1 Bayesian inference in simple conjugate families

(A) Suppose that we take independent observations $x_1, \ldots, x_N$ 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 the w is given a $Beta(a, b)$ prior distribution:

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

Where $\Gamma(\bullet)$ denotes the Gamma function. Derive the posterior distribution $p(w|x_1, \ldots, x_N)$.

We are given that:

$$p(x_1, \ldots, x_N|w) = \prod_{i=1}^{N} w^{x_i}(1-w)^{1-x_i}$$

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

Then:

$$\begin{aligned}
p(w|x_1, \ldots, x_N) &= \frac{p(x_1, \ldots, x_N|w)p(w)}{p(x_1, \ldots, x_N)} && (1)\\
&= \frac{w^{\sum x_i}(1-w)^{\sum(1-x_i)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}w^{a-1}(1-w)^{b-1}}{\int_w w^{\sum x_i}(1-w)^{\sum(1-x_i)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}w^{a-1}(1-w)^{b-1}dw} && (2)\\
&\propto w^{\sum x_i}(1-w)^{\sum(1-x_i)}w^{a-1}(1-w)^{b-1} && (3)\\
&= w^{\sum x_i + a - 1}(1-w)^{\sum(1-x_i)+b-1} && (4)
\end{aligned}$$

So, $w|x_1, \ldots, x_N$ is $Beta(\sum x_i + a, \sum(1-x_i) + b)$:

$$p(w|x_1, \ldots, x_N) = \frac{\Gamma(a+b+N)}{\Gamma(\sum x_i + a)\Gamma(\sum(1-x_i)+b)} w^{\sum x_i + a - 1}(1-w)^{\sum(1-x_i)+b-1}$$

(B) The probability density function of a gamma random variable, $x \sim Ga(a, b)$, is:

$$p(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

Suppose that $x_1 \sim Ga(a_1, 1)$ and that $x_2 \sim Ga(a_2, 1)$. Define two new random variables $y_1 = \frac{x_1}{(x_1+x_2)}$ and $y_2 = x_1 + x_2$. Find the joint density 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.

Assuming x_1 and x_2 are independent:

$$p_x(x_1, x_2) = p(x_1)p(x_2) = \frac{1}{\Gamma(a_1)} x_1^{a_1-1} e^{-x_1} \frac{1}{\Gamma(a_2)} x_2^{a_2-1} e^{-x_2}$$

If $y_1 = \frac{x_1}{x_1+x_2}$ then $y_1 \in (0, 1)$, and if $y_2 = x_1 + x_2$ then $y_2 \in (0, \infty)$. Solving for x_1 and x_2, we find that $x_1 = y_1 y_2$, and $x_2 = y_2 - y_1 y_2$. This transformation is one-to-one and onto for $x_1, x_2 \in (0, \infty)$. Calculate the Jacobian:

$$J = \begin{vmatrix} \frac{\delta x_1}{\delta y_1} & \frac{\delta x_1}{\delta y_2} \\ \frac{\delta x_2}{\delta y_1} & \frac{\delta x_2}{\delta y_2} \end{vmatrix} = \begin{vmatrix} y_2 & y_1 \\ -y_2 & 1 - y_1 \end{vmatrix} = y_2$$

Then,

$$\begin{aligned} p_y(y_1, y_2) &= p_x(x_1 = g_1(y_1, y_2), x_2 = g_2(y_1, y_2))|J| && (5) \\ &= \frac{1}{\Gamma(a_1)} (y_1 y_2)^{a_1-1} e^{-(y_1 y_2)} \frac{1}{\Gamma(a_2)} (y_2 - y_1 y_2)^{a_2-1} e^{-(y_2 - y_1 y_2)} y_2 && (6) \\ &= \frac{1}{\Gamma(a_1)\Gamma(a_2)} y_1^{a_1-1} y_2^{a_1-1} e^{-y_1 y_2} y_2^{a_2-1} (1 - y_1)^{a_2-1} e^{-y_2} e^{y_1 y_2} y_2 && (7) \\ &= \frac{1}{\Gamma(a_1)\Gamma(a_2)} y_1^{a_1-1} (1 - y_1)^{a_2-1} y_2^{a_1+a_2-1} e^{-y_2} && (8) \\ &= \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} y_1^{a_1-1} (1 - y_1)^{a_2-1} \frac{1}{\Gamma(a_1 + a_2)} e^{-y_2} y_2^{a_1+a_2-1} && (9) \end{aligned}$$

Notice that this can be separated into a function of y_1 and a function of y_2, this means that y_1 and y_2 are independent, where $y_1 \sim Beta(a_1, a_2)$ and $y_2 \sim Gamma(a_1+a_2, 1)$. This means that if you can sample $x_1 \sim Gamma(a_1, 1)$ and $x_2 \sim Gamma(a_2, 1)$, then you can sample a $Beta(a_1, a_2)$ as $\frac{x_1}{x_1+x_2}$.

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

We are given that:

$$p(x_1, \ldots, x_N|\theta, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}}$$

$$p(\theta) = \frac{1}{\sqrt{2\pi v}} e^{-\frac{(\theta-m)^2}{2v}}$$

Then:

$$\begin{aligned}
p(\theta|x_1, \ldots, x_n, \sigma^2) &= \frac{p(x_1, \ldots, x_N|\theta, \sigma^2)p(\theta)}{p(x_1, \ldots, x_N)} \tag{10} \\
&= \frac{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{-\frac{\sum_{i=1}^N (x_i-\theta)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi v}} e^{-\frac{(\theta-m)^2}{2v}}}{\int_\theta \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{-\frac{\sum_{i=1}^N (x_i-\theta)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi v}} e^{-\frac{(\theta-m)^2}{2v}} d\theta} \tag{11} \\
&\propto e^{-\frac{\sum_{i=1}^N (x_i-\theta)^2}{2\sigma^2}} e^{-\frac{\sum_{i=1}^N (x_i-\theta)^2}{2\sigma^2}} \tag{12} \\
&= e^{-\frac{\sum_{i=1}^N x_i^2 + 2\theta \sum_{i=1}^N x_i + N\theta^2}{2\sigma^2}} e^{-\frac{\theta^2+2m\theta+m^2}{2v}} \tag{13} \\
&\propto e^{-\frac{2\theta \sum_{i=1}^N x_i + N\theta^2}{2\sigma^2}} e^{-\frac{\theta^2+2m\theta}{2v}} \tag{14} \\
&= e^{-\frac{1}{2}(\theta^2(\frac{1}{v}+\frac{N}{\sigma^2})+2\theta(\frac{m}{v}+\frac{\sum_{i=1}^N x_i}{\sigma^2}))} \tag{15} \\
&= e^{-\frac{\theta^2+2\theta(\frac{v\sum_{i=1}^N x_i + m\sigma^2}{\sigma^2+vN}))}{2\frac{v\sigma^2}{\sigma^2+vN}}} \tag{16} \\
&\propto e^{-\frac{\left(\theta+\frac{v\sum_{i=1}^N x_i + m\sigma^2}{\sigma^2+vN}\right)^2}{2\frac{v\sigma^2}{\sigma^2+vN}}} \tag{17}
\end{aligned}$$

Notice that this is the kernal of a normal distribution so:

$$\theta|x_1, \ldots, x_n, \sigma^2 \sim N\left(\frac{v\sum_{i=1}^N x_i + m\sigma^2}{\sigma^2+vN}, \frac{v\sigma^2}{\sigma^2+vN}\right)$$

or

$$p(\theta|x_1, \ldots, x_n, \sigma^2) = \frac{1}{\sqrt{2\pi\frac{v\sigma^2}{\sigma^2+vN}}} e^{-\frac{\left(\theta+\frac{v\sum_{i=1}^N x_i + m\sigma^2}{\sigma^2+vN}\right)^2}{2\frac{v\sigma^2}{\sigma^2+vN}}}$$

(D) Suppose that we take independent observations $x_1, \ldots, x_N$ from a normal sampling model with known mean θ but unknown variance σ^2. To make things easier, we will re-express things in terms of the precision, or inverse variance $\omega = \frac{1}{\sigma^2}$:

$$p(x_i|\theta, \omega) = \left(\frac{\omega}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{\omega}{2}(x_i-\theta)^2}.$$

Suppose that 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 $p(\omega|x_1, \ldots, x_N)$. Re-express this as a posterior for σ^2, the variance.

We are given that:

$$p(x_1, \ldots, x_N|\omega, \theta) = \prod_{i=1}^{N} \left(\frac{\omega}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{\omega}{2}(x_i-\theta)^2}$$

$$p(\omega) = \frac{b^a}{\Gamma(a)}\omega^{a-1}e^{-b\omega}$$

So,

$$\begin{aligned}
p(\omega|x_1, \ldots, x_N) &= \frac{p(x_1, \ldots, x_N|\omega, \theta)p(\omega)}{p(x_1, \ldots, x_N)} && (18)\\
&= \frac{\left(\frac{\omega}{2\pi}\right)^{\frac{N}{2}} e^{-\frac{\omega}{2}\sum_{i=1}^{N}(x_i-\theta)^2} \frac{b^a}{\Gamma(a)}\omega^{a-1}e^{-b\omega}}{\int_\omega \left(\frac{\omega}{2\pi}\right)^{\frac{N}{2}} e^{-\frac{\omega}{2}\sum_{i=1}^{N}(x_i-\theta)^2} \frac{b^a}{\Gamma(a)}\omega^{a-1}e^{-b\omega}d\omega} && (19)\\
&\propto \omega^{\frac{N}{2}} e^{-\frac{\omega}{2}\sum_{i=1}^{N}(x_i-\theta)^2}\omega^{a-1}e^{-b\omega} && (20)\\
&= \omega^{\frac{N}{2}+a-1}e^{-\omega(b+\frac{1}{2}\sum_{i=1}^{N}(x_i-\theta)^2)} && (21)
\end{aligned}$$

This is the kernal of a gamma distribution, so:

$$\omega|x_1, \ldots, x_N \sim Gamma\left(\frac{N}{2}+a, b+\frac{1}{2}\sum_{i=1}^{N}(x_i-\theta)^2\right)$$

or,

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

Then, $\sigma^2|x_1,\ldots,x_N$ is Inverse Gamma:

$$p(\sigma^2|x_1,\ldots,x_N) = \frac{\left(b+\frac{1}{2}\sum_{i=1}^N (x_i-\theta)^2\right)^{\frac{N}{2}+a}}{\Gamma(\frac{N}{2}+a)}(\sigma^2)^{-(\frac{N}{2}+a)+1}e^{-\frac{(b+\frac{1}{2}\sum_{i=1}^N (x_i-\theta)^2)}{\sigma^2}}(\sigma^2)^{-2} \tag{22}$$

$$= \frac{\left(b+\frac{1}{2}\sum_{i=1}^N (x_i-\theta)^2\right)^{\frac{N}{2}+a}}{\Gamma(\frac{N}{2}+a)}(\sigma^2)^{-(\frac{N}{2}+a)-1}e^{-\frac{(b+\frac{1}{2}\sum_{i=1}^N (x_i-\theta)^2)}{\sigma^2}} \tag{23}$$

(E) Suppose that, as above, we take independent observations $x_1,\ldots,x_N$ from a normal sampling model with unknown, common mean θ. This time, however, each observation has its own idiosyncratic (but known) variance: $x_i \sim N\left(\theta,\sigma_i^2\right)$. Suppose that θ is given a normal prior distribution with mean m and variance v. Derive the posterior distribution $p(\theta|x_1,\ldots,x_N)$. Express the posterior mean in a form that is clearly interpretable as a weighted average of the observations and the prior mean.

