax+b) &= E[(Ax+b-E[Ax+b])(Ax+b-E[Ax+b])^T] \\
&= E[(Ax+b-A\mu-b)(Ax+b-A\mu-b)^T] \\
&= E[(Ax-A\mu)(Ax-A\mu)^T] \\
&= E[A(x-\mu)(x^TA^T-\mu^TA^T)] \\
&= AE[(x-\mu)(x^T-\mu^T)A^T] \\
&= AE[(x-\mu)(x-\mu)^T]A^T \\
&= A\text{cov}(x)A^T
\end{aligned}$$

B. Consider the random vector $z = (z_1, ..., z_p)^T$, with each entry having an independent standard normal distribution. Derive the PDF and MGF of z, expressed in vector notation. We say that z has a standard multivariate normal distribution.

We know that

$$p(z_i) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z_i^2}{2}\right], \ \forall i \in \{1, ..., p\}.$$

Since z_i are independent, we can get the PDF of z in the following way:

$$\begin{aligned} p(z_1, ..., z_p) &= \prod_{i=1}^{p} p(z_i) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^p \exp\left[-\frac{z_1^2}{2} - \frac{z_2^2}{2} - ... - \frac{z_p^2}{2}\right] \\ &= (2\pi)^{-p/2} \exp\left[-\frac{1}{2}(z_1^2 + z_2^2 + ... + z_p^2)\right] \\ &= (2\pi)^{-p/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{p} z_i^2\right] \end{aligned}$$

In vector notation, we can state the pdf of z as

$$p(z) = (2\pi)^{-p/2} \exp\left[-\frac{z^T z}{2}\right], \ z_i \in \mathbb{R} \ \forall i = 1, ..., p$$

We can obtain z's MGF in the following way:

$$\begin{aligned} M_z(t) &= E(e^{t^T z}) \\ &= E\left[\exp\left(t_1 z_1 + t_2 z_2 + ... + t_p z_p\right)\right] \\ &= E\left[e^{t_1 z_1} \cdot ... \cdot e^{t_p z_p}\right] \end{aligned}$$

By the independence of z_i, we then get

$$\begin{aligned} M_z(t) &= E[e^{t_1 z_1}] \cdot ... \cdot E[e^{t_p z_p}] \\ &= \prod_{i=1}^{p} E(e^{t_i z_i}) = \prod_{i=1}^{p} M_{z_i}(t) \\ &= \prod_{i=1}^{p} e^{t_i^2/2} \\ &= \exp\left[\sum_{i=1}^{p} \frac{t_i^2}{2}\right] \\ &= \exp\left[\frac{t^T t}{2}\right] \end{aligned}$$

C. A vector-valued random variable $x = (x_1, ..., x_p)^T$ has a multivariate normal distribution iff every linear combination of its components is univariate normal. That is, for all vectors a not identically zero, the scalar quantity $z = a^T x$ is normally distributed. From this definition, prove that x is multivariate normal, written $x \sim \mathbf{N}(\mu, \Sigma)$, iff its MGF is of the form $E[e^{t^T x}] = \mathbf{exp}\left[t^T\mu + t^T\Sigma t/2\right]$.

($\Rightarrow$) Let $x \sim \mathrm{N}(\mu, \Sigma)$. By linear combination of normals, $E(z) = a^T\mu$ and $\mathrm{var}(z) = a^T\Sigma a$. We know that $z = a^T x$ has the following mgf

$$\begin{aligned} M_z(\tau) &= E(e^{\tau z}) \\ &= \exp\left[E(z)\tau + \frac{1}{2}\mathrm{var}(z)\tau^2\right] \\ &= \exp\left[a^T\mu\tau + \frac{1}{2}a^T\Sigma a\tau^2\right] \\ &= \exp\left[\tau a^T\mu + \frac{1}{2}(a\tau^T)^T\Sigma a\tau^T\right], \end{aligned}$$

Now, denote $t = a\tau^T$. Then,

$$\begin{aligned} M_x(t) &= E\left[e^{t^T x}\right] \\ &= E\left[e^{\tau a^T x}\right] \\ &= E\left[e^{\tau z}\right] \\ &= M_z(\tau) \\ &= \exp\left[\tau a^T\mu + \frac{1}{2}(a\tau^T)^T\Sigma a\tau^T\right] \\ &= \exp\left[t^T\mu + \frac{1}{2}t^T\Sigma t\right] \end{aligned}$$

