SDS 383D: Homework 1

Giorgio Paulon

February 5, 2017

Problem 1. Bayesian inference in simple conjugate families

We start with a few of the simplest building blocks for complex multivariate statistical models: the beta/binomial, normal, and inverse- gamma conjugate families.

(A) *Suppose that we take independent observations $X_1, \dots, X_N$ 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:*

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

where $\Gamma(\cdot)$ denotes the Gamma function. Derive the posterior distribution $p(w|x_1, \dots, x_N)$.

The following model

$$\begin{aligned} X_1, \dots, X_n | w &\overset{\text{iid}}{\sim} Bernoulli(w) \\ w &\sim Beta(a, b) \end{aligned}$$

leads to the following posterior distribution:

$$\begin{aligned} p(w|\boldsymbol{x}) &\propto p(\boldsymbol{X}|w)p(w) \\ &= \prod_{i=1}^{N} p(X_i|w)p(w) \\ &\propto \prod_{i=1}^{N} \left\{ w^{x_i}(1-w)^{1-x_i} \right\} w^{a-1}(1-w)^{b-1}\mathcal{I}_{[0,1]}(w) \\ &= w^{\sum_{i=1}^N x_i}(1-w)^{n-\sum_{i=1}^N x_i} w^{a-1}(1-w)^{b-1}\mathcal{I}_{[0,1]}(w) \\ &= w^{a+\sum_{i=1}^N x_i - 1}(1-w)^{b+(n-\sum_{i=1}^N x_i)-1}\mathcal{I}_{[0,1]}(w). \end{aligned}$$

Therefore,

$$p(w|\boldsymbol{x}) = \frac{\Gamma(a+b+n)}{\Gamma(a+\sum_{i=1}^N x_i)\Gamma(b+(n-\sum_{i=1}^N x_i))} w^{a+\sum_{i=1}^N x_i - 1}(1-w)^{b+(n-\sum_{i=1}^N x_i)-1}\mathcal{I}_{[0,1]}(w)$$

that is,

$$W|\boldsymbol{x} \sim Beta\left(a + \sum_{i=1}^N x_i, b + (n - \sum_{i=1}^N x_i)\right).$$

This is the so-called **Bernoulli-Beta model**.

(B) *The probability density function (PDF) 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 $X_2 \sim Ga(a_2, 1)$. Define two new random variables $Y_1 = X_1/(X_1 + X_2)$ and $Y_2 = X_1 + X_2$. Find the joint density for (Y_1, Y_2) using a direct PDF transformation (and its Jacobian). 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.*

Let $X_1 \sim Ga(a_1, 1)$, $X_2 \sim Ga(a_2, 1)$ and let us define $(Y_1, Y_2) = g(X_1, X_2) = \Big(\frac{X_1}{X_1+X_2}, X_1 + X_2\Big)$. The joint density of (Y_1, Y_2) can be found via pdf transformation, i.e.

$$\begin{cases} y_1 = g_1(x_1, x_2) = \frac{x_1}{x_1+x_2} \\ y_2 = g_2(x_1, x_2) = x_1 + x_2 \end{cases} \Rightarrow \begin{cases} x_1 = g_1^{-1}(y_1, y_2) = y_1 y_2 \\ x_2 = g_2^{-1}(y_1, y_2) = y_2(1 - y_1). \end{cases}$$

The inverse transformation has a unique solution and therefore the mapping is one-to-one. Moreover, the domain $\mathcal{X}_1 \times \mathcal{X}_2 = [0, \infty)^2$ is mapped to $\mathcal{Y}_1 \times \mathcal{Y}_2 = [0, 1] \times [0, \infty)$.

The Jacobian of the transformation is

$$J = \begin{pmatrix} \dfrac{\partial g_1^{-1}(y_1, y_2)}{\partial y_1} & \dfrac{\partial g_1^{-1}(y_1, y_2)}{\partial y_2} \\ \dfrac{\partial g_2^{-1}(y_1, y_2)}{\partial y_1} & \dfrac{\partial g_2^{-1}(y_1, y_2)}{\partial y_2} \end{pmatrix} = \begin{pmatrix} y_2 & y_1 \\ -y_2 & 1 - y_1 \end{pmatrix}$$

and the absolute value of its determinant is $|J| = |y_2 - y_1 y_2 + y_1 y_2| = y_2$. The joint pdf of (X_1, X_2) is, since $X_1 \perp X_2$,

$$f_{(X_1,X_2)}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2) = \frac{1}{\Gamma(a_1)\Gamma(a_2)} x_1^{a_1-1} e^{-x_1} x_2^{a_2-1} e^{-x_2} \mathcal{I}_{[0,\infty)}(x_1) \mathcal{I}_{[0,\infty)}(x_2).$$

Therefore

$$\begin{aligned} f_{(Y_1,Y_2)}(y_1, y_2) &= f_{(X_1,X_2)}(g_1^{-1}(y_1, y_2), g_2^{-1}(y_1, y_2))|J| \\ &= \frac{1}{\Gamma(a_1)\Gamma(a_2)} y_1^{a_1-1} y_2^{a_1-1} e^{-y_1 y_2} y_2^{a_2-1} (1 - y_1)^{a_2-1} e^{-y_2(1-y_1)} \mathcal{I}_{[0,1]}(y_1) \mathcal{I}_{[0,\infty)}(y_2) y_2 \\ &= \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} y_1^{a_1-1} (1 - y_1)^{a_2-1} \mathcal{I}_{[0,1]}(y_1) \times \frac{1}{\Gamma(a_1 + a_2)} y_2^{a_1+a_2-1} e^{-y_2} \mathcal{I}_{[0,\infty)}(y_2). \end{aligned}$$

The joint density factors, and therefore we recognize that

$$f_{(Y_1,Y_2)}(y_1, y_2) = Beta(a_1, a_2) \times Gamma(a_1 + a_2, 1),$$

that is,

$$\begin{aligned} &Y_1 \sim Beta(a_1, a_2) \\ &Y_2 \sim Gamma(a_1 + a_2, 1) \\ &Y_1 \perp Y_2 \end{aligned}$$

Therefore, in order to sample $Y_1 \sim Beta(a_1, a_2)$ we can sample $X_1 \sim Gamma(a_1, 1)$, then $X_2 \sim Gamma(a_2, 1)$ independently. The resulting $Y_1 = \frac{X_1}{X_1+X_2}$ is a draw from $Beta(a_1, a_2)$. The Gamma samples can be obtained as a sum of a_1 exponential samples from $Exp(1)$. The samples from $Exp(1)$ can be obtained as $-\log(U_i)$, where $U_i \overset{\text{iid}}{\sim} \mathcal{U}(0, 1)$.