We are given that:

$$p(x_1,\ldots,x_N|\sigma_1^2,\ldots,\sigma_N^2,\theta) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma_i^2}}e^{-\frac{(x_i-\theta)^2}{2\sigma_i^2}}$$

$$p(\theta) = \frac{1}{\sqrt{2\pi v}}e^{-\frac{(\theta-m)^2}{2v}}$$

Then,

$$
\begin{aligned}
p(\theta|x_1,\ldots,x_N) &= \frac{p(\theta)p(x_1,\ldots,x_N|\sigma_1^2,\ldots,\sigma_N^2,\theta)}{p(x_1,\ldots,x_N}) && (24)\\
&= \frac{\frac{1}{\sqrt{2\pi v}}e^{-\frac{(\theta-m)^2}{2v}}\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma_i^2}}e^{-\frac{(x_i-\theta)^2}{2\sigma_i^2}}}{\int_\theta \frac{1}{\sqrt{2\pi v}}e^{-\frac{(\theta-m)^2}{2v}}\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma_i^2}}e^{-\frac{(x_i-\theta)^2}{2\sigma_i^2}}d\theta} && (25)\\
&\propto \frac{1}{\sqrt{2\pi v}}e^{-\frac{(\theta-m)^2}{2v}}\prod_{i=1}^{N}\left(\frac{1}{\sqrt{2\pi\sigma_i^2}}\right)e^{-\frac{1}{2}\sum_{i=1}^{N}\frac{(x_i-\theta)^2}{\sigma_i^2}} && (26)\\
&\propto e^{-\frac{1}{2}\left(\frac{\theta^2}{v}-\frac{2m\theta}{v}+\frac{m^2}{v}\right)}e^{-\frac{1}{2}\left(\sum_{i=1}^{N}\frac{x_i^2}{\sigma_i^2}-2\theta\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}+\theta^2\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\right)} && (27)\\
&\propto e^{-\frac{1}{2}\left(\frac{\theta^2}{v}-\frac{2m\theta}{v}-2\theta\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}+\theta^2\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\right)} && (28)\\
&= e^{-\frac{1}{2}\left(\theta^2\left(\frac{1}{v}+\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\right)-2\theta\left(\frac{m}{v}-\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}\right)\right)} && (29)\\
&= exp\left\{-\frac{\theta^2-2\theta\left(\frac{m}{v}-\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}\right)\left(\frac{1}{\frac{1}{v}+\sum_{i=1}^{N}\frac{1}{\sigma_i^2}}\right)}{2\left(\frac{1}{\frac{1}{v}+\sum_{i=1}^{N}\frac{1}{\sigma_i^2}}\right)}\right\} && (30)\\
&\propto exp\left\{-\frac{\left(\theta-\left(\frac{m}{v}-\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}\right)\left(\frac{1}{\frac{1}{v}+\sum_{i=1}^{N}\frac{1}{\sigma_i^2}}\right)\right)^2}{2\left(\frac{1}{\frac{1}{v}+\sum_{i=1}^{N}\frac{1}{\sigma_i^2}}\right)}\right\} && (31)
\end{aligned}
$$

Notice that this is the kernal of a normal distribution, so:

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

or

$$p(\theta|x_1,\ldots,x_N) = \frac{1}{\sqrt{2\pi\left(\frac{1}{\frac{1}{v}+\sum_{i=1}^{N}\frac{1}{\sigma_i^2}}\right)}}exp\left\{-\frac{\left(\theta-\left(\frac{m}{v}-\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}\right)\left(\frac{1}{\frac{1}{v}+\sum_{i=1}^{N}\frac{1}{\sigma_i^2}}\right)\right)^2}{2\left(\frac{1}{\frac{1}{v}+\sum_{i=1}^{N}\frac{1}{\sigma_i^2}}\right)}\right\}$$

(F) Suppose that $(x|\sigma^2) \sim N(0, \sigma^2)$, and that $\frac{1}{\sigma^2}$ has a $Gamma(a, b)$ prior, defined as above. Show that the marginal distribution of x is Student's t. This is why the t distribution is often referred to as a scale mixture of normals.

Let $\omega = \frac{1}{\sigma^2}$, then we have that that:

$$p(x|\omega) = \left(\frac{\omega}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{\omega}{2}x^2}$$

$$p(\omega) = \frac{b^a}{\Gamma(a)}\omega^{a-1}e^{-b\omega}$$

We are interested in finding $p(x)$, where:

$$p(x) = \int_\omega p(x,\omega)d\omega = \int_\omega p(x|\omega)p(\omega)d\omega.$$

So:

$$\begin{aligned}
p(x) &= \int_\omega p(x|\omega)p(\omega)d\omega \tag{32}\\
&= \int_\omega \frac{b^a}{\Gamma(a)}\omega^{a-1}e^{-b\omega}\left(\frac{\omega}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{\omega}{2}x^2}d\omega \tag{33}\\
&= \frac{b^a}{\Gamma(a)\sqrt{2\pi}}\int_\omega \omega^{a-1}e^{-b\omega}\omega^{\frac{1}{2}}e^{-\frac{\omega}{2}x^2}d\omega \tag{34}\\
&= \frac{b^a}{\Gamma(a)\sqrt{2\pi}}\int_\omega \omega^{a-\frac{1}{2}}e^{-\omega\left(b+\frac{x^2}{2}\right)}d\omega \tag{35}\\
&= \frac{b^a}{\Gamma(a)\sqrt{2\pi}}\frac{\Gamma(a+\frac{1}{2})}{\left(b+\frac{x^2}{2}\right)^{a+\frac{1}{2}}}\int_\omega \frac{\left(b+\frac{x^2}{2}\right)^{a+\frac{1}{2}}}{\Gamma(a+\frac{1}{2})}\omega^{a+\frac{1}{2}-1}e^{-\omega\left(b+\frac{x^2}{2}\right)}d\omega \tag{36}\\
&= \frac{b^a}{\Gamma(a)\sqrt{2\pi}}\frac{\Gamma(a+\frac{1}{2})}{\left(b+\frac{x^2}{2}\right)^{a+\frac{1}{2}}} \tag{37}\\
&= \frac{\Gamma(\frac{2a+1}{2})\sqrt{a}}{\Gamma(\frac{2a}{2})\sqrt{2a\pi}}\frac{b^a}{\left(b+\frac{ax^2}{2a}\right)^{\frac{2a+1}{2}}} \tag{38}\\
&= \frac{\Gamma(\frac{2a+1}{2})\sqrt{a}}{\Gamma(\frac{2a}{2})\sqrt{2a\pi}}\frac{b^a}{b^{a+\frac{1}{2}}\left(1+\frac{ax^2}{b2a}\right)^{\frac{2a+1}{2}}} \tag{39}\\
&= \frac{\Gamma(\frac{2a+1}{2})}{\Gamma(\frac{2a}{2})\sqrt{2a\pi}}\sqrt{\frac{a}{b}}\left(1+\frac{ax^2}{b2a}\right)^{-\frac{2a+1}{2}} \tag{40}\\
&= \frac{\Gamma(\frac{2a+1}{2})}{\Gamma(\frac{2a}{2})\sqrt{2a\pi\left(\frac{b}{a}\right)}}\left(1+\frac{x^2}{\left(\frac{b}{a}\right)2a}\right)^{-\frac{2a+1}{2}} \tag{41}
\end{aligned}$$

Let $\nu = 2a$, and $\sigma^2 = \frac{b}{a}$:

$$= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi\sigma^2}} \left(1 + \frac{x^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} \tag{42}$$

This is a non-standardized student's t with degrees of freedom $\nu = \frac{a}{2}$ and scale parameter $\sigma^2 = \frac{b}{a}$. Note that if you set $a = b$ in your Gamma prior, you will get a standard student's t distribution.