($\Leftarrow$) Let $M_x(t) = \exp\left[t^T\mu + \frac{t^T\Sigma t}{2}\right]$. Then,

$$\begin{aligned} M_z(t) &= E[e^{t^T z}] \\ &= E(e^{t^T a^T x}) \\ &= M_x(at) \\ &= \exp\left[(at)^T\mu + \frac{1}{2}(at)^T\Sigma(at)\right] \\ &= \exp\left[t^T a^T\mu + \frac{1}{2}t^T a^T\Sigma at\right] \end{aligned}$$

By uniqueness of MGFs,

$$z \sim \mathrm{N}(a^T\mu, a^T\Sigma a).$$

Therefore, any linear combination of x's components is univariate normal. By the given definition, x is multivariate normal.

D. Another basic theorem is that a random vector is multivariate normal iff it is an affine transformation of independent univariate normals. You will first prove the "if" statement. Let z have a standard multivariate normal distribution, and define the random vector $x = Lz + \mu$ for some $p \times p$ matrix L of full-column rank. Prove that x is multivariate normal. In addition, use the moment identities you proved above to compute the expected value and covariance matrix of x.

Let $z \sim \mathrm{N}(0, I_p)$, where $x = Lz + \mu$. We know that $M_z(t) = \exp\left[\frac{t^T t}{2}\right]$. Let's use MGFs to characterize x:

$$\begin{aligned} M_x(t) &= E[e^{t^T x}] \\ &= E[e^{t^T Lz + t^T \mu}] \\ &= E[e^{t^T Lz} e^{t^T \mu}] \\ &= E[e^{t^T Lz}] \cdot E[e^{t^T \mu}] \\ &= e^{t^T \mu} M_z((t^T L)^T) \\ &= \exp\left[t^T \mu + \frac{1}{2}(t^T L)(t^T L)^T\right] \\ &= \exp\left[t^T \mu + \frac{1}{2}(t^T L L^T t)\right], \end{aligned}$$

where the penultimate equality follows from the MGF for z. Note that this MGF uniquely characterizes x, so

$$x \sim \mathrm{N}_p(\mu, LL^T),$$

with expected value μ and covariance matrix LL^T.

E. Now for the "only if." Suppose that x has a multivariate normal distribution. Prove that x can be written as an affine transformation of standard normal random variables. Use this insight to propose an algorithm for simulating multivariate normal random variables with a specified mean and covariance matrix.

Suppose $x \sim N_p(\mu, \Sigma)$. Since Σ is positive semi-definite, we can use spectral decomposition to obtain

$$\Sigma = P\Lambda P^T,$$

where $\Lambda = \text{diag}(\lambda_1, ..., \lambda_p)$ and P is an orthogonal matrix. Consider the affine transformation $x^* = Lz + \mu$, where z is a vector of standard normal random variables, and $L = P\Lambda^{1/2}$. Note that $r(L) = r(P\Lambda^{1/2}) = p$. From the previous problem, we know that

$$\begin{aligned} x^* &\sim \text{N}_p\left(\mu, P\Lambda^{1/2}(P\Lambda^{1/2})^T\right) \\ &\sim \text{N}_p(\mu, P\Lambda P^T) \\ &\sim \text{N}_p(\mu, \Sigma) \end{aligned}$$

Since mean and variance characterize normal distributions, $x^* = x$. Thus, we showed that we can construct a multivariate normal distribution as an affine transformation of standard normal random variables.

F. Use the previous result and the PDF of a standard multivariate normal to show that the PDF of a multivariate normal $x \sim \mathbf{N}(\mu, \Sigma)$ takes the form $p(x) = C\mathbf{exp}[-Q(x-\mu)/2]$ for some constant C and quadratic form $Q(x-\mu)$.