(C) *Suppose that we take independent observations $X_1, \dots, X_N$ from a normal sampling model with unknown mean θ and known variance σ^2: $X_i \overset{\text{iid}}{\sim} N(\theta, \sigma^2)$. Suppose that θ is given a normal prior distribution with mean m and variance v. Derive the posterior distribution $p(\theta|x_1, \dots, x_N)$.*

The following model

$$\begin{aligned} X_1, \dots, X_n | \theta &\overset{\text{iid}}{\sim} N(\theta, \sigma^2) \\ \theta &\sim N(m, v) \end{aligned}$$

leads to the following posterior distribution:

$$\begin{aligned} p(\theta|\boldsymbol{x}) &\propto p(\boldsymbol{X}|\theta)p(\theta) \\ &\propto \prod_{i=1}^{N} \left\{ \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} e^{-\frac{1}{2\sigma^2}(x_i-\theta)^2} \right\} \left(\frac{1}{2\pi v}\right)^{1/2} e^{-\frac{1}{2v}(\theta-m)^2} \\ &\propto e^{-\frac{1}{2\sigma^2}\sum_{i=1}^N (x_i-\theta)^2} e^{-\frac{1}{2v}(\theta-m)^2} \\ &= e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^N x_i^2 + n\theta^2 - 2\theta\sum_{i=1}^N x_i\right)} e^{-\frac{1}{2v}(\theta^2+m^2-2m\theta)} \\ &\propto e^{-\frac{n}{2\sigma^2}\left(\theta^2-2\overline{x}\theta\right)} e^{-\frac{1}{2v}(\theta^2-2m\theta)} \\ &= e^{-\frac{n}{2\sigma^2 v}\left(v\theta^2-2\overline{x}v\theta+\frac{\sigma^2}{n}\theta^2-2\frac{\sigma^2}{n}m\theta\right)} \\ &= e^{-\frac{n}{2\sigma^2 v}\left(v\theta^2-2\overline{x}v\theta+\frac{\sigma^2}{n}\theta^2-2\frac{\sigma^2}{n}m\theta\right)} \\ &= e^{-\frac{n}{2\sigma^2 v}\left[(v+\frac{\sigma^2}{n})\theta^2-2(\overline{x}v+\frac{\sigma^2}{n}m)\theta\right]} \\ &= e^{-\frac{nv+\sigma^2}{2\sigma^2 v}\left(\theta^2-2\frac{\overline{x}nv+\sigma^2 m}{nv+\sigma^2}\theta\right)} \\ &\propto e^{-\frac{nv+\sigma^2}{2\sigma^2 v}\left(\theta-\frac{\overline{x}nv+\sigma^2 m}{nv+\sigma^2}\theta\right)^2} \\ &\propto N\left(\frac{v}{v+\sigma^2/n}\overline{x} + \frac{\sigma^2/n}{v+\sigma^2/n}m; \left(\frac{n}{\sigma^2}+\frac{1}{v}\right)^{-1}\right) \end{aligned}$$

This is the so-called **Normal-Normal model for unknown mean and known variance**.

The precision is additive in Gaussian models: that is, the posterior precision is the sum of the prior precision $1/v$ and the data precision n/σ^2. The mean, moreover, is a weighted average of prior mean and of sample average, whose weights are the precisions related to them.

(D) *Suppose that we take independent observations $X_1, \dots, X_N$ from a normal sampling model with known mean θ but unknown variance σ^2 (this seems even more artificial than the last, but is conceptually important). To make this easier, we will re-express things in terms of the precision, or inverse variance $\omega = 1/\sigma^2$:*

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

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

The following model

$$\begin{aligned} &X_1, \dots, X_n | \sigma^2 \overset{\text{iid}}{\sim} N(\theta, \sigma^2) \\ &\sigma^2 \sim IG(a, b) \end{aligned}$$

can be rewritten in terms of the precision parameter $\omega = 1/\sigma^2$ as

$$\begin{aligned} &X_1, \dots, X_n | \omega \overset{\text{iid}}{\sim} N(\theta, \omega) \\ &\omega \sim Gamma(a, b). \end{aligned}$$

This model leads to the following posterior distribution:

$$\begin{aligned} p(\omega|\boldsymbol{x}) &\propto p(\boldsymbol{X}|\omega)p(\omega) \\ &\propto \prod_{i=1}^{N} \left\{ \left(\frac{\omega}{2\pi}\right)^{1/2} e^{-\frac{\omega}{2}(x_i-\theta)^2} \right\} \frac{b^a}{\Gamma(a)} \omega^{a-1} e^{-b\omega} \mathcal{I}_{[0,\infty)}(\omega) \\ &\propto \omega^{N/2} e^{-\frac{\omega}{2}\sum_{i=1}^{N}(x_i-\theta)^2} \omega^{a-1} e^{-b\omega} \mathcal{I}_{[0,\infty)}(\omega) \\ &= \omega^{a+N/2-1} e^{-\omega(b+\frac{1}{2}\sum_{i=1}^{N}(x_i-\theta)^2)} \mathcal{I}_{[0,\infty)}(\omega) \\ &\propto Gamma\left(a + N/2; b + \frac{1}{2}\sum_{i=1}^{N}(x_i-\theta)^2)\right). \end{aligned}$$

Therefore,

$$\begin{aligned} \omega|\boldsymbol{x} &\sim Gamma\left(a + N/2; b + \frac{1}{2}\sum_{i=1}^{N}(x_i-\theta)^2)\right) \\ \sigma^2|\boldsymbol{x} &\sim IG\left(a + N/2; b + \frac{1}{2}\sum_{i=1}^{N}(x_i-\theta)^2)\right) \end{aligned}$$

In terms of the pdf of σ^2, let us find the generic pdf of $IG(a,b)$. We know that, if $X \sim Gamma(a,b)$, then $Y = 1/X \sim IG(a,b)$. The transformation $g(x) = 1/x$ is monotone and its inverse is $g^{-1}(y) = 1/y$. The derivative of the inverse transformation is $\frac{\partial}{\partial y} g^{-1}(y) = -\frac{1}{y^2}$. Thus,

$$\begin{aligned} f_Y(y) &= f_X(1/y) \left| \frac{\partial}{\partial y} g^{-1}(y) \right| \\ &= \frac{b^a}{\Gamma(a)} \left(\frac{1}{y}\right)^{a-1} e^{-\frac{b}{y}} \frac{1}{y^2} \mathcal{I}_{[0,\infty)}(y) \\ &= \frac{b^a}{\Gamma(a)} y^{-a-1} e^{-\frac{b}{y}} \mathcal{I}_{[0,\infty)}(y). \end{aligned}$$

Therefore, if $\sigma^2|\boldsymbol{x} \sim IG\left(a + N/2; b + \frac{1}{2}\sum_{i=1}^{N}(x_i-\theta)^2\right)$, the pdf is

$$f_{\sigma^2|\boldsymbol{x}}(\sigma^2|\boldsymbol{x}) = \frac{(b + \frac{1}{2}\sum_{i=1}^{N}(x_i-\theta)^2)^{a+N/2}}{\Gamma(a+N/2)} \sigma^{2(-a-N/2-1)} e^{-\frac{b+\frac{1}{2}\sum_{i=1}^{N}(x_i-\theta)^2}{\sigma^2}} \mathcal{I}_{[0,\infty)}(\sigma^2).$$

This is the so-called **Normal-inverse gamma model for known mean and unknown variance**.

(E) *Suppose that, as above, we take independent observations $X_1, \dots, X_N$ from a normal sampling model with unknown, common mean θ. This time, however, each observation has its own idiosyncratic (but known) variance: $X_i \overset{\text{ind}}{\sim} N(\theta, \sigma_i^2)$. Suppose that θ is given a normal prior distribution with mean m and variance v. Derive the posterior distribution $p(\theta|x_1, \dots, x_N)$. Express the posterior mean in a form that is clearly interpretable as a weighted average of the observations and the prior mean.*

The following model

$$\begin{aligned} &X_1, \dots, X_n|\theta \overset{\text{iid}}{\sim} N(\theta, \sigma_i^2) \\ &\theta \sim N(m, v) \end{aligned}$$

leads to the following posterior distribution:

$$\begin{aligned} p(\theta|\boldsymbol{x}) &\propto p(\boldsymbol{X}|\theta)p(\theta) \\ &\propto \prod_{i=1}^N \left\{ \left(\frac{1}{2\pi\sigma_i^2}\right)^{1/2} e^{-\frac{1}{2\sigma_i^2}(x_i-\theta)^2} \right\} \left(\frac{1}{2\pi v}\right)^{1/2} e^{-\frac{1}{2v}(\theta-m)^2} \\ &\propto e^{-\frac{1}{2}\sum_{i=1}^N \frac{(x_i-\theta)^2}{\sigma_i^2}} e^{-\frac{1}{2v}(\theta-m)^2}. \end{aligned}$$