2 The multivariate normal distribution

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

(1)

$$\begin{aligned}
cov(\mathbf{x}) &= E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] && (43)\\
&= \begin{bmatrix} E[(x_i - \mu_i)(x_j - \mu_j)] \end{bmatrix} && (44)\\
&= \begin{bmatrix} E[x_i x_j - x_i\mu_j - \mu_i x_j - \mu_i\mu_j] \end{bmatrix} && (45)\\
&= \begin{bmatrix} E[x_i x_j] - E[x_i]\mu_j - \mu_i E[x_j] + \mu_i\mu_j \end{bmatrix} && (46)\\
&= \begin{bmatrix} E[x_i x_j] - 2\mu_i\mu_j + \mu_i\mu_j \end{bmatrix} && (47)\\
&= \begin{bmatrix} E[x_i x_j] - \mu_i\mu_j \end{bmatrix} && (48)\\
&= E(\mathbf{x}\mathbf{x}^T) - \boldsymbol{\mu}\boldsymbol{\mu}^T && (49)
\end{aligned}$$

(2)

$$\begin{aligned}
cov(\mathbf{A}\mathbf{x} + \mathbf{b}) &= E[(\mathbf{A}\mathbf{x} + \mathbf{b} - E[\mathbf{A}\mathbf{x} + \mathbf{b}])(\mathbf{A}\mathbf{x} + \mathbf{b} - E[\mathbf{A}\mathbf{x} + \mathbf{b}])^T] && (50)\\
&= E[(\mathbf{A}\mathbf{x} + \mathbf{b} - \mathbf{A}E[\mathbf{x}] - \mathbf{b})(\mathbf{A}\mathbf{x} + \mathbf{b} - \mathbf{A}E[\mathbf{x}] - \mathbf{b})^T] && (51)\\
&= E[(\mathbf{A}(\mathbf{x} - E[\mathbf{x}]))(\mathbf{A}(\mathbf{x} - E[\mathbf{x}]))^T] && (52)\\
&= E[\mathbf{A}(\mathbf{x} - E[\mathbf{x}])(\mathbf{x} - E[\mathbf{x}])^T\mathbf{A}^T] && (53)\\
&= \mathbf{A}E[(\mathbf{x} - E[\mathbf{x}])(\mathbf{x} - E[\mathbf{x}])^T]\mathbf{A}^T && (54)\\
&= \mathbf{A}cov(\mathbf{x})\mathbf{A}^T && (55)
\end{aligned}$$

(B) Consider the random vector $\mathbf{z} = (z_1, \ldots, z_p)^T$, with each entry having an independent standard normal distribution. Derive the probability density function (PDF) and moment-generating function (MGF) of $\mathbf{z}$, expressed in vector notation. We say that z has a standard multivariate normal distribution.

PDF: We know that $p(z_i) = \frac{1}{\sqrt{2\pi}} exp(\frac{z_i^2}{2})$. Since all the z_i's are independent, $p(z_1, z_2, \ldots, z_p) = p(z_1)p(z_2)\ldots p(z_p)$. So:

$$\begin{aligned}
p(z_1, \ldots, z_p) &= p(z_1)p(z_2)\ldots p(z_p) && (56)\\
&= \frac{1}{\sqrt{2\pi}} e^{\frac{z_1^2}{2}} \frac{1}{\sqrt{2\pi}} e^{\frac{z_2^2}{2}} \ldots \frac{1}{\sqrt{2\pi}} e^{\frac{z_p^2}{2}} && (57)\\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^p e^{\frac{z_1^2}{2}} e^{\frac{z_2^2}{2}} \ldots e^{\frac{z_p^2}{2}} && (58)\\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^p e^{\frac{1}{2}\sum_{i=1}^p z_i^2} && (59)\\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^p e^{\frac{1}{2}\mathbf{z}^T\mathbf{z}} && (60)
\end{aligned}$$

MGF: The mgf is $E[e^{\mathbf{t}^T\mathbf{z}}]$.

$$\begin{aligned}
E[e^{\mathbf{t}^T\mathbf{z}}] &= E[e^{\sum_{i=1}^p t_i z_i}] && (61)\\
&= E[\prod_{i=1}^p e^{t_i z_i}] && (62)\\
&= \prod_{i=1}^p E[e^{t_i z_i}] && (63)\\
&= \prod_{i=1}^p e^{\frac{t_i^2}{2}} && (64)\\
&= e^{\sum_{i=1}^p \frac{t_i^2}{2}} && (65)\\
&= e^{\frac{\mathbf{t}^T\mathbf{t}}{2}} && (66)
\end{aligned}$$

Line (63) follows from the independence of the z_i's, and line (64) follows from $E[e^{t_i z_i}]$ being the definition of the mgf of z_i, which is $exp\left(\frac{t_i^2}{2}\right)$.

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

Let us first assume that $\mathbf{x}$ is multivariate normal, $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We know that for all vectors $\mathbf{a}$ not identically zero, $z = \mathbf{a}^T\mathbf{x}$ is normally distributed. Let $z \sim N(m, v)$. Then,

$$m = E[z] = E[\mathbf{a}^T\mathbf{x}] = \mathbf{a}E[\mathbf{x}] = \mathbf{a}^T\boldsymbol{\mu}$$

and

$$v = var(z) = var(\mathbf{a}^T\mathbf{x}) = \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a}.$$

Since,

$$E[exp\{tz\}] = exp\{mt + \frac{vt^2}{2}\},$$

it follows that

$$E[exp\{t\mathbf{a}^T\mathbf{x}\}] = exp\{t\mathbf{a}^T\boldsymbol{\mu} + \frac{t\mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a}t}{2}\}.$$

Since t can take on any real number and $\mathbf{a}$ can be any vector that is not identically zero, let $\mathbf{a}t = \mathbf{t}$. Then, we have that

$$E[exp\{\mathbf{t}^T\mathbf{x}\}] = exp\{\mathbf{t}^T\boldsymbol{\mu} + \frac{\mathbf{t}^T\boldsymbol{\Sigma}\mathbf{t}}{2}\},$$

where $\mathbf{x}$ is multivariate normal.

Since an mgf uniquely defines a distribution, having shown that the mgf of a multivariate normal is of the form $exp\{\mathbf{t}^T\boldsymbol{\mu} + \mathbf{t}^T\boldsymbol{\Sigma}\mathbf{t}/2\}$, it follows that an mgf of this form describes a multivariate normal.

(D) Another basic theorem is that a random vector is multivariate normal if and only if 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 $\mathbf{x} = \mathbf{Lz} + \boldsymbol{\mu}$ for some $p \times p$ matrix L of full column rank. Prove that $\mathbf{x}$ is multivariate normal. In addition, use the moment identities you proved above to compute the expected value and covariance matrix of x.

Let us try to find the moment generating function of $x = Lz + \mu$:

$$E\left[e^{\mathbf{t}^T\mathbf{x}}\right] = E\left[e^{\mathbf{t}^T(\mathbf{Lz}+\boldsymbol{\mu})}\right] \tag{67}$$

$$= E\left[e^{\mathbf{t}^T\mathbf{Lz}+\mathbf{t}\boldsymbol{\mu}}\right] \tag{68}$$

$$= e^{\mathbf{t}^T\boldsymbol{\mu}}E\left[e^{\mathbf{t}^T\mathbf{Lz}}\right] \tag{69}$$

Let $\mathbf{L}^T\mathbf{t} = \mathbf{w}$, then consider $E\left[e^{\mathbf{t}^T\mathbf{Lz}}\right] = E\left[e^{\mathbf{w}^T\mathbf{z}}\right]$. Since we know the mgf of a standard multivariate normal, z, we know $E\left[e^{\mathbf{w}^T\mathbf{z}}\right] = e^{\frac{\mathbf{w}^T\mathbf{w}}{2}}$. Plugging back in we have, $E\left[e^{\mathbf{t}^T\mathbf{Lz}}\right] = e^{\frac{\mathbf{t}^T\mathbf{LL}^T\mathbf{t}}{2}}$. Then:

$$E\left[e^{\mathbf{t}^T\mathbf{x}}\right] = e^{\mathbf{t}^T\boldsymbol{\mu}}e^{\frac{\mathbf{t}^T\mathbf{LL}^T\mathbf{t}}{2}} \tag{70}$$

$$= e^{\mathbf{t}^T\boldsymbol{\mu}+\frac{\mathbf{t}^T\mathbf{LL}^T\mathbf{t}}{2}} \tag{71}$$

This is multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = \mathbf{LL}^T$.

(E) Now for the "only if." Suppose that $\mathbf{x}$ 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.

Assume $\mathbf{x}$ has a multivariate normal distribution, then the mgf of $\mathbf{x}$ can be written as:

$$exp\{\mathbf{t}^T\boldsymbol{\mu} + \frac{\mathbf{t}^T\boldsymbol{\Sigma}\mathbf{t}}{2}\}.$$

Since $\boldsymbol{\Sigma}$ is by definition a symmetric square matrix, take the Cholesky decomposition of $\boldsymbol{\Sigma}$, so $\boldsymbol{\Sigma} = \mathbf{LL}^T$ for some $\mathbf{L}$. Now rewrite the mgf with the Cholesky decomposition:

$$\begin{aligned}
exp\{\mathbf{t}^T\boldsymbol{\mu} + \frac{\mathbf{t}^T\boldsymbol{\Sigma}\mathbf{t}}{2}\} &= exp\{\mathbf{t}^T\boldsymbol{\mu} + \frac{\mathbf{t}^T\mathbf{L}\mathbf{L}^T\mathbf{t}}{2}\} && (72)\\
&= exp\{\mathbf{t}^T\boldsymbol{\mu}\}exp\{\frac{\mathbf{t}^T\mathbf{L}\mathbf{L}^T\mathbf{t}}{2}\} && (73)\\
&= e^{\mathbf{t}^T\boldsymbol{\mu}}E[e^{\mathbf{t}^T\mathbf{L}\mathbf{z}}] && (74)\\
&= E[e^{\mathbf{t}^T(\mathbf{L}\mathbf{z}+\boldsymbol{\mu})}] && (75)
\end{aligned}$$

Here $\mathbf{z}$ is standard multivariate normal, and we see that we have arrived at the mgf of an affine transformation of a standard multivariate normal, so our original multivariate normal must be an affine transformation of a standard multivariate normal.

We can use this information to create draws for any multivariate normal if we can take draws from a standard multivariate normal. Take draws from a standard multivariate normal, $\mathbf{z}$, then take the transformation $\mathbf{L}\mathbf{z} + \boldsymbol{\mu}$, where $\boldsymbol{\mu}$ is the mean of your desired multivariate normal, and $\mathbf{L}$ is the Cholesky decomposition, or any other decomposition, of your desired covariance matrix.