From the previous problem, we know how to transform a multivariate normal distribution to standard normal variables:

$$x = Lz + \mu \Leftrightarrow z = L^{-1}(x-\mu),$$

where $L = P\Lambda^{1/2}$ and $L^{-1} = (P\Lambda^{1/2})^{-1} = \Lambda^{-1/2}P^T$. Then, $z = \Lambda^{-1/2}P^T(x-\mu)$. We can obtain the PDF for x using our handy-dandy transformation method:

$$\begin{aligned}
f_x(x) &= f_z(\Lambda^{-1/2}P^T(x-\mu))|J| \\
&= \left|\Lambda^{-1/2}P^T\right|(2\pi)^{-p/2}\text{exp}\left[-\frac{1}{2}(x-\mu)^T P\Lambda^{-1/2}\Lambda^{-1/2}P^T(x-\mu)\right] \\
&= \left|\Lambda^{-1/2}\right| \cdot \left|P^T\right|(2\pi)^{-p/2}\text{exp}\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right] \\
&= |\Lambda|^{-1/2}(2\pi)^{-p/2}\text{exp}\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right] \\
&= |\Sigma|^{-1/2}(2\pi)^{-p/2}\text{exp}\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right],
\end{aligned}$$

since $|\Sigma| = |P\Lambda P^T| = |P||\Lambda||P| = |\Lambda|$. Clearly, the above is the desired PDF with

$$\begin{aligned}
C &= (2\pi)^{-p/2}|\Sigma|^{-1/2} \\
Q(x-\mu) &= (x-\mu)^T\Sigma^{-1}(x-\mu)
\end{aligned}$$

G. Let $x_1 \sim \mathbf{N}(\mu_1, \Sigma_1)$ and $x_2 \sim \mathbf{N}(\mu_2, \Sigma_2)$, where x_1 and x_2 are independent of each other. Let $y = Ax_1 + Bx_2$ for matrices A, B of full column rank and appropriate dimension. Note that x_1 and x_2 need not have the same dimension, as long as Ax_1 and Bx_2 do. Use your previous results to characterize the distribution of y.

Since there's a good chance that y will be a multivariate normal distribution, we can attempt to characterize the distribution of y with MGFs. Since $\forall a \in \mathbb{R}^p / \{0\} \implies a^T Ax_1$ and $a^T Bx_2$ are univariate normal random variables, then both Ax_1 and Bx_2 are multivariate normal. That is,

$$M_{Ax_1}(t) = \exp\left[t^T A\mu_1 + \frac{1}{2}t^T A\Sigma_1 A^T t\right]$$
$$M_{Bx_2}(t) = \exp\left[t^T B\mu_2 + \frac{1}{2}t^T B\Sigma_2 B^T t\right]$$

By independence,

$$\begin{aligned} M_{Ax_1+Bx_2}(t) &= M_{Ax_1}(t) \cdot M_{Bx_2}(t) \\ &= \exp\left[t^T(A\mu_1 + B\mu_2) + \frac{1}{2}t^T(A\Sigma_1 A^T + B\Sigma_2 B^T)t\right] \end{aligned}$$

By uniqueness of MGFs, we see that

$$y = Ax_1 + Bx_2 \sim \mathrm{N}\left(A\mu_1 + B\mu_2, A\Sigma_1 A^T + B\Sigma_2 B^T\right)$$

A. Derive the marginal distribution of x_1.

Recall our affine transformation

$$x = Lz + \mu.$$

We can write this as

$$\begin{aligned}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &= \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \\ &= \begin{bmatrix} L_{11}z_1 + L_{12}z_2 + \mu_1 \\ L_{21}z_1 + L_{22}z_2 + \mu_2, \end{bmatrix}\end{aligned}$$

where $L = \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix}$ and z_1, z_2 are independent standard multivariate normal random variables. So, $x_1 = L_{11}z_1 + L_{12}z_2 + \mu_1$. We can obtain its marginal distribution with MGFS, since we have an inkling that it will be a known – most likely Normal– distribution:

$$\begin{aligned} M_{x_1}(t) &= E[e^{t^T x_1}] \\ &= E\left[\exp\left(t^T(L_{11}z_1 + L_{12}z_2 + \mu_1)\right)\right] \\ &= E\left[\exp\left(t^T L_{11}z_1\right)\exp\left(t^T L_{12}z_2\right)\exp\left(t^T\mu_1\right)\right] \\ &= E\left[\exp\left(t^T L_{11}z_1\right)\right] \cdot E\left[\exp\left(t^T L_{12}z_2\right)\right] e^{t^T\mu_1} && \text{(by ind.)} \\ &= M_{z_1}(t^T L_{11}) M_{z_2}(t^T L_{12}) e^{t^T\mu_1} && \text{(def. of MGF)} \\ &= \exp\left[\frac{1}{2}t^T L_{11}L_{11}^T t\right]\exp\left[\frac{1}{2}t^T L_{12}L_{12}^T t\right] e^{t^T\mu_1} && \text{(MGF of standard MVN)} \\ &= \exp\left[t^T\mu_1 + \frac{1}{2}t^T\left(L_{11}L_{11}^T + L_{12}L_{12}^T\right)t\right], \end{aligned}$$

which uniquely characterizes x_1 as a multivariate normal distribution with mean μ_1 and variance $L_{11}L_{11}^T + L_{12}L_{12}^T$. Note that

$$\begin{aligned}\Sigma &= LL^T \\ &= \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix}\begin{bmatrix} L_{11}^T & L_{21}^T \\ L_{12}^T & L_{22}^T \end{bmatrix} \\ &= \begin{bmatrix} L_{11}L_{11}^T + L_{12}L_{12}^T & L_{11}L_{21}^T + L_{12}L_{22}^T \\ L_{21}L_{11}^T + L_{22}L_{12}^T & L_{21}L_{21}^T + L_{22}L_{22}^T \end{bmatrix} \\ &= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\end{aligned}$$

Therefore, we concisely write the marginal distribution of x_1 as

$$x_1 \sim \mathrm{N}(\mu_1, \Sigma_{11})$$

B. Let $\Omega = \Sigma^{-1}$ be the inverse covariance matrix, or precision matrix, of x. Using identities for the inverse of a partitioned matrix, express each block of Ω in terms of blocks of Σ.

We know that

$$\begin{aligned}
\Sigma &= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \\
&= \begin{bmatrix} L_{11}L_{11}^T + L_{12}L_{12}^T & L_{11}L_{21}^T + L_{12}L_{22}^T \\ L_{21}L_{11}^T + L_{22}L_{12}^T & L_{21}L_{21}^T + L_{22}L_{22}^T \end{bmatrix}, \\
\Omega &= \Sigma^{-1} \\
&= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1}.
\end{aligned}$$

Since Σ is symmetric and positive semi-definite by virtue of it being a covariance matrix, the blocks of Σ are invertible, as well as their respective Schur complements. Therefore, we can use one of the well-established identities to obtain the inverse of a block covariance matrix:

$$\begin{aligned}
\Omega &= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \\
&= \begin{bmatrix} \left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}\left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1} \\ -\left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21}\right)^{-1}\Sigma_{21}\Sigma_{11}^{-1} & \left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1} \end{bmatrix}
\end{aligned}$$

C. Derive the conditional distribution for x_1 given x_2 in terms of the partitioned elements of $x, \mu,$ and Σ. There are several keys to inner peace: work with densities on a log scale, ignore constants that don't affect x_1, and remember the cute trick of completing the square from basic algebra. Explain briefly how one may interpret this conditional distribution as a linear regression on x_2, where the regression matrix can be read off the precision matrix.

First, we "find" the joint distribution of $x = (x_1, x_2)^T$, which was already given. Then, we can ignore all factors that don't involve x_1:

$$\begin{aligned} p(x_1|x_2) &\propto p(x_1, x_2) \\ &\propto \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right] \end{aligned}$$

Note that we can rewrite

$$\begin{aligned} &(x-\mu)^T\Sigma^{-1}(x-\mu) \\ &= \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}^T \begin{bmatrix} \left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}\left(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1} \\ -\left(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21}\right)^{-1}\Sigma_{21}\Sigma_{11}^{-1} & \left(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1} \end{bmatrix} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix} \\ &= (x_1-\mu_1)^T\Omega_{11}(x_1-\mu_1) - (x_2-\mu_2)^T\Omega_{21}(x_1-\mu_1) - (x_1-\mu_1)^T\Omega_{12}(x_2-\mu_2) + (x_2-\mu_2)^T\Omega_{22}(x_2-\mu_2) \end{aligned}$$