The exponents can be rewritten as

$$\begin{aligned} &-\frac{1}{2}\left[\sum_{i=1}^N \left(\frac{x_i-\theta}{\sigma_i}\right)^2 + \frac{(\theta-m)^2}{v}\right] \\ =& -\frac{1}{2}\left[\sum_{i=1}^N \left(\frac{x_i^2}{\sigma_i^2} + \frac{\theta^2}{\sigma_i^2} - 2\frac{x_i\theta}{\sigma_i^2}\right) + \frac{\theta^2 + m^2 - 2m\theta}{v}\right] \\ \propto& -\frac{1}{2}\left[\theta^2\left(\frac{1}{v} + \sum_{i=1}^N \frac{1}{\sigma_i^2}\right) - 2\theta\left(\frac{m}{v} + \sum_{i=1}^N \frac{x_i}{\sigma_i^2}\right)\right]. \end{aligned}$$

Therefore, we get

$$\theta|\boldsymbol{x} \sim N\left(\frac{\frac{m}{v} + \sum_{i=1}^N \frac{x_i}{\sigma_i^2}}{\frac{1}{v} + \sum_{i=1}^N \frac{1}{\sigma_i^2}}; \left(\frac{1}{v} + \sum_{i=1}^N \frac{1}{\sigma_i^2}\right)^{-1}\right).$$

This is the so-called **Normal-Normal model for unknown mean and known idiosyncratic variances**.

The precision is the sum of the prior precision and of each single idiosyncratic precision of the data. The mean is again an average of the prior mean and of the weighted average of the data (weighted by the precisions). This heteroschedastic model is useful when dealing with robust regression.

(F) *Suppose that $(X|\sigma^2) \sim N(0, \sigma^2)$, and that $1/\sigma^2$ 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.*

Given the model

$$
\begin{aligned}
&X_1, \dots, X_n | \omega \overset{\text{iid}}{\sim} N(0, \omega) \\
&\omega \sim Gamma(a, b)
\end{aligned}
$$

The marginal distribution of the data is

$$
\begin{aligned}
f_X(x) &= \int f_{X|\omega}(x|\omega) f_\omega(\omega) d\omega \\
&= \int \left(\frac{\omega}{2\pi}\right)^{1/2} e^{-\frac{\omega}{2}x^2} \frac{b^a}{\Gamma(a)} \omega^{a-1} e^{-b\omega} d\omega \\
&= \left(\frac{1}{2\pi}\right)^{1/2} \frac{b^a}{\Gamma(a)} \int \underbrace{\omega^{(a+1/2)-1} e^{-(b+\frac{1}{2}x^2)\omega}}_{Beta(a+\frac{1}{2}, b+\frac{1}{2}x^2)} d\omega \\
&= \left(\frac{1}{2\pi}\right)^{1/2} \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+\frac{1}{2})}{(b+\frac{1}{2}x^2)^{a+1/2}} \\
&= \frac{\Gamma(\frac{2a+1}{2})}{\Gamma(\frac{2a}{2})} \left(\frac{1}{2\pi}\right)^{1/2} \frac{b^a}{\left[b\left(1+\frac{ax^2}{2ab}\right)\right]^{a+1/2}} \\
&= \frac{\Gamma(\frac{2a+1}{2})}{\Gamma(\frac{2a}{2})} \left(\frac{1}{2\pi b}\right)^{1/2} \frac{1}{\left(1+\frac{ax^2}{2ab}\right)^{a+1/2}} \\
&= \frac{\Gamma(\frac{2a+1}{2})}{\Gamma(\frac{2a}{2})} \left(\frac{1}{2\pi a\frac{b}{a}}\right)^{1/2} \frac{1}{\left(1+\frac{x^2}{2a\frac{b}{a}}\right)^{a+1/2}}
\end{aligned}
$$

which is the pdf of $t_{2a}\left(\mu = 0; \sigma^2 = \frac{b}{a}\right)$.

In a Bayesian framework the hyperparameters of the precision parameters are often given as

$$
\begin{aligned}
&X_1, \dots, X_n | \omega \overset{\text{iid}}{\sim} N(\mu, \omega) \\
&\omega \sim Gamma\left(\frac{\nu}{2}; \frac{\nu\sigma_0^2}{2}\right)
\end{aligned}
$$

In that case the marginal of the data is

$$
X \sim t_\nu(\mu; \sigma_0^2).
$$

Problem 2. The multivariate normal distribution

Basics

We all know the univariate normal distribution, whose long history began with de Moivre's 18th-century work on approximating the (analytically inconvenient) binomial distribution. This led to the probability density function

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

for the normal random variable with mean m and variance v, written $X \sim N(m, v)$.

Here's an alternative characterization of the univariate normal distribution in terms of moment-generating functions: a random variable X has a normal distribution if and only if $E\{\exp(tx)\} = \exp(mt + vt^2/2)$ for some real m and positive real v. Remember that $E(\cdot)$ denotes the expected value of its argument under the given probability distribution. We will generalize this definition to the multivariate normal.

(A) *First, some simple moment identities. The covariance matrix cov($\boldsymbol{X}$) of a vector-valued random variable $\boldsymbol{X}$ is defined as the matrix whose (i, j) entry is the covariance between X_i and X_j. In matrix notation, cov($\boldsymbol{X}$) $= E\{(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T\}$, where $\boldsymbol{\mu}$ is the mean vector whose ith component is $E(X_i)$. Prove the following: (1) cov($\boldsymbol{X}$) $= E(\boldsymbol{X}\boldsymbol{X}^T) - \boldsymbol{\mu}\boldsymbol{\mu}^T$; and (2) cov($A\boldsymbol{X} + \boldsymbol{b}$) $= A$cov($\boldsymbol{X}$)A^T for matrix A and vector $\boldsymbol{b}$.*

Let $\boldsymbol{\mu} = E[\boldsymbol{X}]$. We can write

$$\begin{aligned} Cov(\boldsymbol{X}) &= E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T] \\ &= E[\boldsymbol{X}\boldsymbol{X}^T - \boldsymbol{X}\boldsymbol{\mu}^T - \boldsymbol{\mu}\boldsymbol{X}^T\boldsymbol{\mu}\boldsymbol{\mu}^T] \\ &= E[\boldsymbol{X}\boldsymbol{X}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T. \end{aligned}$$

For the second relation,

$$\begin{aligned} Cov(A\boldsymbol{X} + \boldsymbol{b}) &= E[(A\boldsymbol{X} + \boldsymbol{b} - A\boldsymbol{\mu} - \boldsymbol{b})(A\boldsymbol{X} + \boldsymbol{b} - A\boldsymbol{\mu} - \boldsymbol{b})^T] \\ &= E[A(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T A^T] \\ &= AE[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T]A^T \\ &= ACov(\boldsymbol{X})A^T. \end{aligned}$$

(B) *Consider the random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^T$, with each entry having an independent standard normal distribution (that is, mean 0 and variance 1). Derive the probability density function (PDF) and moment-generating function (MGF) of $\boldsymbol{Z}$, expressed in vector notation. We say that $\boldsymbol{Z}$ has a standard multivariate normal distribution.*

If $Z_1, \ldots, Z_p \overset{\text{iid}}{\sim} N(0, 1)$ then the joint density can be obtained as the product of the marginals (independence), that is

$$\begin{aligned} f_{\boldsymbol{Z}}(z_1, \ldots, z_p) &= \prod_{i=1}^{p} \left(\frac{1}{2\pi}\right)^{1/2} e^{-\frac{1}{2}z_i^2} \\ &= \left(\frac{1}{2\pi}\right)^{p/2} e^{-\frac{1}{2}\sum_{i=1}^{p} z_i^2} \\ &= \left(\frac{1}{2\pi}\right)^{p/2} e^{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{z}}. \end{aligned}$$