(F) Use this last result, together with the PDF of a standard multivariate normal, to show that the PDF of a multivariate normal $x \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ takes the form $p(\mathbf{x}) = \mathbf{c}exp\{-\frac{\mathbf{Q}(\mathbf{x}-\boldsymbol{\mu})}{2}\}$ for some constant $\mathbf{C}$ and quadratic form $\mathbf{Q}(\mathbf{x} - \boldsymbol{\mu})$

Consider $\mathbf{z} = (z_1, \ldots, z_p)^T$, where $\mathbf{z}$ is distributed standard multivariate normal, and take $\mathbf{x}$ to be the transformation $\mathbf{x} = \mathbf{L}\mathbf{z} + \boldsymbol{\mu}$. Then, $g^{-1}(\mathbf{x}) = \mathbf{L}^{-1}(\mathbf{x} - \boldsymbol{\mu})$, since this is an affine transformation it is clearly one-to-one and onto. The Jacobian, $\mathbf{J} = \mathbf{L}^{-1}$. We have shown that:

$$p(\mathbf{z}) = \left(\frac{1}{\sqrt{2\pi}}\right)^p e^{-\frac{1}{2}\mathbf{z}^T\mathbf{z}}.$$

Then:

$$\begin{aligned}
p(\mathbf{x}) &= p(g^{-1}(\mathbf{x}))det(\mathbf{J}) && (76)\\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^p e^{-\frac{1}{2}(\mathbf{L}^{-1}(\mathbf{x}-\boldsymbol{\mu}))^T(\mathbf{L}^{-1}(\mathbf{x}-\boldsymbol{\mu}))}det(\mathbf{L}^{-1}) && (77)\\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^p \frac{1}{det(\mathbf{L})}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{L}^T\mathbf{L})^{-1}(\mathbf{x}-\boldsymbol{\mu})} && (78)
\end{aligned}$$

This is the desired form with $\mathbf{C} = \left(\frac{1}{\sqrt{2\pi}}\right)^p \frac{1}{det(\mathbf{L})}$, and $\mathbf{Q} = (\mathbf{L}^T\mathbf{L})^{-1} = \boldsymbol{\Sigma}^{-1}$

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

Since $\mathbf{x_1}$ and $\mathbf{x_2}$ are both multivariate normal, let us express them each explicitly as a linear combination of independent univariate normals.

$$\mathbf{x_1} = \mathbf{L_1}\mathbf{z_1} + \boldsymbol{\mu_1}$$

$$\mathbf{x_2} = \mathbf{L_2}\mathbf{z_2} + \boldsymbol{\mu_2}.$$

Then,

$$\begin{aligned}
\mathbf{y} &= \mathbf{A}\mathbf{x_1} + \mathbf{B}\mathbf{x_2} && (79)\\
&= \mathbf{A}(\mathbf{L_1}\mathbf{z_1} + \boldsymbol{\mu_1}) + \mathbf{B}(\mathbf{L_2}\mathbf{z_2} + \boldsymbol{\mu_2}) && (80)\\
&= \mathbf{A}\mathbf{L_1}\mathbf{z_1} + \mathbf{B}\mathbf{L_2}\mathbf{z_2} + \mathbf{A}\boldsymbol{\mu_1} + \mathbf{B}\boldsymbol{\mu_2} && (81)\\
&= \begin{pmatrix} \mathbf{A}\mathbf{L_1} & \mathbf{B}\mathbf{L_2} \end{pmatrix} \begin{pmatrix} \mathbf{z_1} \\ \mathbf{z_2} \end{pmatrix} + \mathbf{A}\boldsymbol{\mu_1} + \mathbf{B}\boldsymbol{\mu_2} && (82)
\end{aligned}$$

Since $\mathbf{A}\boldsymbol{\mu_1} + \mathbf{B}\boldsymbol{\mu_2}$ is just a vector of constants, $\mathbf{y}$ is an affine transformation of independent univariate normals, this makes $\mathbf{y}$ multivariate normal by a previously proven result. Since $\mathbf{y}$ is multivariate normal, it suffices to find the mean and variance of $\mathbf{y}$ and plug them into the pdf for a multivariate normal.

$$E[\mathbf{y}] = E[\mathbf{A}\mathbf{x_1} + \mathbf{B}\mathbf{x_2}] = \mathbf{A}E[\mathbf{x_1}] + \mathbf{B}E[\mathbf{x_2}] = \mathbf{A}\boldsymbol{\mu_1} + \mathbf{B}\boldsymbol{\mu_2}$$

$$\begin{aligned}
Cov(\mathbf{y}) &= Cov(\mathbf{A}\mathbf{x_1} + \mathbf{B}\mathbf{x_2}) && (83)\\
&= \mathbf{A}Cov(\mathbf{x_1})\mathbf{A}^T + \mathbf{A}Cov(\mathbf{x_1}, \mathbf{x_2})\mathbf{B}^T && (84)\\
&\quad + \mathbf{B}Cov(\mathbf{x_2}, \mathbf{x_1})\mathbf{A}^T + \mathbf{B}Cov(\mathbf{x_2})\mathbf{B}^T && (85)\\
&= \mathbf{A}Cov(\mathbf{x_1})\mathbf{A}^T + \mathbf{B}Cov(\mathbf{x_2})\mathbf{B}^T && (86)\\
&= \mathbf{A}\boldsymbol{\Sigma_1}\mathbf{A}^T + \mathbf{B}\boldsymbol{\Sigma_2}\mathbf{B}^T && (87)
\end{aligned}$$

So,

$$\mathbf{y} \sim N(\mathbf{A}\boldsymbol{\mu_1} + \mathbf{B}\boldsymbol{\mu_2}, \mathbf{A}\boldsymbol{\Sigma_1}\mathbf{A}^T + \mathbf{B}\boldsymbol{\Sigma_2}\mathbf{B}^T)$$

or, let $\boldsymbol{\mu} = \mathbf{A}\boldsymbol{\mu_1} + \mathbf{B}\boldsymbol{\mu_2}$ and $\boldsymbol{\Sigma} = \mathbf{A}\boldsymbol{\Sigma_1}\mathbf{A}^T + \mathbf{B}\boldsymbol{\Sigma_2}\mathbf{B}^T$, and p be the dimension of the vector $\boldsymbol{\mu}$ then:

$$p(\mathbf{y}) = \left(\frac{1}{\sqrt{2\pi}}\right)^p (det(\boldsymbol{\Sigma}))^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T(\boldsymbol{\Sigma})^{-1}(\mathbf{y}-\boldsymbol{\mu})}.$$

3 Conditionals and marginals

Suppose $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has a multivariate normal distribution. Let $\mathbf{x_1}$ and $\mathbf{x_2}$ denote an arbitrary partition of $\mathbf{x}$ into two sets of components. Because we can relabel the compnents of $\mathbf{x}$ without changing their distribution, we can safely assume that $\mathbf{x_1}$ comprises the first k elements of $\mathbf{x}$, and $\mathbf{x_2}$ the last $p-k$. We will also assume that $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ have been partitioned comfortably with $\mathbf{x}$:

$$\boldsymbol{\mu} = (\boldsymbol{\mu_1}, \boldsymbol{\mu_2})^T \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{pmatrix}$$

Clearly $\boldsymbol{\Sigma_{12}} = \boldsymbol{\Sigma_{21}}^T$ as $\boldsymbol{\Sigma}$ is symmetric.

(A) Derive the marginal distribution of $\mathbf{x_1}$:

We can rewrite $\mathbf{x_1}$ as:

$$\mathbf{x_1} = (\mathbf{I}_{k\times k}, \mathbf{0}_{k\times(p-k)})\mathbf{x},$$

where $\mathbf{I}_{k\times k}$ is the $k \times k$ identity matrix, and $\mathbf{0}_{k\times(p-k)}$ is a $k \times (p-k)$ matrix of zeros. By an earlier proof, a random vector is multivariate normal if and only if it is an affine transformation of independent univariate normals. So, $\mathbf{x}$ is an affine transformation of independent univariate normals, and it follows that $(\mathbf{I}_{k\times k}, \mathbf{0}_{k\times(p-k)})\mathbf{x}$ is an affine transformation of independent univariate normals, therefore $\mathbf{x_1}$ is multivariate normal, so it suffices to find the mean and variance of $\mathbf{x_1}$. First we find the mean of $\mathbf{x_1}$:

$$\begin{aligned} E[\mathbf{x_1}] &= E[(\mathbf{I}_{k\times k}, \mathbf{0}_{k\times(p-k)})\mathbf{x}] && (88)\\ &= (\mathbf{I}_{k\times k}, \mathbf{0}_{k\times(p-k)})E[\mathbf{x}] && (89)\\ &= (\mathbf{I}_{k\times k}, \mathbf{0}_{k\times(p-k)})\boldsymbol{\mu} && (90)\\ &= (\mathbf{I}_{k\times k}, \mathbf{0}_{k\times(p-k)})(\boldsymbol{\mu_1}, \boldsymbol{\mu_2})^T && (91)\\ &= \boldsymbol{\mu_1} && (92) \end{aligned}$$

Next we find the variance of $\mathbf{x_1}$:

$$\begin{aligned} Var(\mathbf{x_1}) &= Var((\mathbf{I}_{k\times k}, \mathbf{0}_{k\times(p-k)})\mathbf{x}) && (93)\\ &= (\mathbf{I}_{k\times k}, \mathbf{0}_{k\times(p-k)})Var(\mathbf{x})(\mathbf{I}_{k\times k}, \mathbf{0}_{k\times(p-k)})^T && (94)\\ &= (\mathbf{I}_{k\times k}, \mathbf{0}_{k\times(p-k)})\begin{pmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{pmatrix}(\mathbf{I}_{k\times k}, \mathbf{0}_{k\times(p-k)})^T && (95)\\ &= (\boldsymbol{\Sigma_{11}}, \boldsymbol{\Sigma_{12}})(\mathbf{I}_{k\times k}, \mathbf{0}_{k\times(p-k)})^T && (96)\\ &= \boldsymbol{\Sigma_{11}} && (97) \end{aligned}$$

So, $\mathbf{x_1} \sim N(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_{11}})$.

(B) Let $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ be the inverse covariance matrix, or precision matrix, of $\mathbf{x}$, and partition $\mathbf{\Omega}$ just as you did $\mathbf{\Sigma}$:

$$\mathbf{\Omega} = \begin{pmatrix} \mathbf{\Omega_{11}} & \mathbf{\Omega_{12}} \\ \mathbf{\Omega_{21}} & \mathbf{\Omega_{22}} \end{pmatrix}$$

Express each block of $\mathbf{\Omega}$ as blocks of $\mathbf{\Sigma}$.