Now, we only consider terms containing x_1. Thus,

$$\begin{aligned} p(x_1|x_2) &\propto \exp\left[-\frac{1}{2}\left((x_1-\mu_1)^T\Omega_{11}(x_1-\mu_1) - (x_2-\mu_2)^T\Omega_{21}(x_1-\mu_1) - (x_1-\mu_1)^T\Omega_{12}(x_2-\mu_2)\right)\right] \\ &\propto \exp\left[-\frac{1}{2}\left(x_1^T\Omega_{11}x_1 - 2\mu_1^T\Omega_{11}x_1 - x_2^T\Omega_{21}x_1 + \mu_2^T\Omega_{21}x_1 - x_1^T\Omega_{12}x_2 + x_1^T\Omega_{12}\mu_2\right)\right] \\ &\propto \exp\left[-\frac{1}{2}\left(x_1^T\Omega_{11}x_1 - 2\mu_1^T\Omega_{11}x_1 + 2\mu_2^T\Omega_{21}x_1 - 2x_2^T\Omega_{21}x_1\right)\right] \\ &\propto \exp\left[-\frac{1}{2}\left(-2\underbrace{\left(\mu_1^T\Omega_{11} + x_2^T\Omega_{21} - \mu_2^T\Omega_{21}\right)}_{b^T}x_1 + x_1^T\underbrace{\Omega_{11}}_{A}x_1\right)\right] \end{aligned}$$

Therefore, the conditional distribution of x_1 is

$$\begin{aligned} p(x_1|x_2) &= \mathrm{N}(A^{-1}b, A^{-1}), \\ A^{-1} &= \Omega_{11}^{-1} \\ &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}, \\ b &= \Omega_{11}^T\mu_1 + \Omega_{21}^Tx_2 - \Omega_{21}^T\mu_2, \\ A^{-1}b &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2) \end{aligned}$$

Interpretation:

This conditional distribution can be interpreted as a regression model on x_2 because we can write

$$x_1 = \underbrace{\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2}_{\text{intercept}} + \underbrace{\Sigma_{12}\Sigma_{22}^{-1}}_{\text{slope}} \cdot \underbrace{x_2}_{\text{regressor}} + \epsilon,$$

where $\epsilon \sim \mathrm{N}(0, \Omega_{11})$ is the stochastic error based on the precision matrix.

A. Show that MLE, LSE, and MoM all lead to the same estimator under SLR. What is the variance of this estimator under the assumption that each ϵ_i is independent and identically distribution with variance σ^2?

First, we consider the least squares estimator (LSE) for β:

$$\hat{\beta}_{LSE} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \right\}, \tag{1}$$

which we can derive with respect to β in matrix notation as follows:

$$\begin{aligned} \frac{\partial}{\partial \beta} \left[(Y - X\beta)^T (Y - X\beta) \right] &= \frac{\partial}{\partial \beta} \left[Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta^T X^T X \beta \right] \\ &= -2X^T Y + 2X^T X\beta \stackrel{set}{=} 0 \\ \hat{\beta}_{LSE} &= (X^T X)^{-1} X^T Y \end{aligned}$$

Now, we obtain the maximum likelihood estimate (MLE) by – of course – maximizing the likelihood!

$$\begin{aligned} \hat{\beta}_{MLE} &= \arg\max_{\beta \in \mathbb{R}^p} \left\{ \prod_{i=1}^{n} p(y_i | \beta, \sigma^2) \right\} \\ &= \arg\max_{\beta \in \mathbb{R}^p} \left\{ \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \right] \right\} \\ &= \arg\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \right\} \quad (= (1)) \\ &= \hat{\beta}_{LSE} \end{aligned}$$

Last but not least: the method of moments (MoM) estimator. Under MoM, we want to choose the β that satisfies $\text{cov}(\epsilon, x_j) = 0$.

$$\begin{aligned} 0 &= \sum_{i=1}^{n} (e_i - \bar{e})(x_{ij} - \bar{x}_j) \\ &= \sum_{i=1}^{n} \left(e_i x_{ij} - e_i \bar{x}_j - \bar{e} x_{ij} + \bar{e} \bar{x}_j \right) \\ &= \sum_{i=1}^{n} e_i x_{ij} - \bar{x}_j \sum_{i=1}^{n} e_i - \bar{e} \sum_{i=1}^{n} x_{ij} + \bar{e} \bar{x}_j \end{aligned}$$