The mgf of the multivariate standard normal distribution is

$$\begin{aligned} M_{\boldsymbol{Z}}(\boldsymbol{t}) &= E[e^{\boldsymbol{t}^T \boldsymbol{Z}}] \\ &= E[e^{t_1 Z_1} \dots e^{t_p Z_p}] \\ &= E[e^{t_1 Z_1}] \dots E[e^{t_p Z_p}] \\ &= e^{\frac{1}{2}t_1^2} \dots e^{\frac{1}{2}t_p^2} \\ &= e^{\frac{1}{2}\boldsymbol{t}^T \boldsymbol{t}}. \end{aligned}$$

(C) *A vector-valued random variable $\boldsymbol{X} = (X_1, \dots, X_p)^T$ has a multivariate normal distribution if and only if every linear combination of its components is univariate normal. That is, for all vectors $\boldsymbol{a}$ not identically zero, the scalar quantity $Z = \boldsymbol{a}^T \boldsymbol{X}$ is normally distributed. From this definition, prove that $\boldsymbol{X}$ is multivariate normal, written $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$, if and only if its moment-generating function is of the form $E(\exp\{\boldsymbol{t}^T \boldsymbol{X}\}) = \exp(\boldsymbol{t}^T \boldsymbol{\mu} + \boldsymbol{t}^T \Sigma \boldsymbol{t}/2)$. Hint: what are the mean, variance, and moment-generating function of Z, expressed in terms of moments of $\boldsymbol{X}$?*

In order to prove that the mgf of the generic multivariate normal distribution has the form $E(\exp\{\boldsymbol{t}^T \boldsymbol{X}\}) = \exp(\boldsymbol{t}^T \boldsymbol{\mu} + \boldsymbol{t}^T \Sigma \boldsymbol{t}/2)$, we will use the definition of multivariate normal distribution and the results on the univariate mgf of a normal distribution. Let $Z = \boldsymbol{a}^T \boldsymbol{X}$, then we know that:

- $m = E[Z] = E[\boldsymbol{a}^T \boldsymbol{X}] = \boldsymbol{a}^T E[X] = \boldsymbol{a}^T \boldsymbol{\mu}$;
- $v = Var(Z) = Var(\boldsymbol{a}^T \boldsymbol{X}) = \boldsymbol{a}^T Cov(\boldsymbol{X}) \boldsymbol{a} = \boldsymbol{a}^T \Sigma \boldsymbol{a}$.

We prove the two directions of the 'if and only if" statement:

- $\Rightarrow$: we know that $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$ and therefore, for definition, $Z = \boldsymbol{a}^T \boldsymbol{X} \sim N(m, v)$. Thus, using the mgf of a univariate normal distribution

$$M_Z(t) = E[e^{tZ}] = E[e^{t\boldsymbol{a}^T \boldsymbol{X}}] = e^{mt + \frac{1}{2}vt^2} = e^{t\boldsymbol{a}^T \boldsymbol{\mu} + \frac{1}{2} t \boldsymbol{a}^T \Sigma \boldsymbol{a} t}.$$

  Taking $t = 1$ in the equation above leads to

$$M_{\boldsymbol{X}}(\boldsymbol{a}) = E[e^{\boldsymbol{a}^T \boldsymbol{X}}] = e^{\boldsymbol{a}^T \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{a}^T \Sigma \boldsymbol{a}}.$$

- $\Leftarrow$: we now prove that, if a distribution has the mgf above, then it is a multivariate normal distribution. Suppose

$$M_{\boldsymbol{X}}(\boldsymbol{a}) = E[e^{\boldsymbol{a}^T \boldsymbol{X}}] = e^{\boldsymbol{a}^T \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{a}^T \Sigma \boldsymbol{a}},$$

  and define $Z = \boldsymbol{a}^T \boldsymbol{X}$. Then, for $t \in \mathbb{R}$,

$$M_Z(t) = E[e^{tZ}] = E[e^{t\boldsymbol{a}^T \boldsymbol{X}}] = M_{\boldsymbol{X}}(t\boldsymbol{a}) = e^{t\boldsymbol{a}^T \boldsymbol{\mu} + \frac{1}{2}t^2 \boldsymbol{a}^T \Sigma \boldsymbol{a}} = e^{tm + \frac{1}{2}t^2 v},$$

  that is, $Z \sim N(m, v)$. Therefore, $\boldsymbol{X}$ is a multivariate normal distribution.

(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 $\boldsymbol{Z}$ have a standard multivariate normal distribution, and define the random vector $\boldsymbol{X} = L\boldsymbol{Z} + \boldsymbol{\mu}$ for some $p \times p$ matrix L of full column rank. Prove that $\boldsymbol{X}$ is multivariate normal. In addition, use the moment identities you proved above to compute the expected value and covariance matrix of $\boldsymbol{X}$.*

Let $\boldsymbol{Z} \sim N(\boldsymbol{0}, \mathbb{I}_p)$ and let $\boldsymbol{X} = L\boldsymbol{Z} + \boldsymbol{\mu}$, where $L \in \mathbb{R}^{p \times p}$ non-singular matrix. To prove that affine transformations of standard multivariate normals are generic multivariate normals, we use the mgf.

The mgf of the standard multivariate normal is

$$M_{\boldsymbol{Z}}(\boldsymbol{t}) = e^{\frac{\boldsymbol{t}^T\boldsymbol{t}}{2}}$$

and the corresponding mgf of $\boldsymbol{X}$ is

$$\begin{aligned}
M_{\boldsymbol{X}}(\boldsymbol{t}) &= E[e^{\boldsymbol{t}^T\boldsymbol{X}}] \\
&= E[e^{\boldsymbol{t}^T(L\boldsymbol{Z}+\boldsymbol{\mu})}] \\
&= E[e^{\boldsymbol{t}^TL\boldsymbol{Z}+\boldsymbol{t}^T\boldsymbol{\mu}}] \\
&= e^{\boldsymbol{t}^T\boldsymbol{\mu}}E[e^{(L^T\boldsymbol{t})^T\boldsymbol{Z}}] \\
&= e^{\boldsymbol{t}^T\boldsymbol{\mu}}e^{\frac{\boldsymbol{t}^TLL^T\boldsymbol{t}}{2}} \\
&= e^{\boldsymbol{t}^T\boldsymbol{\mu}+\frac{\boldsymbol{t}^TLL^T\boldsymbol{t}}{2}}
\end{aligned}$$

and therefore $\boldsymbol{X} \sim N(\boldsymbol{\mu}, LL^T)$. LL^T is symmetric positive definite (it is a valid covariance matrix) because L has full rank.

(E) *Now for the "only if", Suppose that $\boldsymbol{X}$ has a multivariate normal distribution. Prove that $\boldsymbol{X}$ can be written as an affine transformation of standard normal random variables. (Note: a good way to prove that something can be done is to do it!) Use this insight to propose an algorithm for simulating multivariate normal random variables with a specified mean and covariance matrix.*

To prove that every generic multivariate normal can be expressed as the affine combination of multivariate standard normal distributions, let $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$. Since Σ is a symmetric positive definite matrix, we can write its Cholesky decomposition, that is, $\Sigma = LL^T$. We can define $\boldsymbol{Z} = L^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$ because L is a non-singular matrix.