So, $\mathbf{\Sigma}\mathbf{\Sigma}^{-1} = \mathbf{\Sigma}\mathbf{\Omega} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Then:

$$\begin{pmatrix} \mathbf{\Sigma_{11}} & \mathbf{\Sigma_{12}} \\ \mathbf{\Sigma_{21}} & \mathbf{\Sigma_{22}} \end{pmatrix} \begin{pmatrix} \mathbf{\Omega_{11}} & \mathbf{\Omega_{12}} \\ \mathbf{\Omega_{21}} & \mathbf{\Omega_{22}} \end{pmatrix} = \mathbf{I} \tag{98}$$

$$\begin{pmatrix} \mathbf{\Sigma_{11}\Omega_{11}} + \mathbf{\Sigma_{12}\Omega_{21}} & \mathbf{\Sigma_{11}\Omega_{12}} + \mathbf{\Sigma_{12}\Omega_{22}} \\ \mathbf{\Sigma_{21}\Omega_{11}} + \mathbf{\Sigma_{22}\Omega_{21}} & \mathbf{\Sigma_{21}\Omega_{12}} + \mathbf{\Sigma_{22}\Omega_{22}} \end{pmatrix} = \mathbf{I} \tag{99}$$

This gives us the following system of equations:

$$\mathbf{\Sigma_{11}\Omega_{11}} + \mathbf{\Sigma_{12}\Omega_{21}} = \mathbf{I} \tag{100}$$

$$\mathbf{\Sigma_{11}\Omega_{12}} + \mathbf{\Sigma_{12}\Omega_{22}} = \mathbf{0} \tag{101}$$

$$\mathbf{\Sigma_{21}\Omega_{11}} + \mathbf{\Sigma_{22}\Omega_{21}} = \mathbf{0} \tag{102}$$

$$\mathbf{\Sigma_{21}\Omega_{12}} + \mathbf{\Sigma_{22}\Omega_{22}} = \mathbf{I}. \tag{103}$$

Let us start with equation (102):

$$\mathbf{\Sigma_{21}\Omega_{11}} + \mathbf{\Sigma_{22}\Omega_{21}} = \mathbf{0} \tag{104}$$

$$\mathbf{\Sigma_{21}\Omega_{11}} = -\mathbf{\Sigma_{22}\Omega_{21}} \tag{105}$$

$$\mathbf{\Omega_{11}} = -\mathbf{\Sigma_{21}}^{-1}\mathbf{\Sigma_{22}\Omega_{21}}. \tag{106}$$

Now consider equation (100):

$$\mathbf{\Sigma_{11}\Omega_{11}} + \mathbf{\Sigma_{12}\Omega_{21}} = \mathbf{I} \tag{107}$$

$$\mathbf{\Omega_{11}} + \mathbf{\Sigma_{11}}^{-1}\mathbf{\Sigma_{12}\Omega_{21}} = \mathbf{\Sigma_{11}}^{-1}. \tag{108}$$

Combining equations (106) and (108), we have:

$$-\mathbf{\Sigma_{21}}^{-1}\mathbf{\Sigma_{22}\Omega_{21}} + \mathbf{\Sigma_{11}}^{-1}\mathbf{\Sigma_{12}\Omega_{21}} = \mathbf{\Sigma_{11}}^{-1} \tag{109}$$

$$(-\mathbf{\Sigma_{21}}^{-1}\mathbf{\Sigma_{22}} + \mathbf{\Sigma_{11}}^{-1}\mathbf{\Sigma_{12}})\mathbf{\Omega_{21}} = \mathbf{\Sigma_{11}}^{-1} \tag{110}$$

$$\mathbf{\Omega_{21}} = (-\mathbf{\Sigma_{21}}^{-1}\mathbf{\Sigma_{22}} + \mathbf{\Sigma_{11}}^{-1}\mathbf{\Sigma_{12}})^{-1}\mathbf{\Sigma_{11}}^{-1}. \tag{111}$$

Now, plug equation (111) into equation (106):

$$\mathbf{\Omega_{11}} = -\mathbf{\Sigma_{21}}^{-1}\mathbf{\Sigma_{22}}(-\mathbf{\Sigma_{21}}^{-1}\mathbf{\Sigma_{22}} + \mathbf{\Sigma_{11}}^{-1}\mathbf{\Sigma_{12}})^{-1}\mathbf{\Sigma_{11}}^{-1}.$$

Similar to before, consider equation (101):

$$\mathbf{\Sigma_{11}\Omega_{12}} + \mathbf{\Sigma_{12}\Omega_{22}} = \mathbf{0} \tag{112}$$

$$\mathbf{\Sigma_{12}\Omega_{22}} = -\mathbf{\Sigma_{11}\Omega_{12}} \tag{113}$$

$$\mathbf{\Omega_{22}} = -\mathbf{\Sigma_{12}}^{-1}\mathbf{\Sigma_{11}\Omega_{12}}. \tag{114}$$

Now consider equation (103):

$$\mathbf{\Sigma_{21}\Omega_{12}} + \mathbf{\Sigma_{22}\Omega_{22}} = \mathbf{I} \tag{115}$$

$$\mathbf{\Sigma_{22}}^{-1}\mathbf{\Sigma_{21}\Omega_{12}} + \mathbf{\Omega_{22}} = \mathbf{\Sigma_{22}}^{-1}. \tag{116}$$

Combining equations (114) and (116):

$$\mathbf{\Sigma_{22}}^{-1}\mathbf{\Sigma_{21}\Omega_{12}} - \mathbf{\Sigma_{12}}^{-1}\mathbf{\Sigma_{11}\Omega_{12}} = \mathbf{\Sigma_{22}}^{-1} \tag{117}$$

$$(\mathbf{\Sigma_{22}}^{-1}\mathbf{\Sigma_{21}} - \mathbf{\Sigma_{12}}^{-1}\mathbf{\Sigma_{11}})\mathbf{\Omega_{12}} = \mathbf{\Sigma_{22}}^{-1} \tag{118}$$

$$\mathbf{\Omega_{12}} = (\mathbf{\Sigma_{22}}^{-1}\mathbf{\Sigma_{21}} - \mathbf{\Sigma_{12}}^{-1}\mathbf{\Sigma_{11}})^{-1}\mathbf{\Sigma_{22}}^{-1}. \tag{119}$$

Plugging equation (119) into (114):

$$\mathbf{\Omega_{22}} = -\mathbf{\Sigma_{12}}^{-1}\mathbf{\Sigma_{11}}(\mathbf{\Sigma_{22}}^{-1}\mathbf{\Sigma_{21}} - \mathbf{\Sigma_{12}}^{-1}\mathbf{\Sigma_{11}})^{-1}\mathbf{\Sigma_{22}}^{-1}.$$

All together we have:

$$\mathbf{\Omega_{11}} = -\mathbf{\Sigma_{21}}^{-1}\mathbf{\Sigma_{22}}(-\mathbf{\Sigma_{21}}^{-1}\mathbf{\Sigma_{22}} + \mathbf{\Sigma_{11}}^{-1}\mathbf{\Sigma_{12}})^{-1}\mathbf{\Sigma_{11}}^{-1} \tag{120}$$

$$\mathbf{\Omega_{12}} = (\mathbf{\Sigma_{22}}^{-1}\mathbf{\Sigma_{21}} - \mathbf{\Sigma_{12}}^{-1}\mathbf{\Sigma_{11}})^{-1}\mathbf{\Sigma_{22}}^{-1} \tag{121}$$

$$\mathbf{\Omega_{21}} = (-\mathbf{\Sigma_{21}}^{-1}\mathbf{\Sigma_{22}} + \mathbf{\Sigma_{11}}^{-1}\mathbf{\Sigma_{12}})^{-1}\mathbf{\Sigma_{11}}^{-1} \tag{122}$$

$$\mathbf{\Omega_{22}} = -\mathbf{\Sigma_{12}}^{-1}\mathbf{\Sigma_{11}}(\mathbf{\Sigma_{22}}^{-1}\mathbf{\Sigma_{21}} - \mathbf{\Sigma_{12}}^{-1}\mathbf{\Sigma_{11}})^{-1}\mathbf{\Sigma_{22}}^{-1}. \tag{123}$$

Note, there are many different parameterizations that are all equivalent.

(C) Derive the conditional distribution $\mathbf{x_1}$, given $\mathbf{x_2}$, in terms of the partitioned elements of $\mathbf{x}$, $\boldsymbol{\mu}$, and $\mathbf{\Sigma}$. Explain briefly how one may interpret this conditional distribution as a linear regression on $\mathbf{x_2}$, where the regression matrix can be read off the precision matrix.

We know that

$$p(\mathbf{x_1}, \mathbf{x_2}) = \left(\frac{1}{\sqrt{2\pi}}\right)^p |\mathbf{\Sigma}|^{-\frac{1}{2}} e^{\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

and from part (A) of this section we know:

$$p(\mathbf{x_2}) = \left(\frac{1}{\sqrt{2\pi}}\right)^p |\mathbf{\Sigma_{22}}|^{-\frac{1}{2}} e^{\frac{1}{2}(\mathbf{x_2}-\boldsymbol{\mu_2})^T(\mathbf{\Sigma_{22}})^{-1}(\mathbf{x_2}-\boldsymbol{\mu_2})}$$