Suppose we center our data such that $\bar{x}_j = 0$. We also know that $\bar{e} = 0$. There-

fore, the above simplifies to

$$\begin{aligned} 0 &= \sum_{i=1}^{n} e_i x_{ij} \\ &= e^T Xbe \\ &= (Y - X\beta)^T X \\ &= Y^T X - \beta^T X^T X \\ &= X^T Y - X^T X\beta \\ \hat{\beta}_{MoM} &= (X^T X)^{-1} X^T Y, \end{aligned}$$

which is identical to the LSE and MLE estimate. Now, we want to find the variance of these estimates:

$$\begin{aligned} \text{var}(\hat{\beta}) &= \text{var}\left[(X^T X)^{-1} X^T Y\right] \\ &= (X^T X)^{-1} X^T \text{var}(Y) \left((X^T X)^{-1} X^T\right)^T \\ &= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} \end{aligned}$$

As a further reference, one can read section 2.6 in *Plane Answers to Complex Questions* by Ronald Christensen.

B. As mentioned above, the estimator in the previous part corresponds to the assumption that $y \sim \mathbf{N}(X\beta, \sigma^2 I)$. What happens if we instead postulate that $y \sim \mathbf{N}(X\beta, \Sigma)$, where Σ is an arbitrary known covariance matrix, not necessarily proportional to the identity? What is the MLE for β now, and what is the variance of this estimator?

First, we obtain the form of the likelihood and the log-likelihood:

$$f(y|\beta, \sigma^2) = \prod_{i=1}^{n} |2\pi\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(y_i - x_i^T\beta)^T\Sigma^{-1}(y_i - x_i^T\beta)\right]$$

$$\log f(y|\beta, \sigma^2) = \sum_{i=1}^{n} \log\left(|2\pi\Sigma|^{-1/2}\right) - \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^T\Sigma^{-1}(y_i - x_i^T\beta)$$

With the log-likelihood in hand, we can easily obtain the MLE for β by minimizing the second term, since the first is constant with respect to β:

$$\begin{aligned}
\frac{\partial}{\partial\beta}\left[\sum_{i=1}^{n}(y_i - x_i^T\beta)^T\Sigma^{-1}(y_i - x_i^T\beta)\right] &= -2\sum_{i=1}^{n} y_i^T\Sigma^{-1}x_i + 2\sum_{i=1}^{n}\beta^T x_i^T\Sigma^{-1}x_i \overset{set}{=} 0 \\
Y^T\Sigma^{-1}X &= \beta^T X^T\Sigma^{-1}X \\
X^T\Sigma^{-1}Y &= X^T\Sigma^{-1}X\beta \\
\hat{\beta} &= (X^T\Sigma^{-1}X)^- X^T\Sigma^{-1}Y,
\end{aligned}$$

where we use a generalized inverse because a standard inverse is not guaranteed. Now, we find the variance of $\hat{\beta}$:

$$\begin{aligned}
\text{var}\left[(X^T\Sigma^{-1}X)^- X^T\Sigma^{-1}Y\right] &= (X^T\Sigma^{-1}X)^- X^T\Sigma^{-1}\text{var}(Y)\left((X^T\Sigma^{-1}X)^- X^T\Sigma^{-1}\right)^T \\
&= (X^T\Sigma^{-1}X)^- X^T\Sigma^{-1}\Sigma\left((X^T\Sigma^{-1}X)^- X^T\Sigma^{-1}\right)^T \\
&= (X^T\Sigma^{-1}X)^- X^T\Sigma^{-1}X(X^T\Sigma^{-1}X)^- \\
&= (X^T\Sigma^{-1}X)^-,
\end{aligned}$$

where we used a Penrose property to obtain the last equality.

C. Show that in the special case where Σ is a diagonal matrix, that the MLE is the familiar *weighted least squares* estimator. That is, show that $\hat{\beta}$ is the solution to the following linear system of P equations in P unknowns:

$$(X^T W X)\hat{\beta} = X^T W y,$$

where W is a diagonal matrix of weights that you should relate to the σ_i^2's.