We know that

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = E[e^{\boldsymbol{t}^T\boldsymbol{X}}] = e^{\boldsymbol{t}^T\boldsymbol{\mu}+\frac{\boldsymbol{t}^TLL^T\boldsymbol{t}}{2}}.$$

The mgf of the standardized random variable $\boldsymbol{Z}$ is

$$
\begin{aligned}
M_{\boldsymbol{Z}}(\boldsymbol{t}) &= E[e^{\boldsymbol{t}^T\boldsymbol{Z}}] = E[e^{\boldsymbol{t}^T L^{-1}(\boldsymbol{X}-\boldsymbol{\mu})}] \\
&= E[e^{\boldsymbol{t}^T L^{-1}\boldsymbol{X}}]e^{-\boldsymbol{t}^T L^{-1}\boldsymbol{\mu}} \\
&= E[e^{(L^{-T}\boldsymbol{t})^T\boldsymbol{X}}]e^{-\boldsymbol{t}^T L^{-1}\boldsymbol{\mu}} \\
&= M_{\boldsymbol{X}}(L^{-T}\boldsymbol{t})e^{-\boldsymbol{t}^T L^{-1}\boldsymbol{\mu}} \\
&= e^{(L^{-T}\boldsymbol{t})^T\boldsymbol{\mu}+\frac{(L^{-T}\boldsymbol{t})^T LL^T L^{-T}\boldsymbol{t}}{2}}e^{-\boldsymbol{t}^T L^{-1}\boldsymbol{\mu}} \\
&= e^{\frac{\boldsymbol{t}^T L^{-1}LL^T L^{-T}\boldsymbol{t}}{2}} \\
&= e^{\frac{\boldsymbol{t}^T\boldsymbol{t}}{2}}.
\end{aligned}
$$

Therefore, $\boldsymbol{Z} \sim N(\boldsymbol{0}, \mathbb{I})$. In other words, $\boldsymbol{X}$ is the linear combination of standard normal distributions.

In order to sample from a multivariate normal distribution, we can draw $Z_i \overset{\text{iid}}{\sim} N(0,1)$, then stack them in a vector and perform a rotation given by the Cholesky decomposition of the matrix of covariance Σ desired. In practice, when the covariance matrix is ill-conditioned (one eigenvalue is approximately 0) it is preferable to use the singular values decomposition because it is a more robust method.

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

We know the pdf of a standard multivariate normal distribution, that is,

$$
f_{\boldsymbol{Z}}(\boldsymbol{z}) = \left(\frac{1}{2\pi}\right)^{p/2} \exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{z}\right)
$$

and we know that the generic $\boldsymbol{X} = L\boldsymbol{Z} + \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \Sigma)$., where $\Sigma = LL^T$ Therefore, we can use the transformation theorem. The inverse transformation is $\boldsymbol{Z} = L^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$, which is one-to-one because the matrix L is non-singular. The determinant of the Jacobian is

$$
\det(J) = \det(L^{-1}) = \det(L)^{-1}.
$$

Moreover, we can express this quantity as a function of the covariance matrix Σ. In fact

$$
\det(\Sigma) = \det(LL^T) = \det(L)\det(L^T) = \det(L)^2
$$

and therefore

$$
\det(L)^{-1} = \det(\Sigma)^{-1/2}.
$$

For this reason, the transformation leads to

$$\begin{aligned}
f_{\boldsymbol{X}}(\boldsymbol{x}) &= f_{\boldsymbol{Z}}(L^{-1}(\boldsymbol{x}-\boldsymbol{\mu})) \\
&= \left(\frac{1}{2\pi}\right)^{p/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(L^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))^T L^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \\
&= \left(\frac{1}{2\pi}\right)^{p/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T L^{-T} L^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \\
&= \left(\frac{1}{2\pi}\right)^{p/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T (LL^T)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \\
&= \left(\frac{1}{2\pi}\right)^{p/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)
\end{aligned}$$

(G) *Let $\boldsymbol{X}_1 \sim N(\boldsymbol{\mu}_1, \Sigma_1)$ and $\boldsymbol{X}_2 \sim N(\boldsymbol{\mu}_2, \Sigma_2)$, where $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent of each other. Let $\boldsymbol{Y} = A\boldsymbol{X}_1 + B\boldsymbol{X}_2$ for matrices A, B of full column rank and appropriate dimension. Note that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ need not have the same dimension, as long as $A\boldsymbol{X}_1$ and $B\boldsymbol{X}_2$ do. Use your previous results to characterize the distribution of $\boldsymbol{Y}$.*

Let $\boldsymbol{X}_1 \sim N_r(\boldsymbol{\mu}_1, \Sigma_1)$ and $\boldsymbol{X}_2 \sim N_s(\boldsymbol{\mu}_2, \Sigma_2)$. Let A be a $n \times r$ matrix and B be a $n \times s$ matrix. The distribution of $\boldsymbol{Y} = A\boldsymbol{X}_1 + B\boldsymbol{X}_2$ can be found via mgf:

$$\begin{aligned}
M_{\boldsymbol{Y}}(\boldsymbol{t}) &= E[e^{\boldsymbol{t}^T\boldsymbol{Y}}] = E[e^{\boldsymbol{t}^T(A\boldsymbol{X}_1+B\boldsymbol{X}_2)}] \\
&= E[e^{\boldsymbol{t}^T A\boldsymbol{X}_1} e^{\boldsymbol{t}^T B\boldsymbol{X}_2}] \\
&= E[e^{\boldsymbol{t}^T A\boldsymbol{X}_1}] E[e^{\boldsymbol{t}^T B\boldsymbol{X}_2}] \\
&= M_{\boldsymbol{X}_1}(A^T\boldsymbol{t}) M_{\boldsymbol{X}_2}(B^T\boldsymbol{t}) \\
&= e^{\boldsymbol{t}^T A\boldsymbol{\mu}_1 + \frac{\boldsymbol{t}^T A\Sigma_1 A^T \boldsymbol{t}}{2}} e^{\boldsymbol{t}^T B\boldsymbol{\mu}_2 + \frac{\boldsymbol{t}^T B\Sigma_2 B^T \boldsymbol{t}}{2}} \\
&= e^{\boldsymbol{t}^T(A\boldsymbol{\mu}_1 + B\boldsymbol{\mu}_2) + \frac{\boldsymbol{t}^T(A\Sigma_1 A^T + B\Sigma_2 B^T)\boldsymbol{t}}{2}}.
\end{aligned}$$

Therefore

$$\boldsymbol{Y} \sim N_n(A\boldsymbol{\mu}_1 + B\boldsymbol{\mu}_2, A\Sigma_1 A^T + B\Sigma_2 B^T).$$

Conditionals and marginals

Suppose that $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$ has a multivariate normal distribution. Let $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ denote an arbitrary partition of $\boldsymbol{X}$ into two sets of components. Because we can relabel the components of $\boldsymbol{X}$ without changing their distribution, we can safely assume that $\boldsymbol{X}_1$ comprises the first k elements of $\boldsymbol{X}$, and $\boldsymbol{X}_2$ the last pk. We will also assume that $\boldsymbol{\mu}$ and Σ have been partitioned conformably with $\boldsymbol{X}$:

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_1 \quad \boldsymbol{\mu}_2)^T \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Clearly $\Sigma_{21} = \Sigma_{12}^T$, as Σ is a symmetric matrix.

(A) *Derive the marginal distribution of $\boldsymbol{X}_1$. (Remember your result about affine transformations.)*

Let us suppose that $\boldsymbol{X} \sim (\boldsymbol{\mu}, \Sigma)$ is a p-dimensional random vector, and let us decompose it in the k-dimensional $\boldsymbol{X}_1$ and in the $p-k$-dimensional $\boldsymbol{X}_2$. Then the marginal distribution of $\boldsymbol{X}_1$ can be found via the following transformation,

$$\boldsymbol{X}_1 = A\boldsymbol{X}$$

where is a $k \times p$ matrix such that

$$A = \left(\begin{array}{cc} \mathcal{I}_{k \times k} & \mathcal{O}_{k \times (p-k)} \end{array} \right).$$

Then, using the result about affine transformations, we know that

$$\boldsymbol{X}_1 \sim N(A\boldsymbol{\mu}, A\Sigma A^T) = N(\boldsymbol{\mu}_1, \Sigma_{11}).$$

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

$$\Omega = \left(\begin{array}{cc} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{array} \right).$$

Using (or deriving!) identities for the inverse of a partitioned matrix, express each block of Ω in terms of blocks of Σ.