We will use the $\mathbf{\Omega}$ notation from part(B), where $\mathbf{\Sigma}^{-1} = \mathbf{\Omega}$. Then:

$$p(\mathbf{x_1}|\mathbf{x_2}) = \frac{p(\mathbf{x_1}, \mathbf{x_2})}{p(\mathbf{x_2})} \tag{124}$$

$$= \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^p |\mathbf{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}}{\left(\frac{1}{\sqrt{2\pi}}\right)^p |\mathbf{\Sigma_{22}}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x_2}-\boldsymbol{\mu_2})^T (\mathbf{\Sigma_{22}})^{-1}(\mathbf{x_2}-\boldsymbol{\mu_2})}} \tag{125}$$

$$\propto e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{126}$$

$$= \exp\left(-\frac{1}{2}\begin{pmatrix}\mathbf{x_1}-\boldsymbol{\mu_1}\\ \mathbf{x_2}-\boldsymbol{\mu_2}\end{pmatrix}^T \begin{pmatrix}\mathbf{\Sigma_{11}} & \mathbf{\Sigma_{12}}\\ \mathbf{\Sigma_{21}} & \mathbf{\Sigma_{22}}\end{pmatrix}^{-1} \begin{pmatrix}\mathbf{x_1}-\boldsymbol{\mu_1}\\ \mathbf{x_2}-\boldsymbol{\mu_2}\end{pmatrix}\right) \tag{127}$$

$$= \exp\left(-\frac{1}{2}\begin{pmatrix}\mathbf{x_1}-\boldsymbol{\mu_1}\\ \mathbf{x_2}-\boldsymbol{\mu_2}\end{pmatrix}^T \begin{pmatrix}\mathbf{\Omega_{11}} & \mathbf{\Omega_{12}}\\ \mathbf{\Omega_{21}} & \mathbf{\Omega_{22}}\end{pmatrix} \begin{pmatrix}\mathbf{x_1}-\boldsymbol{\mu_1}\\ \mathbf{x_2}-\boldsymbol{\mu_2}\end{pmatrix}\right) \tag{128}$$

$$\begin{aligned}&= \exp\Bigg(-\frac{1}{2}\Big((\mathbf{x_1}-\boldsymbol{\mu_1})^T \mathbf{\Omega_{11}}(\mathbf{x_1}-\boldsymbol{\mu_1}) + 2(\mathbf{x_1}-\boldsymbol{\mu_1})^T \mathbf{\Omega_{12}}(\mathbf{x_2}-\boldsymbol{\mu_2})\\ &\quad + (\mathbf{x_2}-\boldsymbol{\mu_2})^T \mathbf{\Omega_{22}}(\mathbf{x_2}-\boldsymbol{\mu_2})\Big)\Bigg)\end{aligned} \tag{129}$$

$$\propto \exp\Bigg(-\frac{1}{2}\Big((\mathbf{x_1}-\boldsymbol{\mu_1})^T \mathbf{\Omega_{11}}(\mathbf{x_1}-\boldsymbol{\mu_1}) + 2(\mathbf{x_1}-\boldsymbol{\mu_1})^T \mathbf{\Omega_{12}}(\mathbf{x_2}-\boldsymbol{\mu_2})\Big)\Bigg) \tag{130}$$

$$\begin{aligned}&= \exp\Bigg(-\frac{1}{2}\Big(\mathbf{x_1}^T \mathbf{\Omega_{11}}\mathbf{x_1} - 2\mathbf{x_1}^T \mathbf{\Omega_{11}}\boldsymbol{\mu_1} + \boldsymbol{\mu_1}^T \mathbf{\Omega_{11}}\boldsymbol{\mu_1}\\ &\quad + 2(\mathbf{x_1}^T \mathbf{\Omega_{12}}\mathbf{x_2} - \mathbf{x_1}^T \mathbf{\Omega_{12}}\boldsymbol{\mu_2} + \boldsymbol{\mu_1}^T \mathbf{\Omega_{12}}\mathbf{x_2} + \boldsymbol{\mu_1}^T \mathbf{\Omega_{12}}\boldsymbol{\mu_2})\Big)\Bigg)\end{aligned} \tag{131}$$

$$\begin{aligned}&\propto \exp\Bigg(-\frac{1}{2}\Big(\mathbf{x_1}^T \mathbf{\Omega_{11}}\mathbf{x_1} - 2\mathbf{x_1}^T \mathbf{\Omega_{11}}\boldsymbol{\mu_1}\\ &\quad + 2(\mathbf{x_1}^T \mathbf{\Omega_{12}}\mathbf{x_2} - \mathbf{x_1}^T \mathbf{\Omega_{12}}\boldsymbol{\mu_2})\Big)\Bigg)\end{aligned} \tag{132}$$

$$= \exp\Bigg(-\frac{1}{2}\Big(\mathbf{x_1}^T \mathbf{\Omega_{11}}\mathbf{x_1} - 2\mathbf{x_1}^T(\mathbf{\Omega_{11}}\boldsymbol{\mu_1} - \mathbf{\Omega_{12}}(\mathbf{x_2}-\boldsymbol{\mu_2})\mathbf{x_1})\Big)\Bigg) \tag{133}$$

$$= \exp\Bigg(-\frac{1}{2}\Big(\mathbf{x_1}^T \mathbf{\Omega_{11}}\mathbf{x_1} - 2\mathbf{x_1}^T \mathbf{\Omega_{11}}(\boldsymbol{\mu_1} - \mathbf{\Omega_{11}}^{-1}\mathbf{\Omega_{12}}(\mathbf{x_2}-\boldsymbol{\mu_2})\mathbf{x_1})\Big)\Bigg) \tag{134}$$

$$\propto \exp\Bigg(-\frac{1}{2}\Big(\mathbf{x_1} - (\boldsymbol{\mu_1} - \mathbf{\Omega_{11}}^{-1}\mathbf{\Omega_{12}}(\mathbf{x_2}-\boldsymbol{\mu_2}))\Big)^T \mathbf{\Omega_{11}}\Big(\mathbf{x_1} - (\boldsymbol{\mu_1} - \mathbf{\Omega_{11}}^{-1}\mathbf{\Omega_{12}}(\mathbf{x_2}-\boldsymbol{\mu_2}))\Big)\Bigg) \tag{135}$$

This is a multivariate normal distribution with mean $\boldsymbol{\mu_1} - \boldsymbol{\Omega_{11}}^{-1}\boldsymbol{\Omega_{12}}(\mathbf{x_2} - \boldsymbol{\mu_2})$ and variance $\boldsymbol{\Omega_{11}}^{-1}$.

This can be viewed as a regression, where

$$\mathbf{x_1} = \boldsymbol{\mu_1} + \beta(\mathbf{x_2} - \boldsymbol{\mu_2}) + \boldsymbol{\varepsilon} \;\; \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Omega_{11}}^{-1})$$

where $\boldsymbol{\Omega_{11}}^{-1}\boldsymbol{\Omega_{12}}$ is the estimate for β.

To express this distribution in terms of $\boldsymbol{\Sigma}$, recall part (B), where we showed that:

$$\boldsymbol{\Omega_{11}} = -\boldsymbol{\Sigma_{21}}^{-1}\boldsymbol{\Sigma_{22}}(-\boldsymbol{\Sigma_{21}}^{-1}\boldsymbol{\Sigma_{22}} + \boldsymbol{\Sigma_{11}}^{-1}\boldsymbol{\Sigma_{12}})^{-1}\boldsymbol{\Sigma_{11}}^{-1} \tag{136}$$

$$\boldsymbol{\Omega_{12}} = (\boldsymbol{\Sigma_{22}}^{-1}\boldsymbol{\Sigma_{21}} - \boldsymbol{\Sigma_{12}}^{-1}\boldsymbol{\Sigma_{11}})^{-1}\boldsymbol{\Sigma_{22}}^{-1} \tag{137}$$

We can rearrange these equations to get the following:

$$\boldsymbol{\Omega_{11}} = (\boldsymbol{\Sigma_{11}} - \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}}^{-1}\boldsymbol{\Sigma_{21}})^{-1} \tag{138}$$

$$\boldsymbol{\Omega_{12}} = -(\boldsymbol{\Sigma_{11}} - \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}}^{-1}\boldsymbol{\Sigma_{21}})^{-1}\boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}}^{-1} \tag{139}$$

Plugging in for $\boldsymbol{\Omega_{11}}$ and $\boldsymbol{\Omega_{12}}$, we get that:

$$\mathbf{x_1}|\mathbf{x_2} \sim N(\boldsymbol{\mu_1} + \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}}^{-1}(\mathbf{x_2} - \boldsymbol{\mu_2}), \boldsymbol{\Sigma_{11}} - \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}}^{-1}\boldsymbol{\Sigma_{21}})$$

4 Multiple regression: three classical principles for inference

Suppose we observe data that we believe to follow a linear model, where $y_i = \mathbf{x_i}^T\boldsymbol{\beta} + \epsilon_i$, for $i = 1, \ldots, n$.

(A) Show that these three principles lead to the same estimator.

Least Squares: make the sum of squared errors as small as possible.

$$\hat{\boldsymbol{\beta}} = \text{arg min}_{\beta \in R^p} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x_i}^T\boldsymbol{\beta})^2 \right\}$$

To find

$$\text{arg min}_{\beta \in R^p} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x_i}^T\boldsymbol{\beta})^2 \right\}$$

take the derivative with respect to $\boldsymbol{\beta}$ and set it equal to zero.

$$\frac{\delta}{\delta\boldsymbol{\beta}} \left(\sum_{i=1}^{n} (y_i - \mathbf{x_i}^T\boldsymbol{\beta})^2 \right) = 0 \tag{140}$$

$$\frac{\delta}{\delta\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^2 = 0 \tag{141}$$

$$-2\mathbf{X}^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \tag{142}$$

$$-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = 0 \tag{143}$$

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y} \tag{144}$$

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{145}$$

Take a second derivative to assure this is a minimum:

$$\frac{\delta^2}{\delta\boldsymbol{\beta}^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^2 = \frac{\delta}{\delta\boldsymbol{\beta}}(-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}) = 2\mathbf{X}^T\mathbf{X}$$

Since $\mathbf{X}^T\mathbf{X}$ is always positive, this is a minimum. So, our estimator $\hat{\boldsymbol{\beta}}$ is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Maximum likelihood under Gaussianity: assume that the errors are independent, mean-zero normal random variables with common variance σ^2, choose $\hat{\boldsymbol{\beta}}$ to maximize the likelihood:

$$\hat{\boldsymbol{\beta}} = \text{arg max}_{\beta \in R^p} \left\{ \prod_{i=1}^{n} p(y_i|\boldsymbol{\beta}, \sigma^2) \right\}$$

Since log is a monotone function, maximizing $f(x)$ is equivalent to maximizing $\log(f(x))$, so to find

$$\hat{\boldsymbol{\beta}} = \arg\,\max_{\beta \in R^p} \left\{ \prod_{i=1}^{n} p(y_i|\boldsymbol{\beta}, \sigma^2) \right\}$$

take the derivative with respect to $\boldsymbol{\beta}$ of

$$\log \left(\prod_{i=1}^{n} p(y_i|\boldsymbol{\beta}, \sigma^2) \right)$$

and set it equal to zero.