First, we start with the given equation:

$$X^T W Y = X^T W X \hat{\beta}$$

Now, consider the form of $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2)$. Recall that the inverse of a diagonal matrix is simply a diagonal matrix with its diagonal entries as reciprocals of the original matrix. That is,

$$\Sigma^{-1} = \text{diag}\left(\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, ..., \frac{1}{\sigma_n^2}\right).$$

By setting $W = \Sigma^{-1}$, the weights in W are simply the reciprocal of the variances in Σ, and we obtain the following:

$$X^T \Sigma^{-1} Y = X^T \Sigma^{-1} X \hat{\beta} \tag{2}$$

Finally, we see that equation (2) resembles our work in the previous problem. That is, the weighted LSE is simply the MLE.

A. Let's continue with the weighted least-squares estimator you just characterized, i.e. the solution to the linear system

$$(X^T W X)\hat{\beta} = X^T W y.$$

One way to calculate $\hat{\beta}$ is to: (1) recognize that, trivially, the solution to the above linear system must satisfy $\hat{\beta} = (X^T W X)^{-1} X^T W y$; and (2) to calculate this directly, i.e. by inverting $X^T W X$. Let's call this the "inversion method" for calculating the WLS solution. Numerically speaking, is the inversion method the fastest and most stable way to actually solve the above linear system? Do some independent sleuthing on this question. Summarize what you find, and provide pseudo-code for at least one alternate method based on matrix factorizations – call it "your method" for short.

Is inversion quicker than factorization?

For this problem, we consider matrix factorization using LU decomposition, inspired by Gunderson's blog post. Solving a linear system directly for $\hat{\beta}$ using factorization is roughly $\frac{2}{3}n^3 + 2n^2$ flops, whereas inverting a matrix, then multiplying the matrix out will require at least $\frac{14}{3}n^3$ flops. Therefore, inversion is much slower than directly solving the system of linear equations; in fact, the inversion method could be up to 7 times slower than using factorization to solve for $\hat{\beta}$.

Is inversion more stable than factorization?

Again, we have a resounding "no!" It may be the case that $X^T W X$ is an ill-conditioned matrix, which would result in sharp decays in singular values that lead to many instabilities. With factorization, we can work with ill-conditioned matrices with more stability, as inverting an ill-conditioned matrix comes with some instabilities. Additionally, $X^T W X$ may be singular, in which case we cannot directly take the inverse of $X^T W X$. Rather, we would have to use a generalized inverse, which are not unique and could lead to different inference.

Matrix Factorization Algorithm

Algorithm 1 "My method"

1: **GOAL:** Solve for $\hat{\beta}$ in $X^T W X \hat{\beta} = X^T W y$
2: Factor $X^T W X$ as LU using LU decomposition, where L and U are lower- and upper-triangular matrices, respectively.
3: Solve for z in $Lz = X^T W y$ using forward substitution.
4: Solve for $\hat{\beta}$ in $U\hat{\beta} = z$ using backward substitution.

B. Code up functions that implement the inversion method and "your method." Simulate some silly data from the linear model for a range of values N and P. Benchmark the performance of the inversion solver and your solver across a range of scenarios.

Given various values for N and P, I initialized the following variables:

- $W = I_n$
- $X_{ii} \sim \mathrm{N}(0, 1)$
- $y \sim \mathrm{N}(0.3 * X[, 1] + 0.5 * X[, 2], 1)$

The mean execution times for both of the methods are reported in the table below under various values for N and P.

N	P	Inversion Time (ms)	Factorization Time (ms)
10	2	0.5518008	0.4652355
100	50	165.27649	11.32176
500	100	1974.4714	112.4707
800	200	31664.8722	800.1262

As can be seen in the above table, the factorization method is more computationally efficient under various scenarios. Therefore, we have experimental evidence that the LU decomposition method is quicker than the inversion method. My personal computer was not able to implement a combination of N and P greater that 800 and 200, respectively; my R session would kill itself. [1] However, even with these various scenarios, we can see that as the parameter space and dimension of data increases, the discrepancy between mean execution time of the two methods increases.

[1] Don't worry- it has since been resurrected.