Using the identity $\Sigma\Omega = \mathcal{I}_{p \times p}$, we can write the following system of equations:

$$\begin{cases} \Sigma_{11}\Omega_{11} + \Sigma_{12}\Omega_{12}^T = \mathcal{I}_{k \times k} \\ \Sigma_{11}\Omega_{12} + \Sigma_{12}\Omega_{22} = \mathcal{O}_{k \times (p-k)} \\ \Sigma_{12}^T\Omega_{11} + \Sigma_{22}\Omega_{12}^T = \mathcal{O}_{(p-k) \times k} \\ \Sigma_{12}^T\Omega_{12} + \Sigma_{12}\Omega_{22} = \mathcal{I}_{(p-k) \times (p-k)} \end{cases}$$

$$\Rightarrow \begin{cases} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T)\Omega_{11} = \mathcal{I} \\ \Omega_{12} = -\Sigma_{11}^{-1}\Sigma_{12}\Omega_{22} \\ \Omega_{12}^T = -\Sigma_{22}^{-1}\Sigma_{12}^T\Omega_{11} \\ (\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12})\Omega_{22} = \mathcal{I} \end{cases}$$

$$\Rightarrow \begin{cases} \Omega_{11} = (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T)^{-1} \\ \Omega_{12} = -\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12})^{-1} \\ \Omega_{12}^T = -\Sigma_{22}^{-1}\Sigma_{12}^T(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T)^{-1} \\ \Omega_{22} = (\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12})^{-1} \end{cases}$$

(C) *Derive the conditional distribution for $\boldsymbol{X}_1$, given $\boldsymbol{X}_2$, in terms of the partitioned elements of $\boldsymbol{X}$, $\boldsymbol{\mu}$, and Σ. There are several keys to inner peace: work with densities on a log scale, ignore constants that don't affect $\boldsymbol{X}_1$, 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 $\boldsymbol{X}_2$, where the regression matrix can be read off the precision matrix.*

Let us now suppose that the vector $\boldsymbol{X}$ is partitioned in $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, and that both the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ have been partitioned accordingly.

In order to find the distribution of the conditional law $\boldsymbol{X}_2|\boldsymbol{X}_1$, we use the pdf:

$$\begin{aligned}
\log f_{X_2|X_1}(x_2|x_1) &= \log f_{(X_1,X_2)}(x_1,x_2) - \log f_{X_1}(x_1) \\
&\propto \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^T\Sigma_{22}^{-1}(\boldsymbol{x}_2-\boldsymbol{\mu}_2) \\
&= \frac{1}{2}(\boldsymbol{x}_1-\boldsymbol{\mu}_1)^T\Omega_{11}(\boldsymbol{x}_1-\boldsymbol{\mu}_1) + (\boldsymbol{x}_2-\boldsymbol{\mu}_2)^T\Omega_{21}(\boldsymbol{x}_1-\boldsymbol{\mu}_1) + \\
&\quad \frac{1}{2}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^T\Omega_{22}(\boldsymbol{x}_2-\boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^T\Omega_{22}(\boldsymbol{x}_2-\boldsymbol{\mu}_2) \\
&= \frac{1}{2}[(\boldsymbol{x}_1-\boldsymbol{\mu}_1)^T\Omega_{11} + 2(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^T\Omega_{21}](\boldsymbol{x}_1-\boldsymbol{\mu}_1) \\
&\propto \frac{1}{2}\left[\boldsymbol{x}_1^T\Omega_{11}\boldsymbol{x}_1 - \boldsymbol{\mu}_1^T\Omega_{11}\boldsymbol{x}_1 + 2(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^T\Omega_{21}\boldsymbol{x}_1 - \boldsymbol{x}_1^T\Omega_{11}\boldsymbol{\mu}_1\right] \\
&\propto \frac{1}{2}\boldsymbol{x}_1^T\Omega_{11}\boldsymbol{x}_1 + [(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^T\Omega_{21} - \boldsymbol{\mu}_1^T\Omega_{11}]\boldsymbol{x}_1,
\end{aligned}$$

where at each step we dropped the terms not depending on $\boldsymbol{x}_1$. Using the "completing the square" trick, we get:

$$\log f_{X_2|X_1}(x_2|x_1) = \frac{1}{2}(\boldsymbol{x}_1-\boldsymbol{m})^TM(\boldsymbol{x}_1-\boldsymbol{m}) + v,$$

where

$$\boldsymbol{m} = \boldsymbol{\mu}_1 - \Omega_{11}^{-1}\Omega_{12}(\boldsymbol{x}_2-\boldsymbol{\mu}_2) = \boldsymbol{\mu}_1 - \Omega_{11}^{-1}\Omega_{12}(\boldsymbol{x}_2-\boldsymbol{\mu}_2) = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)$$

and

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

Therefore,

$$\boldsymbol{X}_2|\boldsymbol{X}_1 \sim N(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2-\boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T).$$

Problem 3. Multiple regression: three classical principles for inference

Suppose we observe data that we believe to follow a linear model, where $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i, i = 1, \ldots, n$. To fix notation: y_i is a scalar response; $\boldsymbol{x}_i$ is a p-vector of predictors or features; and the ε_i are errors. By convention we write vectors as column vectors. Thus $\boldsymbol{x}_i^T\boldsymbol{\beta}$ will be our typical way of writing the inner product between the vectors $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$.

Consider three classic inferential principles that are widely used to estimate $\boldsymbol{\beta}$, the vector of regression coefficients. In this context we will let $\hat{\boldsymbol{\beta}}$ denote an estimate of $\boldsymbol{\beta}$ - think, it wears a hat because it's masquerading as the true value.

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

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 \right\}.$$

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:

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

Here $p_i(y_i|\boldsymbol{\beta}, \sigma^2)$ is the conditional probability density function of y_i, given the model parameters $\boldsymbol{\beta}$ and σ^2.

Method of moments: Choose $\hat{\boldsymbol{\beta}}$ so that the sample covariance between the errors and each of the p predictors is exactly zero. This gives you a system of p equations and p unknowns.

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

Via **least squares**, we get

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}) \right\} \\
&= \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} (\boldsymbol{Y} - X\boldsymbol{\beta})^T (\boldsymbol{Y} - X\boldsymbol{\beta}) \\
&= \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}}\, l(\boldsymbol{\beta}).
\end{aligned}$$

We compute the gradient of the loss function that we want to minimize

$$\nabla_{\boldsymbol{\beta}} l(\boldsymbol{\beta}) = -2X^T (\boldsymbol{Y} - X\boldsymbol{\beta})$$

and we set it to 0, getting

$$\begin{aligned}
&X^T (\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}) = 0 \\
\Rightarrow &X^T \boldsymbol{Y} - X^T X \hat{\boldsymbol{\beta}} = 0 \\
\Rightarrow &\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{Y}.
\end{aligned}$$

With the **maximum likelihood** method, we obtain

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmax}} \left\{ \prod_{i=1}^n p(y_i|\boldsymbol{\beta}, \sigma^2) \right\} \\
&= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmax}} \prod_{i=1}^n \left\{ \left(\frac{1}{2\pi\sigma^2} \right)^{1/2} e^{-\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2} \right\} \\
&= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmax}}\, e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2} \\
&= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmax}}\, e^{-\frac{1}{2\sigma^2}(\boldsymbol{Y} - X\boldsymbol{\beta})^T(\boldsymbol{Y} - X\boldsymbol{\beta})} \\
&= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} (\boldsymbol{Y} - X\boldsymbol{\beta})^T(\boldsymbol{Y} - X\boldsymbol{\beta}) \\
&= (X^TX)^{-1}X^T\boldsymbol{Y}.
\end{aligned}$$