$$\frac{\delta}{\delta\boldsymbol{\beta}} \log \left(\prod_{i=1}^{n} p(y_i|\boldsymbol{\beta}, \sigma^2) \right) = 0 \tag{146}$$

$$\frac{\delta}{\delta\boldsymbol{\beta}} \log \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mathbf{x_i}^T\boldsymbol{\beta})^2}{2\sigma^2}} \right) = 0 \tag{147}$$

$$\frac{\delta}{\delta\boldsymbol{\beta}} \log \left(\left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x_i}^T\boldsymbol{\beta})^2} \right) = 0 \tag{148}$$

$$\frac{\delta}{\delta\boldsymbol{\beta}} \left(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x_i}^T\boldsymbol{\beta})^2 \right) = 0 \tag{149}$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbf{x_i}(y_i - \mathbf{x_i}^T\boldsymbol{\beta}) = 0 \tag{150}$$

$$\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = 0 \tag{151}$$

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y} \tag{152}$$

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{153}$$

Take the second derivative to ensure this is a maximum:

$$\begin{aligned} \frac{\delta^2}{\delta\boldsymbol{\beta}^2} \log \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mathbf{x_i}^T\boldsymbol{\beta})^2}{2\sigma^2}} \right) &= \frac{\delta}{\delta\boldsymbol{\beta}} \left(\frac{1}{\sigma^2}(\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \right) \\ &= \frac{1}{\sigma^2}(-\mathbf{X}^T\mathbf{X}) \end{aligned}$$

Since, $\mathbf{X}^T\mathbf{X}$ is always positive, the second derivative is always negative, therefore this is a maximum, and our estimator $\hat{\boldsymbol{\beta}}$ is once again:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Method of Moments: Choose $\hat{\boldsymbol{\beta}}$ so that the sample covariance between the errors and each of the p predictors is exactly zero. Let e_i denote the errors and

$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i$, and let $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$.Then the method of moments generates the following equation for each predictor j.

$$\sum_{i=1}^{n}(e_i - \bar{e})(x_{ij} - \bar{x}_j) = 0$$

Let us assume without loss of generality that $\bar{x}_j$ is zero for each predictor. Practically this can be achieved by centering each predictor on its mean. Then the equations become:

$$\sum_{i=1}^{n}(e_i - \bar{e})x_{ij} = 0 \tag{154}$$

$$\sum_{i=1}^{n} e_i x_{ij} - \bar{e}\sum_{i=1}^{n} x_{ij} = 0 \tag{155}$$

$$\sum_{i=1}^{n} e_i x_{ij} - n\bar{e}\bar{x}_j = 0 \tag{156}$$

$$\sum_{i=1}^{n} e_i x_{ij} = 0 \tag{157}$$

Translating this into matrix notation to represent the entire set of equations simultaneously, we have:

$$\mathbf{e}^T\mathbf{X} = 0 \tag{158}$$

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{X} = 0 \tag{159}$$

$$\mathbf{y}^T\mathbf{X} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X} = 0 \tag{160}$$

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y} \tag{161}$$

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{162}$$

So, once again we get the estimator:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

(B) Now suppose you trust some observations more than others, and will estimate $\boldsymbol{\beta}$ by minimizing the weighted sum of squared errors,

$$\hat{\boldsymbol{\beta}} = \arg\ \min_{\beta \in R^p}\left\{\sum_{i=1}^{n} w_i(y_i - \mathbf{x_i}^T\boldsymbol{\beta})^2\right\}$$

where the w_i are weights. Derive the estimator, and show that is corresponds to the maximum-likelihood solution under heteroscedastic Gaussian error:

$$\hat{\boldsymbol{\beta}} = \arg\ \max_{\beta \in R^p}\left\{\prod_{i=1}^{n} p(y_i|\boldsymbol{\beta}, \sigma_i^2)\right\}$$

Make sure to explicitly connect the weights w_i and the idiosyncratic variances σ_i^2.

Let

$$\mathbf{W} = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix}$$

Then,

$$\text{arg min}_{\beta \in R^p} \left\{ \sum_{i=1}^{n} w_i (y_i - \mathbf{x_i}^T \boldsymbol{\beta})^2 \right\} = \text{arg min}_{\beta \in R^p} \left\{ \mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^2 \right\}.$$

As before, take the derivative and set it equal to zero,

$$\frac{\delta}{\delta \boldsymbol{\beta}} \left(\mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^2 \right) = 0 \tag{163}$$

$$-2\mathbf{X}^T \mathbf{W} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right) = 0 \tag{164}$$

$$-2\mathbf{X}^T \mathbf{W} \mathbf{y} + 2\mathbf{X}^T \mathbf{W} \mathbf{X} \boldsymbol{\beta} = 0 \tag{165}$$

$$\mathbf{X}^T \mathbf{W} \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{W} \mathbf{y} \tag{166}$$

$$\boldsymbol{\beta} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y} \tag{167}$$

The second derivative is $2\mathbf{X}^T \mathbf{W} \mathbf{X}$, which is positive, so this is a minimum, and:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}$$

Now let,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

Then,

$$\text{arg max}_{\beta \in R^p} \left\{ \prod_{i=1}^{n} p(y_i | \boldsymbol{\beta}, \sigma_i^2) \right\} = \text{arg max}_{\beta \in R^p} \left\{ \left(\frac{1}{\sqrt{2\pi}} \right)^n |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})} \right\}.$$

As before, we will work with the derivative of the log:

$$\frac{\delta}{\delta\boldsymbol{\beta}}\log\left(\left(\frac{1}{\sqrt{2\pi}}\right)^{n}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}e^{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{T}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}\right)=0 \tag{168}$$

$$\frac{\delta}{\delta\boldsymbol{\beta}}(-\frac{n}{2}\log(2\pi)-\frac{1}{2}\log(|\boldsymbol{\Sigma}|)-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{T}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}))=0 \tag{169}$$

$$\frac{\delta}{\delta\boldsymbol{\beta}}-\frac{1}{2}(\mathbf{y}^{T}\boldsymbol{\Sigma}^{-1}\mathbf{y}-2\mathbf{y}^{T}\boldsymbol{\Sigma}^{-1}\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\beta}^{T}\mathbf{X}^{T}\boldsymbol{\Sigma}^{-1}\mathbf{X}\boldsymbol{\beta})=0 \tag{170}$$

$$\mathbf{y}^{T}\boldsymbol{\Sigma}^{-1}\mathbf{X}-\frac{1}{2}(\mathbf{X}^{T}\boldsymbol{\Sigma}^{-1}\mathbf{X}+(\mathbf{X}^{T}\boldsymbol{\Sigma}^{-1}\mathbf{X})^{T})\boldsymbol{\beta}=0 \tag{171}$$

$$\mathbf{X}^{T}\boldsymbol{\Sigma}^{-1}\mathbf{X}\boldsymbol{\beta}=2\mathbf{X}^{T}\boldsymbol{\Sigma}^{-1}\mathbf{y} \tag{172}$$

$$(\mathbf{X}^{T}\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^{T}\boldsymbol{\Sigma}^{-1}\mathbf{y}=\boldsymbol{\beta} \tag{173}$$

The second derivative is $-\mathbf{X}^{T}\boldsymbol{\Sigma}^{-1}\mathbf{X}$, which is always negative, so this is a maximum, and:

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}^{T}\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^{T}\boldsymbol{\Sigma}^{-1}\mathbf{y}.$$

And, $\mathbf{W}=\boldsymbol{\Sigma}^{-1}$, so $w_i=\frac{1}{\sigma_i^2}$.

5 Quantifying uncertainty: some basic frequentist ideas

Suppose that we observe data from a linear regression model with Gaussian error:

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}\ \ ,\ \ \boldsymbol{\epsilon}\sim N(0,\sigma^{2}\mathbf{I})$$

(A) Derive the sampling distribution of your estimator for $\boldsymbol{\beta}$ from the previous problem.

We know that:

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}\ \ ,\ \ \boldsymbol{\epsilon}\sim N(\mathbf{0},\sigma^{2}\mathbf{I}).$$

In other words:

$$\mathbf{y}|\mathbf{X}\sim N(\mathbf{X}\boldsymbol{\beta},\sigma^{2}\mathbf{I}).$$

We have shown previously that

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{y}$$

So, $\hat{\boldsymbol{\beta}}$ is an affine transformation of $\mathbf{y}$, and is therefore normally distributed. So, it suffices to find the mean and variance of $\hat{\boldsymbol{\beta}}$.

Mean:

$$E[\hat{\boldsymbol{\beta}}]=E[(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{y}]=(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T}E[\mathbf{y}]=(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{X}\boldsymbol{\beta}=\boldsymbol{\beta}$$

Variance:

$$
\begin{aligned}
Var(\hat{\boldsymbol{\beta}}) &= Var((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) && (174)\\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Var(\mathbf{y})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} && (175)\\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} && (176)\\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} && (177)
\end{aligned}
$$

So,

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}).$$

(B) The sampling distribution depends on σ^2, yet this is unknown. Suppose that you still wanted to quantify your uncertainty about the individual regression coefficients. Propose a strategy for calculating standard errors for each β_j.

First I would calculate $\hat{\boldsymbol{\beta}}$ as $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. Then I would use the residual sum of squares (RSS), as my estimate of σ^2, where

$$RSS = \frac{\sum_{i=1}^{n}(y_i - \mathbf{x_i}^T\hat{\boldsymbol{\beta}})^2}{n-p}$$

p being the number of predictors. We calculate the distribution of $\hat{\boldsymbol{\beta}}$ in part (A). Using the variance from that, and a previous result, it follows that the variance for β_j is $RSS * (\mathbf{X}^T\mathbf{X})^{-1}_{jj}$. Below are the results of my calculations versus those from the lm function in R.