The **method of moments** uses the following relation:

$$\widehat{\text{Cov}}(\boldsymbol{e}, \boldsymbol{x}_j) = 0 \qquad \forall j = 1, \dots p$$

This implies

$$\begin{aligned}
&\frac{1}{n-1}\sum_{i=1}^n (x_{ij} - \overline{\boldsymbol{x}}_j)(e_i - \overline{\boldsymbol{e}}) = 0 \\
\Rightarrow &\sum_{i=1}^n \{x_{ij}e_i\} - \overline{\boldsymbol{x}}_j\overline{\boldsymbol{e}} = 0.
\end{aligned}$$

Without loss of generality, we can assume that the covariates have been centred. In this case, the last term cancels and we can rewrite the previous set of equations in matricial form as $X^T\boldsymbol{e} = 0$. Therefore

$$\begin{aligned}
&X^T\boldsymbol{e} = 0 \\
&X^T(\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}) = 0 \\
&\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\boldsymbol{Y},
\end{aligned}$$

which is the same solution we obtained via Least Squares and Maximum Likelihood.

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

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \sum_{i=1}^n w_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 \right\}$$

where the w_i are weights (trustworthy observations have large weights). Derive this estimator, and show that it corresponds to the maximum-likelihood solution under heteroscedastic Gaussian error:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmax}} \left\{ \prod_{i=1}^n p(y_i|\boldsymbol{\beta}, \sigma^2) \right\}.$$

Here $p_i(y_i|\boldsymbol{\beta}, \sigma_i^2)$. Make sure you explicitly connect the weights w_i and the idiosyncratic variances σ_i^2.

The **weighted least squares** problem can be formulated as follows:

$$
\begin{aligned}
\hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} w_i (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 \right\} \\
&= \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}}\, l(\boldsymbol{\beta}).
\end{aligned}
$$

Let us write the WLS objective (or cost) function

$$
l(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2
$$

in a matricial form. We get, by defining $W = \text{diag}\{w_1, \dots, w_n\}$,

$$
\begin{aligned}
l(\beta) &= (\boldsymbol{Y} - X\boldsymbol{\beta})^T W (\boldsymbol{Y} - X\boldsymbol{\beta}) \\
&= \boldsymbol{Y}^T W \boldsymbol{Y} - \boldsymbol{Y}^T W X \boldsymbol{\beta} - (X\boldsymbol{\beta})^T W \boldsymbol{Y} + (X\boldsymbol{\beta})^T W X \boldsymbol{\beta}.
\end{aligned}
$$

Let us remark here that the second and the third term of the sum can be collected in one single term. In fact

$$
(\boldsymbol{Y}^T W X \boldsymbol{\beta})^T = (X\boldsymbol{\beta})^T W \boldsymbol{Y},
$$

i.e. the first is the transposed of the second but, since they are real numbers, they are also equal. Hence we have

$$
l(\boldsymbol{\beta}) = \boldsymbol{Y}^T W \boldsymbol{Y} - 2\boldsymbol{\beta}^T X^T W \boldsymbol{Y} + \boldsymbol{\beta}^T X^T W X \boldsymbol{\beta}.
$$

Since we are trying to minimize a convex function, we can just find its stationary points. To do this, we first need to calculate the gradient of the WLS objective function equalizing it to 0. Thus,

$$
\begin{aligned}
&\nabla l(\hat{\boldsymbol{\beta}}) = 0 \\
\Rightarrow& -2X^T W \boldsymbol{Y} + 2X^T W X \hat{\boldsymbol{\beta}} = 0 \\
\Rightarrow& (X^T W X)\hat{\boldsymbol{\beta}} = X^T W \boldsymbol{Y}.
\end{aligned}
$$

Solving the linear system requires the matrix $X^T W X$ to be nonsingular (and therefore invertible). The matrix $X^T W X$ is positive definite, and therefore nonsingular, in case X has full rank (i.e. the columns of X are linearly independent). The last equation can be easily solved by inversion, which leads to

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

On the other hand, the **maximum likelihood** problem yields

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmax}} \left\{ \prod_{i=1}^{n} p(y_i|\boldsymbol{\beta}, \sigma_i^2) \right\} \\
&= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmax}} \prod_{i=1}^{n} \left\{ \left(\frac{1}{2\pi\sigma_i^2} \right)^{1/2} e^{-\frac{1}{2\sigma_i^2}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2} \right\} \\
&= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmax}}\, e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma_i^2}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2} \\
&= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{1}{\sigma_i^2}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2
\end{aligned}$$

which is the same minimization problem as WLS if we set $w_i = 1/\sigma_i^2$, and therefore yields to the same solution. Let us remark, in this case, that the weight of the WLS problem correspond to the precisions of each data point.

Problem 4. Quantifying uncertainty: some basic frequentist ideas

In linear regression

In frequentist inference, inferential uncertainty is usually characterized by the sampling distribution, which expresses how one's estimate is likely to change under repeated sampling. The idea is simple: unstable estimators shouldn't be trusted, and should therefore come with large error bars. This should be a familiar concept, but in case it isn't, consult the tutorial on sampling distributions in this chapter's references. Suppose, as in the previous section, that we observe data from a linear regression model with Gaussian error:

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\mathcal{I}).$$

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

We know from the previous exercise that, if

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\mathcal{I}_n)$$

then $\boldsymbol{Y} \sim N(X\boldsymbol{\beta}, \sigma^2\mathcal{I}_n)$. The estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = A\boldsymbol{Y}$, where $A = (X^TX)^{-1}X^T$.

Therefore,

$$\begin{aligned}
&\hat{\boldsymbol{\beta}} \sim N(AX\boldsymbol{\beta}, A\sigma^2\mathcal{I}_nA^T) \\
\Rightarrow &\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1}).
\end{aligned}$$

(B) *This 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. Then consult the data set on ozone concentration in Los Angeles,* *where the goal is to regress daily ozone concentration on a set of other atmospheric variables. This is available from the R package "mlbench", with my R script "ozone.R" giving you a head start on processing things. Calculate standard errors using your method, and then using the pre-packaged* `lm` *function in R. Note: you may have an essentially correct strategy for calculating standard errors that yields something slightly different from the* `lm` *function. If so, that's OK - can you explain the discrepancy?*

We could use the estimated standard deviation instead of σ^2, that is

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{RSS}{n-p-1}$$

and the standard deviation of each component of $\hat{\boldsymbol{\beta}}$ would become

$$\text{sd}(\hat{\beta}_i) = \hat{\sigma}^2 (X^T X)^{-1}_{ii}. \tag{1}$$

This procedure is justified by the fact that the residual sum of squares is an unbiased estimator. In fact, let $H = X(X^T X)^{-1}X^T$ be the perpendicular projection matrix on $C(X)$, then

$$\begin{aligned}
E[RSS] &= E[||\boldsymbol{Y} - X(X^T X)^{-1}X^T\boldsymbol{Y}||^2] \\
&= E[||(I-H)\boldsymbol{Y}||^2] \\
&= E[||(I-H)(X\boldsymbol{\beta}+\boldsymbol{\varepsilon})||^2] \\
&= E[||(I-H)\boldsymbol{\varepsilon}||^2] \\
&= E[\boldsymbol{\varepsilon}^T(I-H)^T(I-H)\boldsymbol{\varepsilon}] \\
&= E[\boldsymbol{\varepsilon}^T(I-H)\boldsymbol{\varepsilon}] \\
&= E[\text{tr}(\boldsymbol{\varepsilon}^T(I-H)\boldsymbol{\varepsilon})] \\
&= E[\text{tr}((I-H)\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon})] \\
&= \text{tr}(I-H)\sigma^2,
\end{aligned}$$

and

$$\text{tr}(I-H) = \text{tr}(I) - \text{tr}(H) = n-p-1.$$

Thus, $E[\hat{\sigma}^2] = \sigma^2$.