Variable	R se	my se
x	38.3	38.3
xV5	0.00725	0.00725
xV6	0.174	0.174
xV7	0.0238	0.0238
xV8	0.0693	0.0693
xV9	0.125	0.125
xV10	0.000394	0.000394
xV11	0.0148	0.0148
xV12	0.119	0.119
xV13	0.00490	0.00490

6 Propagating Uncertainty

Suppose you have taken data and estimated some parameters $\theta_1, \ldots, \theta_p$ of a multivariate statistical model–for example, the regression model of the previous problem. Call your estimate $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)^T$. Suppose that you also have an estimate of the covariance matrix of the sampling distribution $\hat{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{\Sigma}} \approx E\left\{(\hat{\boldsymbol{\theta}} - \bar{\theta})(\hat{\boldsymbol{\theta}} - \bar{\theta})^T\right\}$$

where the expectation is under the sampling distribution for the data, given the true parameter $\hat{\boldsymbol{\theta}}$. Here $\bar{\theta}$ denotes the mean of the sampling distribution.

If you want to report uncertainty about the $\hat{\theta}_j$'s, you can do so by peeling off the diagonal of the estimate covariance: $\hat{\Sigma}_{jj} = \hat{\sigma}_j^2$ is the square of the ordinary standard error of $\hat{\theta}_j$ But what if you want to report uncertainty about some function involving multiple components of the estimate $\hat{\boldsymbol{\theta}}$?

(A) Start with the trivial case where you want to estimate

$$f(\boldsymbol{\theta}) = \theta_1 + \theta_2.$$

Calculate the standard error for $f(\hat{\boldsymbol{\theta}})$, and generalize this to the case where f is the sum of all p components of $\hat{\boldsymbol{\theta}}$.

Let us begin by writing $f(\hat{\boldsymbol{\theta}})$ in matrix notation:

$$f(\hat{\boldsymbol{\theta}}) = \hat{\theta}_1 + \hat{\theta}_2 = (1, 1, 0, \ldots, 0)\hat{\boldsymbol{\theta}}.$$

Then, calculate the variance

$$\begin{aligned}
Var(f(\hat{\boldsymbol{\theta}})) &= Var((1, 1, 0, \ldots, 0)\hat{\boldsymbol{\theta}}) && (178)\\
&= (1, 1, 0, \ldots, 0)Var(\hat{\boldsymbol{\theta}})(1, 1, 0, \ldots, 0)^T && (179)\\
&= (1, 1, 0, \ldots, 0)\hat{\boldsymbol{\Sigma}}(1, 1, 0, \ldots, 0)^T && (180)\\
&= \hat{\Sigma}_{11} + \hat{\Sigma}_{12} + \hat{\Sigma}_{21} + \hat{\Sigma}_{22} && (181)\\
&= \hat{\Sigma}_{11} + 2\hat{\Sigma}_{12} + \hat{\Sigma}_{22}. && (182)
\end{aligned}$$

Then the standard error for $\theta_1 + \theta_2$ is $\sqrt{\hat{\Sigma}_{11} + 2\hat{\Sigma}_{12} + \hat{\Sigma}_{22}}$. Expanding upon this, consider $f(\boldsymbol{\theta}) = \sum_{i=1}^{p} \theta_i$. Then,

$$\begin{aligned}
Var(f(\hat{\boldsymbol{\theta}})) &= Var(\sum_{i=1}^{p} \hat{\theta}_i) && (183)\\
&= Var((1, \ldots, 1)\hat{\boldsymbol{\theta}}) && (184)\\
&= (1, \ldots, 1)Var(\hat{\boldsymbol{\theta}})(1, \ldots, 1)^T && (185)\\
&= (1, \ldots, 1)\hat{\boldsymbol{\Sigma}}(1, \ldots, 1)^T && (186)\\
&= \sum_{j=1}^{p}\sum_{i=1}^{p} \hat{\Sigma}_{ij} && (187)
\end{aligned}$$

The standard error would just be the square root of the variance.

(B) What now if f is a nonlinear function of the $\hat{\theta}_j$'s? Propose an approximation for $var\{f(\hat{\boldsymbol{\theta}})\}$, where f is any sufficiently smooth function.

We are interested in $Var\left(f(\hat{\boldsymbol{\theta}})\right)$, where f is a nonlinear function of $\hat{\boldsymbol{\theta}}$ we have shown, in part (A), how to approximate the variance of a linear function g. Let us approximate f as a linear function of $\hat{\boldsymbol{\theta}}$ using a first order Taylor approximation, and calculate the variance of that.

$$f(\hat{\boldsymbol{\theta}}) \approx f(\hat{\boldsymbol{\theta}}) + f'(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$$

Where $f'(\hat{\boldsymbol{\theta}})$ is the vector of partial derivatives evaluated at $\hat{\boldsymbol{\theta}}$ So, the variance of $f(\hat{\boldsymbol{\theta}}$ is:

$$Var\left(f(\hat{\boldsymbol{\theta}})\right) = Var\left(f(\hat{\boldsymbol{\theta}}) + f'(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right) = f'(\hat{\boldsymbol{\theta}})^2 Var(\hat{\boldsymbol{\theta}}) = f'(\hat{\boldsymbol{\theta}})^2 \hat{\boldsymbol{\Sigma}}$$

The error of this approximation is bounded by the next term in the Taylor series approximation, namely $f''(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^2$. In general, first order Taylor series work well when the area of the curve you are trying to approximate if fairly linear, the more curved it is, the worse this approximation will be.

7 Bootstrapping

(A) Let $\hat{\boldsymbol{\Sigma}}$ denote the covariance matrix of the sampling distribution of $\hat{\boldsymbol{\beta}}$, the least-squares estimator. Write an R function that will estimate $\hat{\boldsymbol{\Sigma}}$ via bootstrapped resampling for a given design matrix $\mathbf{X}$ and response vector $\mathbf{y}$. Use it to compute $\hat{\boldsymbol{\Sigma}}$ for the ozone data set, and compare it to the parametric estimate based on normal theory.

As discussed in class, there are several ways that you can draw a bootstrap sample:

- *Sampling xy pairs*: The first, and to me most intuitive way, is to resample xy pairs. So, for a sample of size N, you would sample N xy pairs with replacement, and calculate $\hat{\boldsymbol{\beta}}$ with your sample. This would be repeated for a number of iterations, in my case I did 10,000, and then you would compute the covariance of the 10,000 betas you calculated. This method requires the fewest assumptions about the data.

- *Sampling residuals*: You can also keep the xy pairs fixed and resample the residuals. So, here you would estimate $\hat{\boldsymbol{\beta}}$ using all of the data, and then calculate the estimate of the residuals. Then you would sample from the residuals and recalculate $\hat{\mathbf{y}} = \mathbf{y} + \hat{\boldsymbol{\epsilon}}$. As before, you would calculate a $\hat{\boldsymbol{\beta}}$ at each iteration. Then compute the variance of your sample of $\hat{\boldsymbol{\beta}}$'s.

- *Parametric* This technique explicitly uses the distribution calculated earlier, namely $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$. So, you could estimate $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$, and then sample from $N(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1})$. I did not implement this as I used $\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$ as my asymptotic estimate of the variance to compare with the other methods.

Below are the estimates of the covariance I got for bootstrap samples of $10,000$.

Asymptotic Estimate

	V1	V5	V6	V7	V8	V9	V10	V11	V12	V13
	1396.72	-0.26	-1.96	-0.15	0.34	1.51	0.00	-0.04	0.40	-0.00
V5	-0.26	0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00	-0.00	-0.00
V6	-1.96	0.00	0.03	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
V7	-0.15	0.00	-0.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	0.00
V8	0.34	-0.00	-0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00
V9	1.51	-0.00	-0.00	-0.00	-0.00	0.01	-0.00	0.00	-0.01	0.00
V10	0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00	0.00	0.00	-0.00
V11	-0.04	0.00	-0.00	-0.00	-0.00	0.00	0.00	0.00	0.00	-0.00
V12	0.40	-0.00	-0.00	-0.00	-0.00	-0.01	0.00	0.00	0.01	-0.00
V13	-0.00	-0.00	-0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.00	0.00

Bootstrap sampling xy pairs

	V1	V5	V6	V7	V8	V9	V10	V11	V12	V13
	1253.06	-0.24	-2.33	-0.04	0.39	1.53	0.00	-0.07	0.19	-0.00
V5	-0.24	0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00	-0.00	-0.00
V6	-2.33	0.00	0.03	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
V7	-0.04	0.00	-0.00	0.00	0.00	-0.00	0.00	-0.00	0.00	0.00
V8	0.39	-0.00	-0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	0.00
V9	1.53	-0.00	-0.00	-0.00	-0.00	0.01	-0.00	0.00	-0.01	0.00
V10	0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00	0.00	0.00	-0.00
V11	-0.07	0.00	-0.00	-0.00	-0.00	0.00	0.00	0.00	0.00	0.00
V12	0.19	-0.00	-0.00	0.00	-0.00	-0.01	0.00	0.00	0.01	-0.00
V13	-0.00	-0.00	-0.00	0.00	0.00	0.00	-0.00	0.00	-0.00	0.00

Bootstrap sampling residuals

	V1	V5	V6	V7	V8	V9	V10	V11	V12	V13
	1387.61	-0.26	-1.88	-0.13	0.37	1.44	0.00	-0.05	0.40	-0.01
V5	-0.26	0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00	-0.00	0.00
V6	-1.88	0.00	0.03	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	-0.00
V7	-0.13	0.00	-0.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	0.00
V8	0.37	-0.00	-0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00
V9	1.44	-0.00	-0.00	-0.00	-0.00	0.01	-0.00	0.00	-0.01	0.00
V10	0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00	0.00	0.00	-0.00
V11	-0.05	0.00	-0.00	-0.00	-0.00	0.00	0.00	0.00	0.00	-0.00
V12	0.40	-0.00	0.00	-0.00	-0.00	-0.01	0.00	0.00	0.01	-0.00
V13	-0.01	0.00	-0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.00	0.00

(B) Write R functions that will accomplish the following:

1. For a specified mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, simulate multivariate normal random variables.
2. For a given sample $x_1, \ldots, x_n$ from a multivariate normal distribution, estimate the mean vector and covariance matrix by maximum likelihood.
3. Bootstrap a given sample $x_1, \ldots, x_n$ to estimate the sampling distribution of the MLE.

Please see R code.