In Table 1 we can check that the estimate obtained with (1) is the same that R's built-in package `lm` provides.

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 of $\hat{\boldsymbol{\theta}}$:

$$\hat{\Sigma} \approx Cov(\hat{\boldsymbol{\theta}}) = E[(\hat{\boldsymbol{\theta}} - \overline{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \overline{\boldsymbol{\theta}})^T],$$

	Proposed estimate	lm package
Intercept	38.328694	38.328694
V5	0.007251	0.007251
V6	0.174143	0.174143
V7	0.023769	0.023769
V8	0.069299	0.069299
V9	0.124714	0.124714
V10	0.000394	0.000394
V11	0.014777	0.014777
V12	0.119292	0.119292
V13	0.004896	0.004896

Table 1: Comparison between the function we implemented (left column) and R's built-in function.

where the expectation is under the sampling distribution for the data, given the true parameter $\boldsymbol{\theta}$. Here $\overline{\boldsymbol{\theta}}$ 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 estimated covariance matrix: $\hat{\Sigma}_{jj} = \sigma_j^2$ 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 $f(\theta) = \theta_1 + \theta_2$. Calculate the standard error of $f(\hat{\theta})$, and generalize this to the case where f is the sum of all p components of $\hat{\theta}$.*

In the trivial case when $f(\boldsymbol{\theta}) = \theta_1 + \theta_2$, we get

$$\begin{aligned}
\text{Var}(f(\boldsymbol{\theta})) &= \text{Var}(\theta_1 + \theta_2) \\
&= E[(\theta_1 + \theta_2)^2] - (E[\theta_1 + \theta_2])^2 \\
&= E[\theta_1^2 + \theta_2^2 + 2\theta_1\theta_2] - (E[\theta_1] + E[\theta_2])^2 \\
&= E[\theta_1^2] + E[\theta_2^2] + 2E[\theta_1\theta_2] - (E[\theta_1])^2 + (E[\theta_2])^2 - 2E[\theta_1]E[\theta_2] \\
&= \text{Var}[\theta_1] + \text{Var}[\theta_2] + 2[E(\theta_1\theta_2) - E(\theta_1)E(\theta_2)] \\
&= \text{Var}[\theta_1] + \text{Var}[\theta_2] + 2\text{Cov}(\theta_1, \theta_2).
\end{aligned}$$

In the general case, the proof is identical, i.e.

$$\begin{aligned}
\text{Var}(f(\boldsymbol{\theta})) &= \text{Var}(\sum_{i=1}^{p} \theta_i) \\
&= E\left[\left(\sum_{i=1}^{p}\right)^2\right] - \left(E\left[\sum_{i=1}^{p}\theta_i\right]\right)^2 \\
&= E\left[\sum_{i=1}^{p}\sum_{j=1}^{p}\theta_i\theta_j\right] - \left(\sum_{i=1}^{p}E[\theta_i]\right)^2 \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p}E\left[\theta_i\theta_j\right] - \sum_{i=1}^{p}\sum_{j=1}^{p}E[\theta_i]E[\theta_j] \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p}\text{Cov}(\theta_i,\theta_j) \\
&= \sum_{i=1}^{p}\text{Var}(\theta_i) + 2\sum_{i<j}\text{Cov}(\theta_i,\theta_j).
\end{aligned}$$

(B) *What now if f is a nonlinear function of the θ_j's? Propose an approximation for var$\{f(\hat{\boldsymbol{\theta}})\}$, where f is any sufficiently smooth function. (As above, the variance is under the sampling distribution of the data, given the true parameter.) There are obviously many potential strategies that might work, but here's one you might find fruitful: try a first-order Taylor approximation of $f(\hat{\theta})$ around the unknown true value θ. Try to bound the size of the likely error of the approximation, or at least talk generally about what kinds of assumptions or features of f or $p(\hat{\theta}|\theta)$ might be relevant. You should also reflect on some of the potential caveats of this approach.*

If f is a generic smooth function, we can write the first-order Taylor approximation of $f(\hat{\boldsymbol{\theta}})$ around the unknown true value $\boldsymbol{\theta}$, that is

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

Therefore,

$$\text{Var}\left\{f(\hat{\boldsymbol{\theta}})\right\} \approx (\nabla f(\boldsymbol{\theta}))^T \cdot \text{Var}(\hat{\boldsymbol{\theta}}) \cdot (\nabla f(\boldsymbol{\theta})).$$

Bootstrapping

The basic idea is to simulate the process of repeated sampling from the population by resampling from your sample (with replacement). The ties and duplicates in your "bootstrapped samples" will mimic the sampling variability of the true data-generating process.

(A) *Let $\hat{\Sigma}$ denote the covariance matrix of the sampling distribution of $\hat{\boldsymbol{\beta}}$, the least-squares estimator. Write an R function that will estimate $\hat{\Sigma}$ via bootstrapped resampling for a given design matrix*

X and response vector $\boldsymbol{Y}$. Use it to compute $\hat{\Sigma}$ for the ozone data set, and compare it to the parametric estimate based on normal theory.

From the normal theory, we know that

$$\hat{\Sigma} = \text{Cov}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(X^T X)^{-1}.$$

We present here two different versions of nonparametric bootstrap. The first one consists in considering the couple $(\boldsymbol{X}_i, Y_i)$ as random, and thus resampling with replacement from the rows of the matrix X and for the corresponding Y's. Each bootstrap sample of size n, say $(X, \boldsymbol{Y})^{(b)}$, leads to an estimate $\hat{\beta}^{(b)}$, whose covariance matrix is $\hat{\Sigma}^{(b)} = \hat{\sigma}^{2(b)}((X^{(b)})^T X^{(b)})^{-1}$. The estimated covariance is the average of each one of the bootstrapped estimated covariance matrices.

The second approach consists in considering X as a fixed design matrix, and resampling only the residuals from their empirical cdf. In fact, we can compute $\boldsymbol{\varepsilon} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}$, and resample with replacement from $\boldsymbol{\varepsilon}$. Once we have the bootstrapped sample $\boldsymbol{\varepsilon}^{(b)}$, we can recompute $\boldsymbol{y}^{(b)}$ and, subsequently, the estimate $\hat{\Sigma}^{(b)}$.

Both the methods provide with a satisfactory estimate of the covariance matrix. In order to evaluate which method performs better, we used the mean squared error between the elements of $\hat{\Sigma}$, the original estimate, and the elements of the bootstrap estimate. In Table 2 the results are reported. As one can see, the first method performs better.

	Bootstrap $(X, Y)_i$	Bootstrap ε_i
MSE	8.7702	54.3932

Table 2: Table reporting the MSE for the two bootstrap techniques.

(B) *Now let's a few of these ideas. Write R functions that will accomplish the following:*

1. *For a specified mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, simulate multivariate normal random variables.*
2. *For a given sample $\boldsymbol{X}_1, \dots, \boldsymbol{X}_N$ from a multivariate normal distribution, estimate the mean vector and covariance matrix by maximum likelihood.*
3. *Bootstrap a given sample $\boldsymbol{X}_1, \dots, \boldsymbol{X}_N$ to estimate the sampling distribution of the MLE.*

Try out your code in $d = 2$ dimensions for a few different sample sizes N. See how well you can recover the true covariance matrix from simulated data.

Appendix A

R code